ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2022, Vol. 17(1) 191–204

DOI: 10.1177/1745691621990369
www.psychologicalscience.org/PPS
SAGE

The Number of Senders and Total Judgments Matter More Than Sample Size in Deception-Detection Experiments

Timothy R. Levine[1], Yasuhiro Daiku[2], and Jaume Masip[3]
[1]Department of Communication Studies, University of Alabama Birmingham; [2]Graduate School of Human Sciences, Osaka University; and [3]Department of Social Psychology and Anthropology, University of Salamanca

Abstract
Hundreds of experiments have examined people's ability to distinguish truths from lies. Meta-analyses suggest that the findings from larger scale experiments converge and that findings discrepant from the meta-analytic average of 54% occur in only smaller experiments. Study size (number of data points, or total number of judgments) is a joint function of the sample size and the number of judgments per research participant. Furthermore, because senders vary more than judges, experiments involving few senders may not be replicable. A number of simulations are reported in which the sample size, the number of unique senders, and the number of judgments per research participant are varied. The findings demonstrate that stability is more a function of the number of judgments than the sample size and that experiments involving too few senders risk idiosyncratic findings that are less likely to be replicable. Implications for research design are discussed.

Keywords
deception detection, number of judgments, sample size, random response error, replicability

There is a large experimental literature on human ability to distinguish lies from honest communication (for a recent review, see Levine, 2020). A meta-analysis of prior findings suggests that people are poor lie detectors, achieving an average of just under 54% accuracy in truth/lie discrimination (Bond & DePaulo, 2006). Although the average is both well documented and widely accepted in the scientific community (Levine, 2020), the findings are tied to a specific conventionally determined set of methodological decisions (e.g., Levine, 2015, 2018). Nevertheless, at least within the confines of commonly used experimental designs, the 54% average is a reliable and reproducible finding.

The typical human-to-human deception-detection experiment involves two types of participants: senders and judges. Senders provide stimulus materials in the form of honest-truthful messages and lies, the ground truth of which is known to the researchers but not the judges. The judges evaluate the senders' messages for honesty-deception. Deception-detection accuracy is the extent to which judges are correct in their evaluations of sender honesty or duplicity. In human-to-human deception-detection experiments, the sample size is conventionally based on the number of judges, and judge accuracy scores are the unit of analysis.

In some experiments, each judge makes a single evaluation of a different sender (e.g., Vernham et al., 2014). The single judgments are just hit or miss, zero or perfect. In other experiments, judges rate multiple messages from the same sender (e.g., Burgoon et al., 1994), single messages by multiple senders (e.g., Levine et al., 2014), or multiple messages each from multiple senders (e.g., Levine et al., 2005). Although less frequent, some experiments cross senders with judges who assess multiple messages from each of multiple senders in round-robin designs in which senders are also judges (e.g., Levine, 2016; Masip et al., 2020). In designs in which judges make multiple judgments, each judge's accuracy score can be averaged across senders

Corresponding Author:
Jaume Masip, Department of Social Psychology and Anthropology, University of Salamanca
E-mail: jmasip@usal.es

and messages like scores on a multiitem true/false test. In such cases, accuracy for each judge is a percentage rather than a dichotomous hit-or-miss variable. Alternatively, for some designs, the data could be understood as multilevel data with dichotomous judgments of specific messages nested within senders and senders nested with judges. Regardless of which approach is used, the total number of judgments in an experiment is the product of the number of judges multiplied by the number of judgments per judge.

Here, we explore two underappreciated research design features that we believe strongly affect experimental results: the total number of judgments and the number of senders. We argue that within the confines of typical deception-detection experiments, study-to-study variability is more strongly a function of the number of senders and the total number of judgments (i.e., number of judgments per judge times number of judges) than of the sample size (i.e., number of judges). Studies involving few senders or few judgments per judge can produce unstable, idiosyncratic results, which are problematic and contribute to the replication crisis. Science, of course, requires results that hold up across studies. Thus, obtaining a sufficient number of judgments from a sufficient sample of senders is critical for forming sound scientific inferences. Analyses of statistical power are insufficient because both the number of judgments and the number of senders are critical issues, not just the sample size (Judd et al., 2012).

Illustrative Examples

To clarify our title and our claims, we have in mind a variety of fairly typical experimental designs. In the first design, there are 50 judges, each of whom interviews one of 50 senders (half honest, half lying) and makes a single truth/lie assessment that is scored for accuracy, yielding 50 total judgments. This type of design is reflected in approximately 40% of the human deception-detection studies in the Vrij et al. (2017) meta-analysis. In the second design, 50 judges view and rate 20 "sender tapes" (half are truths, half are lies) for 1,000 total judgments. Each tape contained one message. Each set of 20 tapes could contain one message from each of 20 senders, two messages from each of 10 senders, four messages from each of five senders, 10 messages from each of two senders, or 20 messages from a single sender. For comparison with actual practice, the median experiment included in the Bond and DePaulo (2006) meta-analysis had 41.5 judges (range = 1–816), 16 senders (range = 1–200), and 16 judgments per judge (range = 1–416).

Although the sample size ($N = 50$) across our hypothetical variations is the same, we argue that the precision and stability of the results are quite different. The first design yields just 50 judgments compared with 1,000 in the second. Given the nature of random error, results from the first design are necessarily much more variable that those of the second. But we also have a second consideration in mind. We think that small samples of senders are more problematic than small samples of judges because senders vary much more (Bond & DePaulo, 2008), and hence extremely small samples of senders can yield results that are different from the population and are unlikely to be replicable. Thus, the 50-judge, 50-judgment design yields highly unstable results (i.e., it is subject to large random variation) despite a sizable sample of 50 senders. A 50-judge, 1,000-judgment design will be much more stable regardless of the number of senders (less random variation), but the results may still be idiosyncratic and may not be replicable if it is based on just one or only a very few senders. The least important consideration among the number of judges, number of judgments, and number of senders is the number of judges because judges have very little variance (Bond & DePaulo, 2008). Ironically, this least important aspect seems to matter most for deception researchers.

There are good practical reasons why researchers focus on recruiting judges and are less concerned about the numbers of senders and judgments. Running senders is more complex than running judges. Judges can be run in groups, whereas senders must be run individually. For some studies, the senders have to participate in an hour-long experimental scenario before lying or telling the truth. Typically, the tapes of the senders' messages have to be edited before being shown to judges. It is also understandable that in studies in which long interviews are conducted, each interviewer does just one interview and renders just one judgment. Thus, it is reasonable for researchers to run few senders if they do not see the benefit of doing otherwise. One goal of this article is to show researchers the negative consequences of having an insufficient number of senders or total judgments in their designs.

Four Underappreciated Insights From Meta-Analyses

Collectively, three comprehensive meta-analyses nicely summarize the deception-detection literature on judges' accuracy (Aamodt & Custer, 2006; Bond & DePaulo, 2006, 2008). We argue that there are four particularly noteworthy but often overlooked conclusions stemming from these meta-analyses.

First, accuracy results are remarkably homogeneous across different experiments and labs. This observation is noteworthy because heterogeneity is ubiquitous in most social-science literatures (Levine & Weber, 2020)

and because the meta-analytic average can be used as a benchmark for evaluating the typicality of findings from individual experiments. Across experiments, accuracy results are normally distributed around the mean of 54%, and accuracy is just a little better than chance across a wide range of potential moderators such as lie stakes, judge expertise, communication medium, the extent of interaction between sender and judge, and sender spontaneity-preparation (Bond & DePaulo, 2006). The bottom line is that the mean reflects the literature well; a vast number of experiments neatly and symmetrically converge around 54%, and results departing from the mean may not be replicable.

Second, standard errors around accuracy averages are often quite small even in primary experiments. Fifty-four percent correct is significantly better than 50–50 chance with a moderate effect size of $d = 0.40$ (Bond & DePaulo, 2006). Thus, it is factually incorrect to round 54% down to mere chance. Accuracy is reliably better than chance even though 50% and 54% appear on the face of it to be much more similar than different. The reason that the small difference between 50% and 54% is statistically meaningful is the small standard error. The reason for the small standard errors, in turn, appears to be very little judge-to-judge variance in accuracy within experiments (Bond & DePaulo, 2008). Individual differences such as age, education, cognitive ability, experience, and personality traits such as neuroticism, extroversion, and self-monitoring are not correlated with deception-detection accuracy (Aamodt & Custer, 2006). Because judges do not vary much, as the number of judgments per judge increases, judges converge with other judges to produce a stable average with a small standard error.

Third, the Bond and DePaulo (2006) meta-analysis provided a funnel plot that graphed the number of judgments onto accuracy. The resulting funnel was nicely symmetrical and showed that virtually every finding that departed more than few points from the across-study average had one design feature in common. All discrepant data points were attributable to studies involving relative fewer judgments. The fewer the judgments, the more study-to-study variably. As the number of judgments increased, findings converged into a small range of values centered on 54%. The strongest predictor of accuracy appears to be the amount of data.

Finally, although judges are remarkably homogeneous, senders are not. Bond and DePaulo (2008) estimated that the standard deviation for senders is approximately 7 times larger than that of judges. Simply put, some senders are more difficult to detect than others. Levine (2020) called these matched and mismatched senders. Any given sender might be matched (demeanor and actual honesty align) and easy to detect or mismatched (demeanor is opposite ground truth) and systematically yield accuracy well below chance. With a larger number of senders, matched and mismatched senders will average out. However, this becomes less likely with fewer senders.

In short, meta-analyses show that judges' lie-detection accuracy is strikingly similar across studies, standard errors are very small, and discrepant values have been found only in studies with few judgments. Few total judgments can be a consequence of either few judges or few judgments per judge. We briefly conceptualize the separate impact of each of these two aspects on variation in the next section. Meta-analyses also show that senders are much less homogeneous than receivers; the implications of this difference in variation are also explained below.

Two Sources of Random Error

There are at least two important nonsystematic reasons why deception detection accuracy scores vary from judge to judge and experiment to experiment. The first source of variation is sampling error, which is a function of the number of judges. The second source of variation is random response error, which is a function of the number of judgments per judge. We argue that the number of judgments per judge is analogous to the number of items on a test or a scale. It has long been known that all other things being equal, reliability increases with the number of items and that single-item measurement is notoriously unreliable (Kuder & Richardson, 1937).

We believe that within deception-detection research, the role of sampling error is more widely recognized and appreciated than the role of random response error. This is unfortunate because random response error may be just as important. Educational psychologists would not endorse assessing student ability with a single true/false question because single-item measurement is unreliable. Likewise, measuring deception detection accuracy with a single truth/lie assessment is similarly unreliable.

Adding Sender Variation Into the Mix

The impact of random error is well known to people with training in statistics. Less obvious is differences in variability between senders and judges. As previously noted, there is more stimulus side variability in senders than judge variability (Bond & DePaulo, 2008).

A second reason why educational psychologists would not endorse assessing student ability with a single true/false question is that single-item measures surely lack content validity. That is, any one item is

unlikely to capture the full bandwidth of the construct being assessed. Likewise, when there is a judgment of a single sender, that sender might be matched, mismatched, or somewhere in between. Multiple senders mitigates against such stimulus-side idiosyncrasies. Thus, the sheer amount of data is not the only consideration because senders and judges are distributed differently. Sampling a range of senders is also important.

Implications for Deception Research

The preceding discussion leads to two testable hypotheses. First, consistent with Bond and DePaulo's (2006) funnel plot and basic statistical theory, accuracy scores (raw percentage correct in truth/lie discrimination) should become more stable as a function of the total number of judgments (Sample Size × Judgments per Judge). Although this is mathematically true, it can be demonstrated. The continued existence of experiments with a single judgment per judge demonstrates that the issue is an ongoing concern.

Second, of the components of the number of judgments, increasing the number of senders should have a greater impact on stability of accuracy score than increasing the number of judges. This is simply because senders vary more than judges. Thus, we predict that if the number of judges is held constant, increasing the number of senders leads to more stable findings.

Beyond accuracy, our analyses also examine the impact of sample size, number of senders, and total number of judgments on truth-bias scores. Presumably, truth-bias will be affected by the number of senders and the number of judges in the same way as accuracy scores.

Method

We examined these issues by conducting a number of analyses and simulations using a data set we collected for a previous study (Masip et al., 2020). To our knowledge, Masip et al.'s study is the largest existing experiment crossing senders and judges; therefore, this data set is most well suited for the current purpose. Other than rereporting basic descriptive statistics for the data, all the analyses reported here are original to this report. The data are available for research purposes upon reasonable requests.

Data content

All participants in Masip et al.'s (2020) study (N = 50) acted as both senders and judges. Senders were asked eight separate short questions about topics such as their favorite vacation destination, least favorite class, parents' occupation, and so on. They answered four questions deceptively and the other four truthfully (randomly determined). Judges made binary truth/lie judgments for each individual response of each sender except themselves. Those judges who already knew the truth because they were friends with the sender wrote an asterisk next to their judgment. Participants were not informed about the 50/50 truth/lie split across all responses. Because 11 senders got momentarily confused in following the instructions, the final data set contained slightly more deceptive (51%) than truthful (49%) replies.

In all, 19,600 truth/lie judgments were collected (50 Judges × 49 Senders × 8 Answers). Two hundred and fifty-two judgments (1.29%) were discarded because they were followed by an asterisk, which left 19,348 usable judgments. We reasoned that this number of judgments was more than sufficient to test the above predictions because the funnel plot in Bond and DePaulo's (2006) meta-analysis suggests that findings stabilize at approximately 4,000 judgments. The single experiment with the most judgments in Bond and DePaulo's meta-analysis had 10,304 judgments (DePaulo & Pfeifer, 1986). Our data set nearly doubled this amount, thus allowing for analyses that apply to but also extend beyond the existing literature.

As shown in Figure 1, accuracy and truth bias were calculated by sender (i.e., across judges) and by judge (across senders). Consistent with the Bond and DePaulo (2006) meta-analysis, judge accuracy was distributed tightly around slightly better than chance, whereas truth bias showed greater variability. Consistent with the Bond and DePaulo (2008) meta-analysis, for accuracy, the standard deviation of judges was smaller than that of senders. Masip et al. (2020) provided more detailed descriptive information.

Simulation procedure

We conducted several simulations that followed the same basic procedure. Imagine long-formatted data. Each row contains one judgment and the corresponding accuracy of that judgment. The columns include judge ID, sender ID, judgment (1 = truth or 0 = lie), and accuracy (1 = accurate or 0 = inaccurate). First, we randomly extracted a specific number of judges and senders and filtered the data by their judge ID and sender ID. Second, we randomly extracted one judgment and the corresponding accuracy score from each of the selected judge-sender pairs and then calculated averages across judgments (i.e., truth bias) and across accuracy scores.

For example, if we pick three judges (Judge IDs 1, 2, and 3) and three senders (Sender IDs 4, 5, and 6),

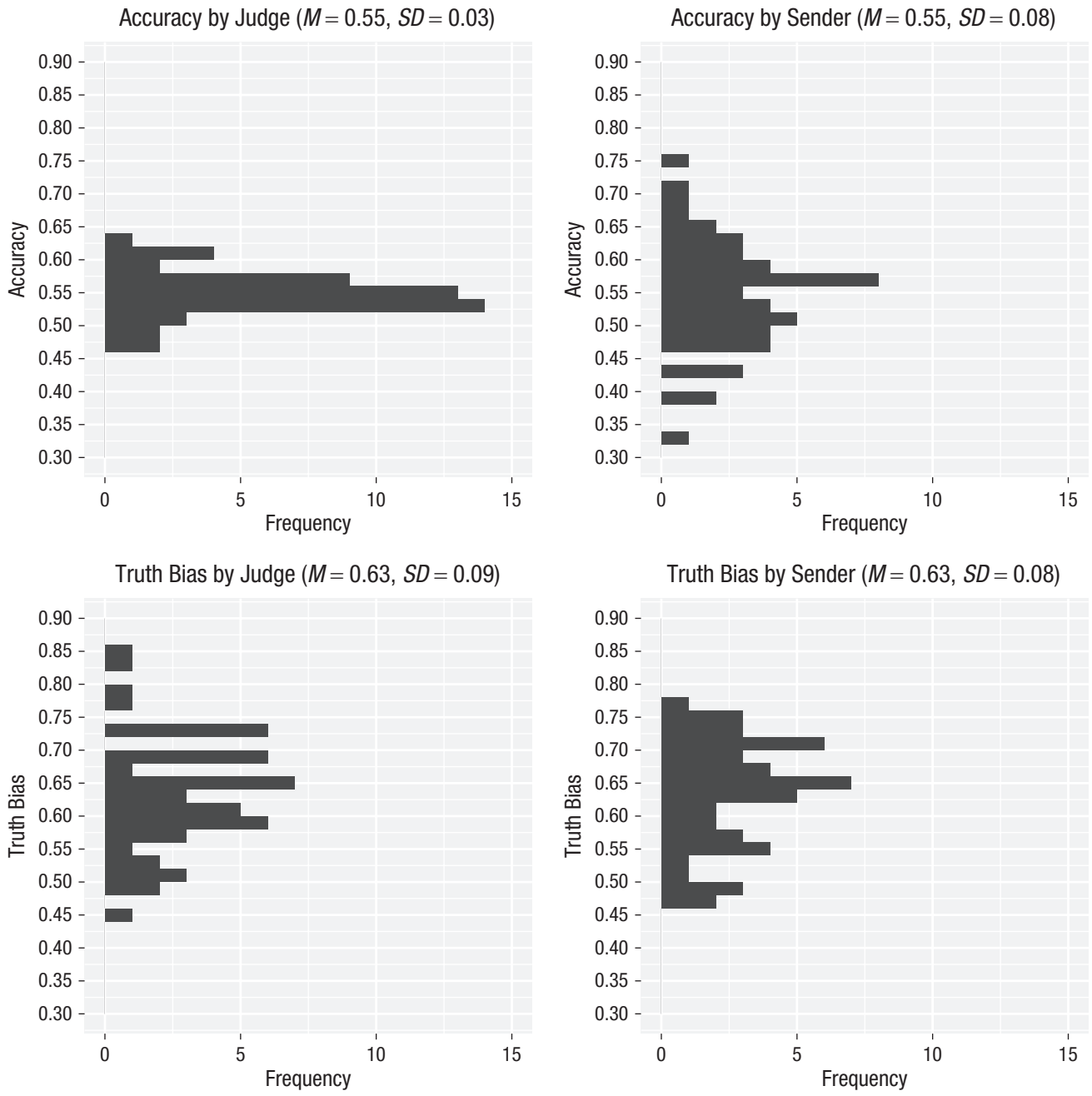

Fig. 1. Distributions of accuracy and truth bias by sender and judge.

we first extract data including only judgments and accuracies of Judges 1, 2, and 3 and of Senders 4, 5, and 6. Then, we randomly extract one judgment and accuracy for Judge 1 with Sender 4, Judge 1 with Sender 5, . . . , Judge 3 with Sender 6. As a result, we can get 3 × 3 (i.e., Number of Judges × Number of Senders) judgments and accuracies. Finally, we calculate averages across the nine judgments (i.e., truth bias) and across the nine accuracy scores.

We repeated these steps, using what we call a "loop," 10,000 times. As a result, we obtained judgments (i.e., truth-bias) and accuracy rates of 10,000 simulated experiments involving a specific number of judges and senders.

Because of the round-robin design of Masip et al. (2020), however, actual implementation of the simulation was slightly modified from the general procedure above. In round-robin designs, a judge can also be a

sender and vice versa. Because of this, judge ID and sender ID can be the same when we randomly extract them simultaneously. When we filter our data by the same judge ID and sender ID, we find no data: Judges in Masip et al.'s experiment did not make veracity judgments of their own messages. This required us to select judge and sender sequentially: We first randomly selected judges and then randomly selected senders for each judge excluding the judge. This procedure does not purely mirror actual experiments because the variable chosen later (i.e., senders) has more randomness than the one chosen earlier (i.e., judges), which will decrease the heterogeneity of the senders. To avoid this problem, we prepared two kinds of loops: one choosing judges first and the other one choosing senders first. We obtained our results by mixing the results of the two kinds of loops. Our reported results included both 10,000 judge-first loops and 10,000 sender-first loops.

Results

Simulation experiment varying the number of judgments

First, we provide a simple demonstration of the importance of total judgments. Figure 2 shows simulations involving 30 or 100 judges with one, two, six, 12, or 100 senders. The dashed lines in Figure 2 indicate the mean ± 10; that is, 55% ± 10% for accuracy and 63% ± 10% for truth bias. Through the simulations, we used these ranges as criteria for stabilization—in other words, as a stability corridor (e.g., Kleinberg et al., 2019). Although it is quite obvious by formula, the results stabilize as a function of total judgments. Adding either more judges or more senders, we can obtain more stabilized results (for additional simulations, see Appendix A, Fig. A1). This is in line with our first prediction above, with Bond and DePaulo's (2006) outcomes, and with statistical theory.

Next, we explore the less obvious issue of whether the number of judges and senders is interchangeable. Recall that our second prediction was drawn from the previous research showing that senders have more variability than judges.

Simulation experiments varying the number of judges and senders (1–50)

To investigate the effect of increasing the number of judges and senders, we first varied the number of judges while holding the number of senders constant, and then we varied the number of senders while holding the number of judges constant. Figures 3 and 4 show funnel-shaped plots for accuracy and truth bias. As expected, findings became more stable as the number of either judges or senders increased. Extreme outliers were observed in the one-judge or one-sender condition, but the boxes and whiskers rapidly shrank and became stable as additional judges or senders were added.

More importantly, the accuracy when changing the number of senders (Fig. 4, top) was more variable than the accuracy when changing the number of judges (Fig. 3, top), and this trend was more pronounced with small numbers of judges and senders. On the other hand, the mean and median accuracy hardly changed at all across all conditions. This suggests that even with 50 judges, we sometimes observed substantial deviations from the slightly better than chance accuracy with a small number of senders.

Simulation experiments fixing total judgments

To see whether increasing the number of senders has a greater impact on the stability of accuracy scores than increasing the number of judges, we conducted simulations that varied both the number of judges and senders while holding the total number of judgments constant. Figure 5 shows the results of the simulations for accuracy with 100, 200, 500, and 1,000 total judgments. With 100 and 200 total judgments, the results never stabilized with any number of judges and senders. This was mostly because the total number of judgments is too small. Researchers should refrain from conducting experiments with fewer than 200 total judgments.

Note the hourglass shape of the plots. This is because the number of senders and the number of judges both matter. There is much more variability with small numbers of either senders or judges. The lack of symmetry in the hourglass shows the greater variability in senders. With 500 and 1,000 total judgments, the results indicated that the number of senders is more important, and simulations with extremely small number of senders had more variance that those with extremely small number of judges. For example, variability for one sender being judged by 500 judges was much larger (SD = .062) than variability for 500 senders being judged by one judge (SD = .032), variability for two senders being judged by 250 judges (SD = .047) was much larger than variability for 250 senders being judged by two judges (SD = .028), and so forth (see Appendix A, Table A1). Because both the number of senders and the number of judges matter, the differences become smaller toward the center of the plot. Of course, accuracy never stabilizes with extremely small numbers of judges and senders. Instead, a moderate number of judges and senders stabilized the results well. This conclusion is very reasonable because even

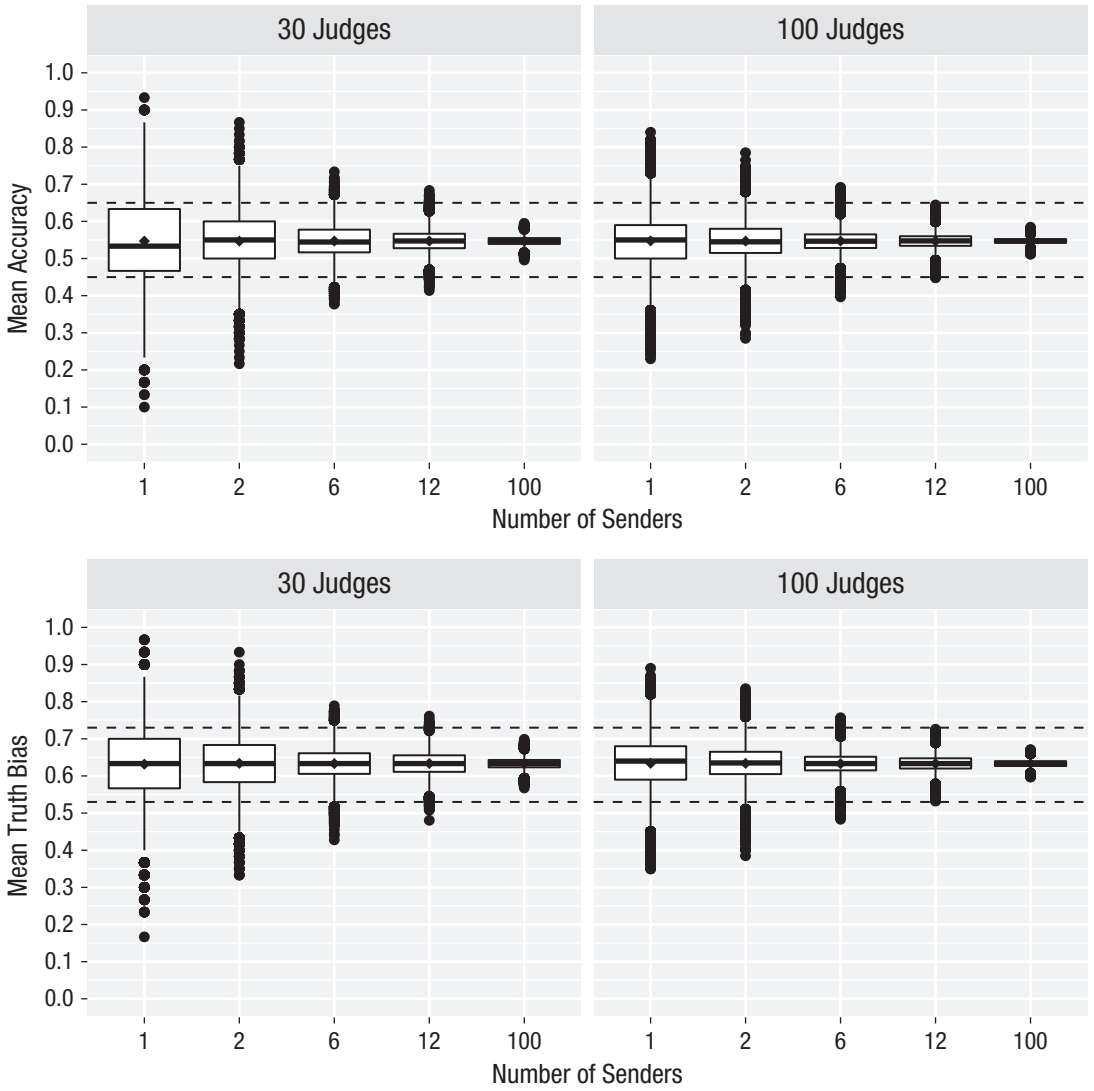

Fig. 2. A demonstration for the relationship among the number of judges, senders, and total judgments. Dashed lines indicate 0.1 above and below the mean. The boxes represent the interquartile range (IQR), and the heavy line in each box represents the median. The diamond represents the mean. The whiskers represent 1.5 times the IQR, and the dots represent outliers.

a moderate number of judges and senders can reduce both sampling error and random response error as long as there are multiple senders per judge.

We did not see clear trends in the simulation results for truth bias as we did for accuracy. For example, we did not observe a big difference between the boxplots of judges (Fig. 3, bottom) and senders (Fig. 4, bottom). Moreover, Figure 6 seems to indicate that the number of judges is more important than the number of senders. This is mostly because judges had more variance than senders for truth bias in the data we used (Fig. 1). The simulation results clearly reflect the structure of the data. Considering that senders have more variance in believability than judges (Bond & DePaulo, 2008), senders may be more important for truth bias, too. But this finding was not replicated in our data.

Conclusions, Implications, and Suggestions

A meta-analysis has shown little variability in accuracy at the level of individual experiments as long as the primary experiments are based on a sufficient amount of data (Bond & DePaulo, 2006). As explained in the introduction, both the prior meta-analyses of deception-detection research and statistical arguments suggest that the stability (or lack thereof) of mean accuracy rates in judging deception depends more strongly on

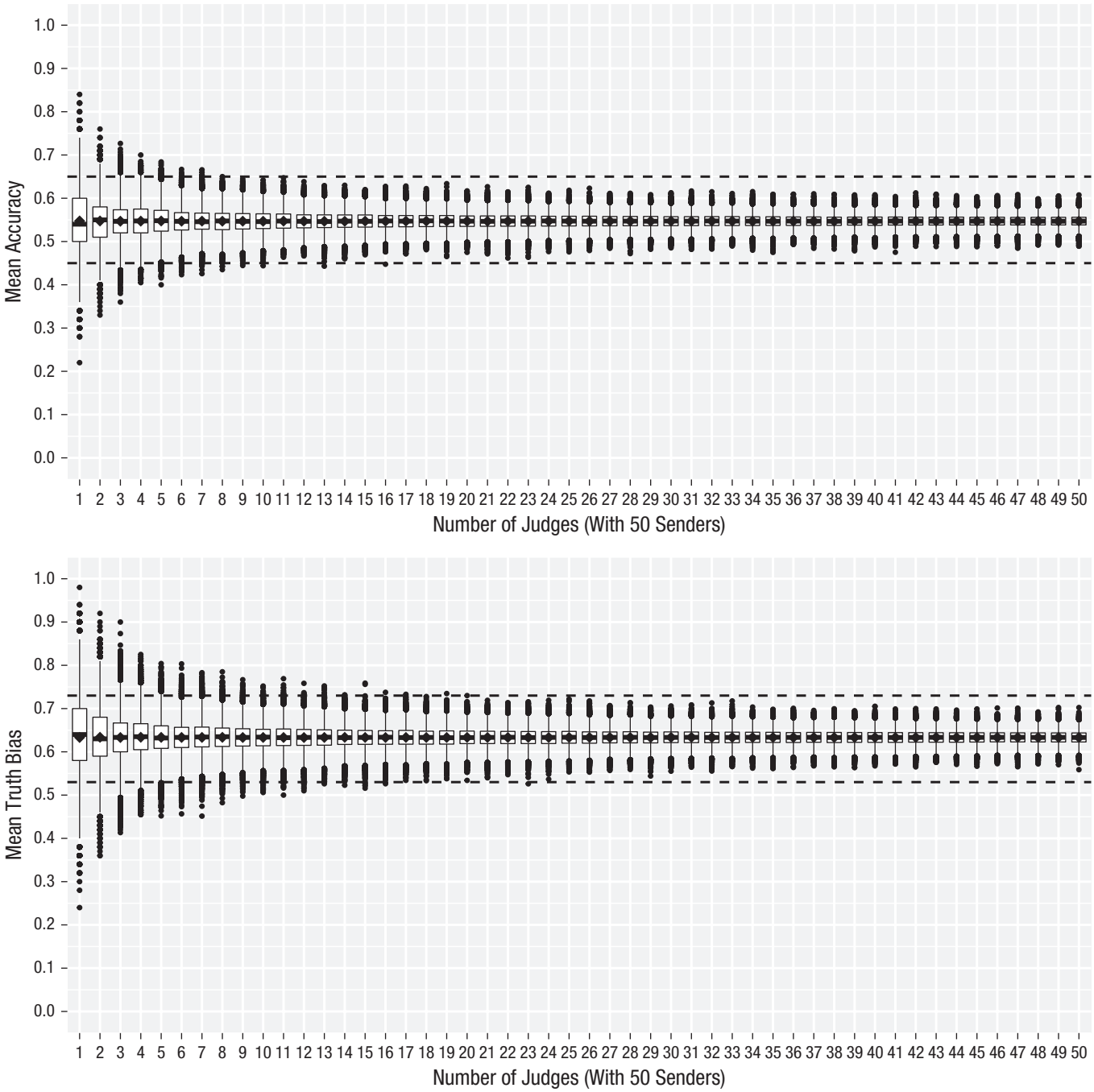

Fig. 3. Box plots for accuracy (top) and truth bias (bottom) in the simulations when we increased the number of judges progressively from one to 50 with 50 senders. Dashed lines indicate 0.1 above and below the mean. The boxes represent the interquartile range (IQR), and the heavy line in each box represents the median. The diamond represents the mean. The whiskers represent 1.5 times the IQR, and the dots represent outliers.

the number of judgments than on the number of judges (sample size). We conducted a series of simulation experiments demonstrating this to be the case. Having only one or two judgments per judge, regardless of the number of senders, can produce unstable results that may not be replicable.

Meta-analysis also demonstrated little variability among judges, especially in accuracy (Bond & DePaulo, 2008). The current data show how trends at the level of experiments apply at the level of individual judges. With a sufficient number of judgments per judge, all judges score similarly.

In contrast, meta-analysis has also found large variance among senders (Bond & DePaulo, 2008). This fact is potentially problematic for researchers because the larger variability in senders generates random response

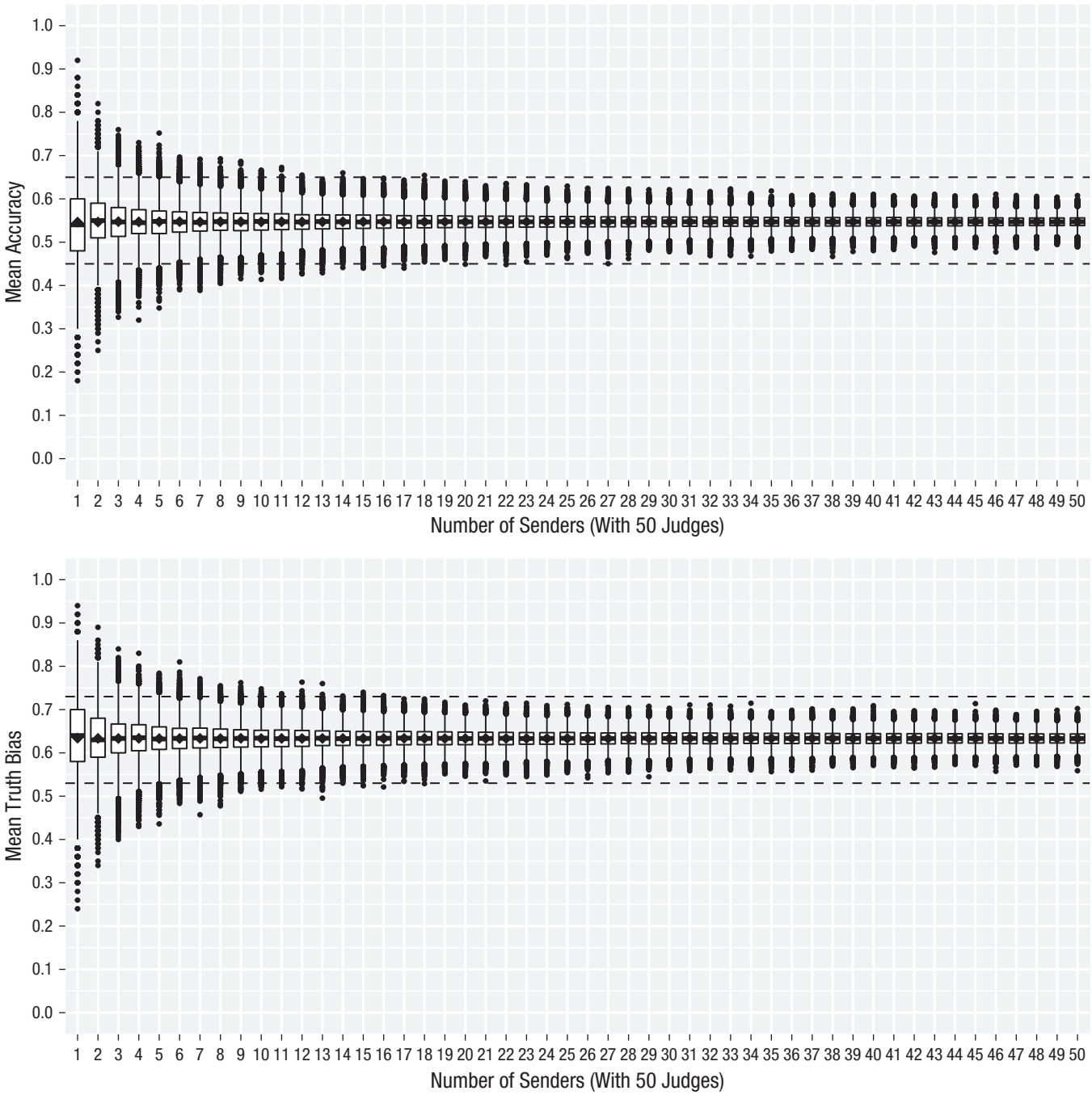

Fig. 4. Box plots for accuracy (top) and truth bias (bottom) in the simulations when we increased the number of senders progressively from one to 50 with 50 judges. Dashed lines indicate 0.1 above and below the mean. The boxes represent the interquartile range (IQR), and the heavy line in each box represents the median. The diamond represents the mean. The whiskers represent 1.5 times the IQR, and the dots represent outliers.

errors and presents issues of content validity (sampling adequacy of stimulus materials). Experiments with few senders, therefore, may yield findings that are not replicable regardless of the number of judges or judgments. Deception researchers are therefore encouraged to be more mindful about both the number of senders and the number of judgments.

The current results have several important implications for both research design and data interpretation. Concerning research design, experiments in which the dependent measure is human truth/lie discrimination (i.e., deception-detection accuracy) need to be based on a sufficient number of judgments. We are reluctant to provide a hard rule of thumb, but we nevertheless offer

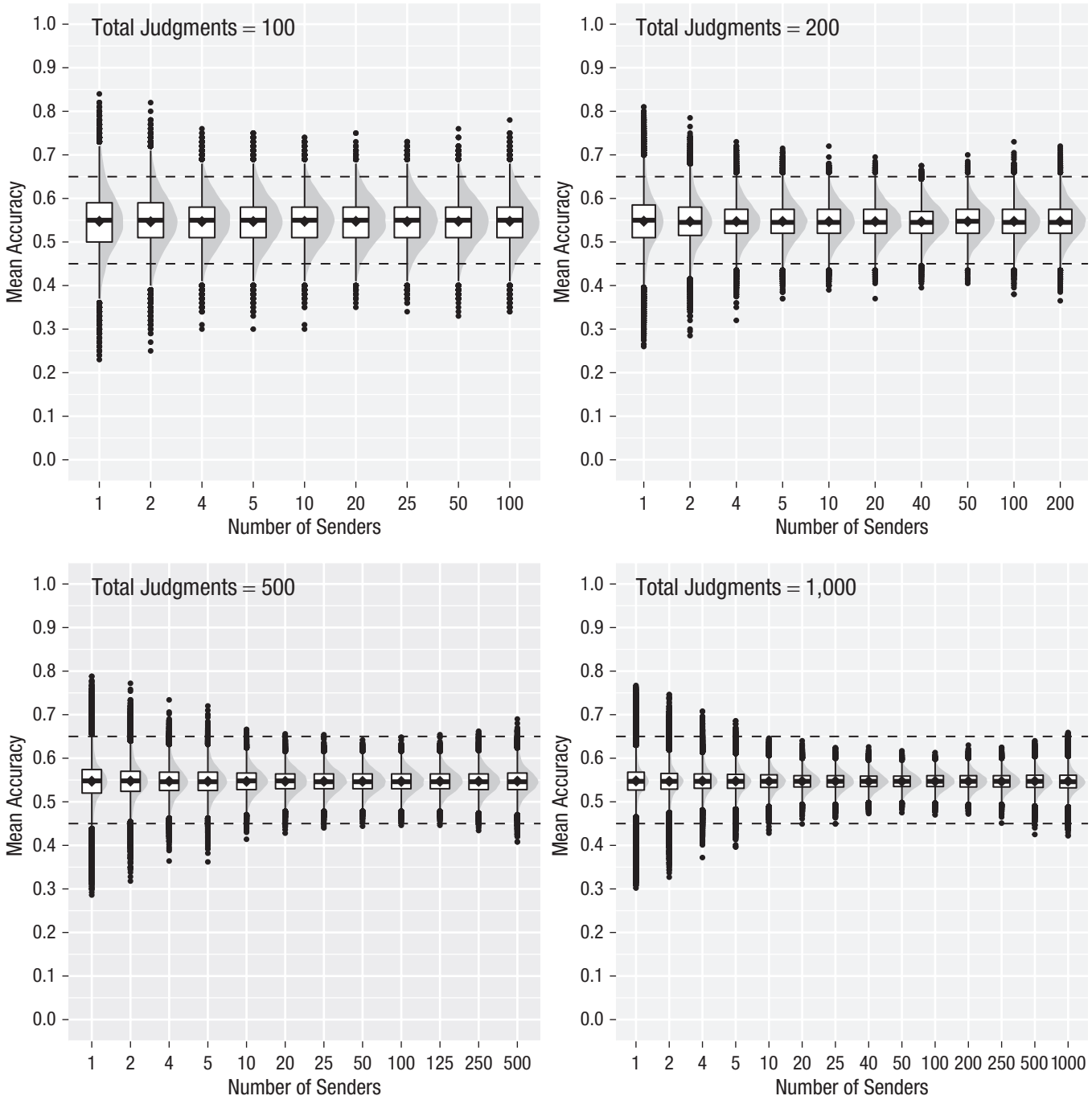

Fig. 5. Box plots for accuracy. Dashed lines indicate 0.1 above and below the mean. The boxes represent the interquartile range (IQR), and the heavy line in each box represents the median. The diamond represents the mean. The whiskers represent 1.5 times the IQR, and the dots represent outliers. The gray areas indicate distributions.

several data-informed suggestions. First, designs with a single judgment per judge should be avoided if at all possible. Results fail to stabilize even at $N = 1{,}000$. Second, increasing the number of judgments per judge provides a more efficient increase in stability than adding additional judges. Third, total number of judgments is important. Fewer than 200 judgments is clearly problematic. At least 500 total judgments are advisable. Thus, if the sample size is 50 judges, at least 10 judgments per judge is recommended. Extremely low numbers of either judges or senders need to be avoided even with 500 or more total judgments—it is apparent in Figure 5 that variability is smallest toward the center of the plots. Fourth, there is a point of diminishing returns in

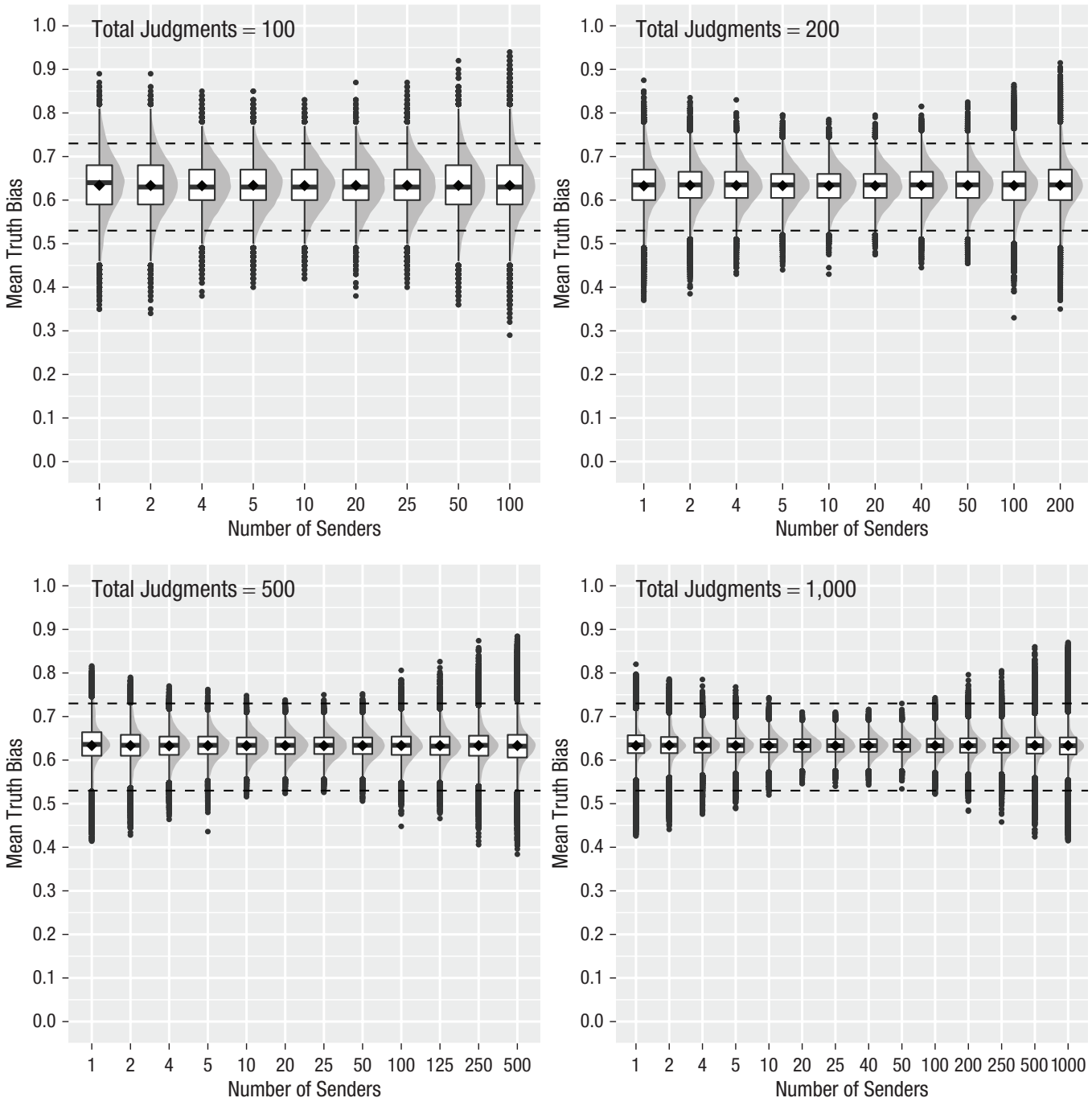

Fig. 6. Box plots for truth bias. Dashed lines indicate 0.1 above and below the mean. The boxes represent the interquartile range (IQR), and the heavy line in each box represents the median. The diamond represents the mean. The whiskers represent 1.5 times the IQR, and the dots represent outliers. The gray areas indicate distributions.

which additional judges, judgments, or senders provide little gain in data stability. Up to that point, more is better. The fewer judgments there are, the more is gained from adding more judgments. If one wants to keep the deviance within 55% ± 10%, we recommend preparing at least 20 senders with 50 judges (see Fig. 4). Of course, researchers should continue to care about statistical power, too, which depends on estimated effect size.

For data interpretation, the implications are that the number of judgments and the number of senders are critical considerations. Findings outside the 55% ± 10% range do not inspire confidence if based on relatively fewer judgments (less than 500) or just a few senders. Researchers should be wary about nonnormative results based on few judgments because publishing such findings without consideration may provide a

distorted picture of the science of deception (e.g., Luke, 2019).

We caution readers that although 54% accuracy in deception detection experiments is a robust finding within the confines of conventional experimental design in the literature, it does not follow that people are therefore poor lie detectors, only that people perform poorly under conventional experimental conditions. It has been argued that the experimental designs that produce 54% do not capture how people detect lies outside the lab (Park et al., 2002). Levine (2015) reviewed evidence for improved lie detection, and Levine (2018) provided an informative discussion of the interplay between design features in various ecologies relevant to applied lie detection.

Finally, researchers from other areas of psychology should also examine the separate effects of sampling error and random response error on the stability of the estimates. There is widespread concern in psychological science about statistical power and sample size. However, the importance of the number of stimuli being presented to experimental participants is less well understood and, probably, underappreciated. Few stimuli make it impossible to obtain sufficient statistical power even with a large sample size (Judd et al., 2012). Moreover, random response error inflates Type I error dramatically without appropriate analysis using mixed models for crossed random effects (Baayen et al., 2008; Judd et al., 2012). Researchers should be conscious about its effect for replicable science.

Appendix A

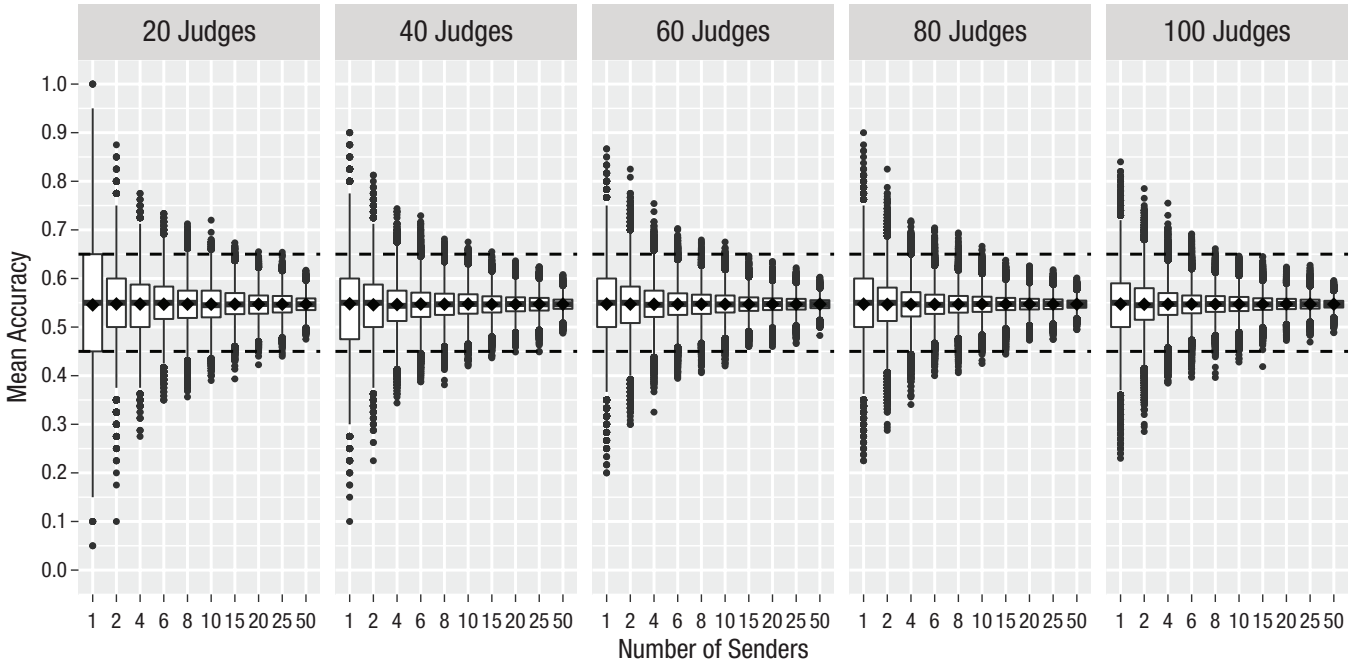

Fig. A1. *(continued on next page)*

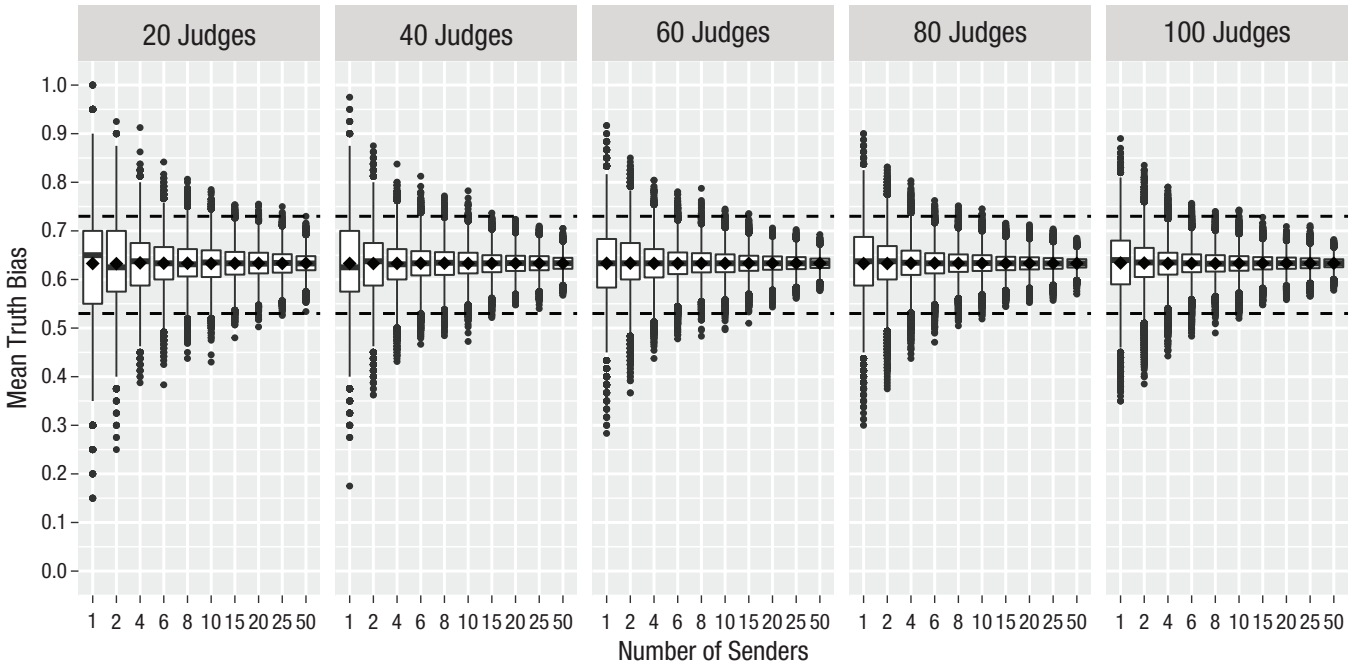

Fig. A1. Simulation results varying number of judges and senders. Dashed lines indicate 0.1 above and below the mean. The boxes represent the interquartile range (IQR), and the heavy line in each box represents the median. The diamond represents the mean. The whiskers represent 1.5 times the IQR, and the dots represent outliers.

Table A1. Mean Accuracy with Standard Deviations in Analyses Holding the Total Number of Judgments Constant and With Reversals of the Number of Judges and the Number of Senders

Total judgments	Number of judges and senders	More judges than senders		More senders than judges		Difference between *M*s	Difference between *SD*s
		M	*SD*	*M*	*SD*		
100	1 × 100	.547	.077	.547	.055	.000	.022
	2 × 50	.546	.064	.547	.053	–.001	.011
	4 × 25	.547	.057	.547	.052	.000	.005
	5 × 20	.547	.056	.547	.052	.000	.004
200	1 × 200	.548	.067	.547	.042	.001	.025
	2 × 100	.547	.054	.547	.039	.000	.014
	4 × 50	.547	.045	.547	.038	.000	.007
	5 × 40	.547	.043	.547	.037	.000	.006
	10 × 20	.547	.040	.547	.038	.000	.002
500	1 × 500	.547	.062	.547	.032	.000	.030
	2 × 250	.548	.047	.547	.028	.001	.019
	4 × 125	.547	.037	.547	.026	.000	.011
	5 × 100	.547	.034	.547	.025	.000	.009
	10 × 50	.547	.029	.547	.025	.000	.004
	20 × 25	.547	.026	.547	.026	.000	.000
1,000	1 × 1,000	.548	.060	.547	.028	.001	.032
	2 × 500	.547	.044	.547	.023	.000	.021
	4 × 250	.548	.033	.547	.020	.001	.013
	5 × 200	.547	.030	.547	.019	.000	.011
	10 × 100	.547	.024	.547	.018	.000	.006
	20 × 50	.547	.021	.547	.018	.000	.003
	25 × 40	.547	.020	.547	.019	.000	.001

Note: The values on either side of the multiplication symbols can represent either senders or judges. For instance, 1 × 100 indicates that we extracted 100 judges and 1 sender ("more judges than senders") or 1 judge and 100 senders ("More senders than judges"). The two final columns display the differences between the means and the standard deviations displayed in the preceding columns, and show the effect of having more senders than judges (compared with having more judges than senders) on the means and standard deviations.

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 iDs

Yasuhiro Daiku https://orcid.org/0000-0002-9816-4219
Jaume Masip https://orcid.org/0000-0002-7783-9547

References

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. *Forensic Examiner*, *15*, 6–11.

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

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*, 214–234. https://doi.org/10.1207/s15327957pspr1003_2

Bond, C. F., Jr., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, *134*, 477–492. https://doi.org/10.1037/0033-2909.134.4.477

Burgoon, J. K., Buller, D. B., Ebesu, A. S., & Rockwell, P. (1994). Interpersonal deception: V. Accuracy in deception detection. *Communication Monographs*, *61*, 303–325. https://doi.org/10.1080/03637759409376340

DePaulo, B. M., & Pfeifer, R. L. (1986). On-the-job experience and skill at detecting deception. *Journal of Applied Social Psychology*, *16*, 249–267. https://doi.org/10.1111/j.1559-1816.1986.tb01138.x

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. https://doi.org/10.1037/a0028347

Kleinberg, B., Arntz, A., & Verschuere, B. (2019). Being accurate about accuracy in verbal deception detection. *PLOS ONE*, *14*(8), Article e0220228. https://doi.org/10.1371/journal.pone.0220228

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160. https://doi.org/10.1007/BF02288391

Levine, T. R. (2015). New and improved accuracy findings in deception detection research. *Current Opinion in Psychology*, *6*, 1–5. doi.org/10.1016/j.copsyc.2015.03.003

Levine, T. R. (2016). Examining sender and judge variability in honesty assessments and deception detection accuracy: Evidence for a transparent liar but no evidence of deception-general ability. *Communication Research Reports*, *33*, 188–194. https://doi.org/10.1080/08824096.2016.1186629

Levine, T. R. (2018). Ecological validity and deception detection research design. *Communication Methods and Measures*, *12*, 45–54. https://doi.org/10.1080/19312458.2017.1411471

Levine, T. R. (2020). *Duped: Truth-default theory and the social science of lying and deception*. University of Alabama Press.

Levine, T. R., Clare, D., Blair, J. P., McCornack, S. A., Morrison, K., & Park, H. S. (2014). Expertise in deception detection involves actively prompting diagnostic information rather than passive behavioral observation. *Human Communication Research*, *40*, 442–462. https://doi.org/10.1111/hcre.12032

Levine, T. R., Feeley, T. H., McCornack, S. A., Hughes, M., & Harms, C. M. (2005). Testing the effects of nonverbal behavior training on accuracy in deception detection with the inclusion of a bogus training control group. *Western Journal of Communication*, *69*, 203–217. https://doi.org/10.1080/10570310500202355

Levine, T. R., & Weber, R. (2020). Unresolved heterogeneity in meta-analysis: Combined construct invalidity, confounding, and other challenges to understanding mean effect sizes. *Human Communication Research*, *46*, 345–354. https://doi.org/10.1093/hcr/hqz019

Luke, T. J. (2019). Lessons from Pinocchio: Cues to deception may be highly exaggerated. *Perspectives on Psychological Science*, *14*, 646–671. https://doi.org/10.1177/1745691619838258

Masip, J., Levine, T. R., Somastre, S., & Herrero, C. (2020). Teaching students about sender and receiver variability in lie detection. *Teaching of Psychology*, *47*, 84–91. https://doi.org/10.1177/0098628319888116

Park, H. S., Levine, T. R., McCornack, S. A., Morrison, K., & Ferrara, M. (2002). How people really detect lies. *Communication Monographs*, *69*, 144–157. https://doi.org/10.1080/714041710

Vernham, Z., Vrij, A., Mann, S., Leal, S., & Hillman, J. (2014). Collective interviewing: Eliciting cues to deceit using a turn-taking approach. *Psychology, Public Policy, and Law*, *20*, 309–324. https://doi.org/10.1037/law0000015

Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, *22*, 1–21. https://doi.org/10.1111/lcrp.12088