Comparing the accuracy of experimental estimates to guessing: a new perspective on replication and the “Crisis of Confidence” in psychology

Clintin P. Davis-Stober · Jason Dana

Published online: 10 May 2013

Abstract We develop a general measure of estimation accuracy for fundamental research designs, called *v*. The *v* measure compares the estimation accuracy of the ubiquitous ordinary least squares (OLS) estimator, which includes sample means as a special case, with a benchmark estimator that randomizes the direction of treatment effects. For sample and effect sizes common to experimental psychology, *v* suggests that OLS produces estimates that are insufficiently accurate for the type of hypotheses being tested. We demonstrate how *v* can be used to determine sample sizes to obtain minimum acceptable estimation accuracy. Software for calculating *v* is included as online supplemental material (R Core Team, 2012).

Keywords Estimation accuracy · Replicability · Statistical power · Improper linear models

Introduction

Research in psychology is currently facing a *quantitative* crisis. Peer-reviewed journals are rife with contradictory findings, potentially attributable to underpowered studies that result in spurious rejection (or acceptance) of hypotheses (e.g., Ioannidis, 2005, 2008; Maxwell, 2000, 2004). Findings of precognition and premonition that have achieved statistical significance and survived peer review (Bem, 2011) have raised the controversial question of when some studies should be believed, while others not. Failures to replicate well-known results have created what some researchers are calling a “crisis of confidence” (Pashler & Wagenmakers, 2012). When evaluating research, how do we know whether our findings are accurate and genuine?

We offer a new perspective on this “crisis” that does not depend upon unreported researcher activities, such as publication bias (Francis, 2012a, b; Ioannidis, 2008) or researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011). Nor does our approach concern itself with null hypothesis testing or traditional methods of statistical inference. Rather, we examine the accuracy of parameter estimation for fundamental research designs in psychology. We study estimation accuracy because of its pertinence to the following question: Over repetitions of an experiment, how good are the estimates and how much do they vary? Satisfactory estimation accuracy is a prepotent problem. If one’s estimates are inaccurate under reported conditions, it is a moot question whether researcher degrees of freedom have been used or publication bias has occurred. Conversely, if a study is free of researcher degrees of freedom and publication bias, it is still meaningless if estimation accuracy is poor. Put differently, the quality and value of our experimental conclusions hinge upon how accurately population values of interest, such as true means, can be estimated.

We demonstrate that for many areas of psychology, study conditions are such that standard estimation methods, such as sample means and regression coefficients, have unacceptably poor accuracy. We compare these methods against a benchmark estimator we call *random least squares* (RLS). RLS determines the relative values of its estimates at random. That is, RLS yields literally random conclusions about whether data show treatment effects. Yet, for sample and effect sizes common to psychology, we show that RLS estimates the population values of interest more accurately than do sample means or linear regression coefficients. We present a measure called *v* that tracks how often a researcher

Electronic supplementary material The online version of this article (doi:10.3758/s13428-013-0342-1) contains supplementary material, which is available to authorized users.

C. P. Davis-Stober (✉)
Department of Psychological Sciences, University of Missouri,
Columbia, MO 65211, USA
e-mail: stoberc@missouri.edu

J. Dana
University of Pennsylvania, Philadelphia, PA, USA

can expect standard estimation methods to be more accurate than RLS. Under one interpretation, v is the probability of standard methods being more accurate than RLS. We argue that studies undertaken with low v cannot meaningfully establish patterns of treatment effects and should not be the basis of theoretical conclusions.

To underscore the relevance of these conclusions for experimentalists, imagine a researcher in the physical sciences who routinely uses some piece of lab equipment. After decades of research, the manufacturer of that equipment informs the researcher that it is faulty and gives readings that are essentially pure noise under laboratory conditions that fall within accepted practice in the field. Surely, the researcher would immediately wonder how many of his or her findings over the years were actually true. If many other labs were also using the same faulty equipment, it would be a scandal. Yet our analysis suggests we are in exactly that position in many areas of psychology. The statistical equipment that researchers routinely use is too inaccurate under some conditions, and these conditions fall within guidelines of accepted practice. As such, nothing prevents the publication of statistically significant findings that are potentially meaningless. While the debate over whether to trust experimental results often focuses on the quality of experimental procedures, publication bias, or the meaning of a *p*-value, we show that, absent all of these issues, there remains a basic problem: Some findings should not be trusted because sample means or regression estimates are too inaccurate under the sample and effect sizes at which the experiment operates.

We focus on evaluating the estimation accuracy of the ordinary least squares (OLS) estimator, due to its ubiquity for evaluating theory in psychology. For example, whenever a researcher assigns participants to different groups and uses the group mean responses on some dependent variable to represent the population means, he or she is using OLS. Thus, the two-sample *t*-test and ANOVA procedures rely on the OLS estimator. OLS is also the default estimator for linear regression. While many accuracy metrics are possible, we compare the accuracy of OLS with that of RLS using mean squared error (*MSE*), a "gold standard" measure of estimation accuracy in the statistics literature (Lehmann & Casella, 1998). *MSE* is defined as the expected sum of squared differences between a set of p-many estimates, $\widehat{\beta}_i$, and the corresponding population values, β_i,

$$MSE_{\widehat{\beta}} = E\left(\sum_{i=1}^{p}\left(\widehat{\beta}_i - \beta_i\right)^2\right). \tag{1}$$

Put differently, *MSE* is how close, on average, an estimator is to the truth.

Whether OLS or RLS is expected to be more accurate depends on the true population values of interest, something the researcher does not know. We solve the problem of how to compare accuracy using a measure that we denote v. Given an upper bound on overall effect size, v measures the proportion of all possible population values for which OLS is more accurate than RLS. If all population values are equally likely, then v is the probability of OLS being more accurate than RLS. Thus, v ranges from 0 to 1, with higher values of v suggesting more robust accuracy. For sample and effect sizes common to experimental psychology, v suggests that OLS produces estimates that are insufficiently accurate for the type of hypotheses being tested. We demonstrate how targeting minimum standards for accuracy such as $v>.5$ (i.e., one's estimates are better than the guessing benchmark at least half the time) can prospectively determine minimum acceptable sample sizes for general experimental designs in psychology. In the following sections, we lay out the logical purpose of having an accuracy benchmark against which to compare OLS. We then describe the logic behind our choice of benchmark and how it is constructed. We offer several illustrative examples of our benchmark. Finally, we present the v measure and discuss how it can be used.

The need for an accuracy benchmark

As was noted above, *MSE* is a standard metric of estimation accuracy in the statistical literature. At the same time, *MSE* values, in and of themselves, are not illuminating for understanding the quality of psychological data because it is unclear that we have, or could have, precise standards of satisfactory *MSE* values. Clearly, less *MSE* is better because we would like our estimates to be accurate. Beyond that, it is difficult to construct anything better than an ad hoc standard for how small *MSE* need be for experimental results to be satisfactory, let alone to tailor that standard according to research questions and practices in various areas of the field.

One difficulty when determining adequate levels of accuracy is that theories in psychology are usually not sufficiently quantified to make point predictions of population values. Where the dependent measure is a Likert scale rating, it is not always clear that a point prediction is even meaningful; the mean scale response may make sense to the researcher only as a way of establishing that experimental groups respond differently. We do not view this level of quantification as a problem. Rather, it is a feature of many psychological hypotheses. Key hypotheses can be formulated simply as directional statements about effects. Examples might be "people primed with words associated with the elderly will subsequently walk slower" or "infants at six months are better at recognizing primate faces then infants at 9 months." Such directional hypotheses, if true, are highly informative. They force us to propose specific mechanisms that could explain the effects—mechanisms that could confirm and disconfirm various theories and propose new theories. Directional hypotheses are important to psychology, and indeed, they are often what researchers care about the most.

We are far less often in the position to hypothesize precise population values. We may predict, for example, that the experimentally primed group will walk more slowly than a control group, but we probably cannot hypothesize the exact mean walking speed for each group. Similarly, one might hypothesize that a regression coefficient will be different than zero with a specific sign in light of other variables in the model. Much less often, if at all, do psychological theories allow us to specify and test a priori point predictions for the coefficients. Directional hypotheses do, however, specify relative values of the parameters of interest. When one hypothesizes a directional effect—for example, that the population mean of one group will be larger than the population mean of another—it follows that the sample means of the experimental groups are predicted to show this pattern. In factorial designs, hypotheses about main effects and interactions specify how most or all of the parameters will be ordered. While psychological theories may not supply exact values, we believe that most researchers are willing to hypothesize the sign of a regression coefficient or the direction of a difference in group means. Even if that does not lend to specifying parameters in an absolute sense, it does lend itself to specifying relationships among the parameters.

The estimated relationships among population values are thus highly important for testing psychological theories, and sometimes these values are of singular interest. Yet, as was described earlier, the quality of experimental results also hinges on estimation accuracy, which is measured in an absolute sense with *MSE*. For example, we would not want to make much of the finding that one sample mean was larger than another if we somehow knew that those sample means were far from the population values. One way to address these issues is to specify benchmarks against which to compare the accuracy of OLS. Suppose that there was an alternative method for estimating population values that psychological theorists, primarily concerned with relative population values, could agree should not be more accurate than standard methods. Perhaps this alternative promises to yield different or even nonsensical relationships among the parameter estimates. If the *MSE* of this alternative method could be calculated, an absolute standard of accuracy for OLS, under a given set of experimental conditions, would be that its *MSE* should be less than the *MSE* of this benchmark. The logic is simple: If we, as specialists, agree that the benchmark is an unsatisfactory method for establishing treatment effects, it would be troubling if the benchmark were more accurate at estimating the population values of interest than our current methods.

Following the above logic, we propose a benchmark for OLS called random least squares (RLS). RLS determines certain features of the relationships among the parameter estimates at random. As we will describe in detail in the following section, RLS specifically fixes the relative signs and magnitudes of its estimates, the very relationships that determine whether and in what direction treatment effects exist, randomly without use of data. Since psychologists are primarily interested in estimating the relative population values—that is, directional treatment effects—it is problematic if there exists an estimator that scrambles this information yet gets closer to the truth, on average, than do typical estimates in the field.

Of course, if one has high-quality data because of large samples and large effects, then standard methods of estimation will be more accurate, on average, than such a benchmark. Indeed, one may have the intuition that standard methods should not lose for even small effects and sample sizes. In the following sections, we precisely develop RLS and solve for sample and effect sizes for which it is more accurate than OLS on average. Such sample and effect sizes indeed exist, and perhaps surprisingly, they are not uncommon to published research in psychology.

Random least squares

Before developing RLS, it is helpful to remind the reader that our analysis deals with estimating the standard linear model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is a fixed (nonrandom) $n \times p$ design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of population weights, and $\boldsymbol{\varepsilon} \sim (\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{\boldsymbol{n}\times\boldsymbol{n}})$ with i.i.d. sampling. Note that we do *not* assume a specific distributional form for $\boldsymbol{\varepsilon}$, such as the normal distribution. We assume that $\boldsymbol{X}$ is of full rank and that every entry of $\boldsymbol{\beta}$ is real-valued. Without loss of generality, we assume that observed values of $\boldsymbol{y}$ and all continuous predictor variables in $\boldsymbol{X}$ are standardized in *z*-score format, as in a standardized multiple regression. We assume orthogonal coding for qualitative predictors. For our one-way ANOVA examples, we assume that $\boldsymbol{X}$ is coded for qualitative variables such that OLS becomes the vector of sample means—that is, cell means coding (see Muller & Fetterman, 2003; Maxwell & Delaney, 2003).

The default estimator for the linear model is OLS, denoted $\widehat{\beta}_{OLS}$, which, using the standard matrix notation, is defined as

$$\widehat{\beta}_{OLS} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2}$$

Depending upon the design of $\boldsymbol{X}$, sample means, regression, and correlation coefficients are special cases of OLS.

Intuitively, we can think of OLS as the best way to portion out weight to the independent variables in an experiment and determine which is most important within the sample. We can also think of OLS as the best way to assign weight to different experimental conditions and determine whether one treatment is more important than another. As was explained earlier, experimental conclusions drawn from data are often, if implicitly, statements about the relationships among OLS coefficients, be they sample means or regression coefficients. That is, the direction and relative importance of experimental effects are determined by the relative signs and magnitudes of the OLS coefficients.

Our concept of a benchmark for OLS is inspired by the idea of *improper linear models* (see Dawes, 1979). Rather than best-fitting the sample, improper linear models assign weight to the independent variables according to a priori decision heuristics that the researcher chooses without the use of data. For example, a great deal of research has shown that a simple equal weighting of all independent variables often outperforms OLS on future samples (Dana & Dawes, 2004; Dawes & Corrigan, 1974; Wainer, 1976). Recent work has thoroughly analyzed the properties of improper linear models as estimates of $\boldsymbol{\beta}$ (Davis-Stober, 2011; Davis-Stober, Dana, & Budescu, 2010a, b). Specifically, one could express an a priori decision heuristic as a vector $\boldsymbol{a}$. For example, if one's heuristic is to equally weight all of p-many independent variables, that heuristic could be represented by an $\boldsymbol{a}$ vector in which all entries are equal to 1. Such a vector captures the decision heuristic's policy of equally weighting all predictors. This vector could not, however, serve as a sensible estimate of $\boldsymbol{\beta}$; indeed, the coefficients of $\boldsymbol{a}$ in this example set a scale that does not conform to the scale of the data. To solve this problem, one could multiply $\boldsymbol{a}$ by a single number k, calculated from experimental data, to rescale the values of $\boldsymbol{a}$—hence, $\boldsymbol{a}k$. Obviously, multiplication by a scalar will not change the relative values of the decision heuristic's weighting policy.

One choice of k that has been well-studied in this context is the least squares scaling factor $k = \frac{\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{y}}{(\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a})}$ (Davis-Stober, 2011; Davis-Stober et al., 2010a, b). This choice of k minimizes the squared error of $\boldsymbol{a}k$ for one's sample, subject to any fixed $\boldsymbol{a}$. To reiterate, the vector $\boldsymbol{a}$ is chosen independently of any observed data; data can be used only to determine the value k, which scales the final estimates, $\boldsymbol{a}k$, but, obviously, cannot change the relative signs and relative magnitudes that are determined by $\boldsymbol{a}$. Indeed, $\boldsymbol{a}k$ can be considered as a special case of general constrained least squares estimation (Amemiya, 1985; Chipman & Rao, 1964).

Following from this logic, we create a performance benchmark for OLS by using this procedure with $\boldsymbol{a}$ chosen *randomly*, a procedure we call RLS. More precisely, let the entries of $\boldsymbol{a}$ be sampled according to a uniform distribution over an interval and, once all of the entries of $\boldsymbol{a}$ are sampled, this randomly determined $\boldsymbol{a}$ vector is multiplied by the scalar value $k = \frac{\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{y}}{(\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a})}$. Due to our choice of scaling factor, k, the entries of $\boldsymbol{a}$ can be sampled uniformly from any symmetric interval centered at 0. The length of the $\boldsymbol{a}$ vector is factored out by k; for example, for any given $\boldsymbol{X}$ and $\boldsymbol{y}$, an $\boldsymbol{a}$ vector with all entries equal to 1 will provide precisely the same final RLS estimates as an $\boldsymbol{a}$ vector with all entries equal to 700. In other words, RLS determines two important properties at random: (1) whether any pair of coefficients agree in sign (relative signs) and (2) the ratio of any pair of coefficients (relative magnitudes). The relative values of the entries in $\boldsymbol{a}$ are the key ingredient, not the overall scale; that is set by k using experimental data. In our illustrations below, we select each entry of $\boldsymbol{a}$ according to a uniform distribution over [−10,10], but we would obtain the very same results selecting the entries of $\boldsymbol{a}$ according to a uniform distribution over [−20000, 20000]. When we consider the v measure, for mathematical convenience, we will sample $\boldsymbol{a}$ uniformly from the surface of a unit sphere of dimension p centered at the origin.

To the extent one is concerned about the relationships among population values, as with directional hypotheses, RLS should be an uncontroversial benchmark for accuracy. By randomizing the relative signs and magnitudes of its estimates, RLS randomizes conclusions about directional hypotheses—that is, the relative orderings of coefficients. The least squares scaling factor, k, which is calculated from experimental data, allows us to compare OLS and RLS on the same accuracy metrics but does not change the important properties of randomness outlined above.

Illustration of RLS

To illustrate the use of RLS, consider a researcher interested in testing a theory concerning response times of subjects in three different conditions. According to the theory, the (standardized) population mean of group 3 (the treatment condition) will be larger than that of group 1 (control condition), which will be larger than the population mean for group 2 (the reverse treatment condition)—that is, $\mu_2 < \mu_1 < \mu_3$. Typically, one would run a carefully controlled experiment and use the observed sample means (i.e., OLS) of the three groups as estimates of the three population means. Once estimated, an appropriate statistical analysis would be conducted, such as an ANOVA procedure. But suppose that this researcher decides to apply RLS instead of sample means. *Before* running the experiment, he or she needs three random numbers to form the basis of the population mean estimates for each of the three treatment groups. He or she decides to generate random numbers from a uniform distribution over [−10,10], but of course the length of the interval is unimportant. For group 1, he or she draws a random number and obtains 7.2. For group 2, he or she obtains −2.89, and for group 3, he or she obtains a value of 1.2. These values constitute his or her $\boldsymbol{a}$ vector—that is, $\boldsymbol{a}' = (7.2\ {-2.89}\ 1.2)$.

This researcher, without collecting any experimental data, already knows that his or her final RLS estimates will not support the theory he or she is testing. The data he or she collects will determine the value of k, but no value of k can change the fact that the estimate for group 3 will lie between the estimates for groups 1 and 2. Even worse, the relative magnitudes of the estimates are already determined. The estimate for group 1 will be six times as large as that for group 3 because 7.2/1.2=6. No value of the scalar k and, therefore, no amount of data will change this ratio for the final RLS estimates. Before the experiment has even been run, this researcher has already arrived at some experimental conclusions about the relative directions and relative magnitudes of the population means. As a final step, the researcher runs the experiment and obtains the following (standardized) data:

Group 1	Group 2	Group 3
1.10	−0.83	0.50
0.10	−1.07	1.70
−1.30	−0.94	1.70
0.50	−0.83	2.30
−0.10	−0.59	1.10
−0.10	−0.83	1.71
1.10	−0.83	−0.11
−1.07	−0.95	−0.50
−0.59	−0.82	−0.11
−0.22	−0.70	−0.11

The OLS estimator, assuming cell means coding, for these data is simply the vector of sample means for the three groups, $\bar{x}_1 = -.08, \bar{x}_2 = -.84$, and $\bar{x}_3 = .92$, respectively. Calculating k for these data yields k=.05. Multiplying k by the random choice of $\boldsymbol{a}$ yields RLS coefficients of .36, −.14, and .06 for groups 1–3, respectively. As was expected, the RLS estimates have scrambled some of the information contained in the sample. While the experimental data scaled the final RLS estimates through the k term, the relative agreement in sign and the relative magnitudes of the randomly determined coefficients in $\boldsymbol{a}$ are preserved. As an example, for groups 3 and 1, the ratios of the RLS estimates are equal to $\frac{.36}{.06} = \frac{7.2*.05}{1.2*.05} = 6$. Raw code for simulating comparisons of RLS with OLS is available as supplemental material, as well as a workable example.

Obviously, RLS is nonsensical science. To the extent a psychological theory is defined in terms of the relationships among population means, RLS uniformly scrambles this information for any experimental data. As the complexity of an experiment increases, yielding more parameters to be estimated, the randomizing nature of $\boldsymbol{a}$ greatly increases. For example, if there are eight parameters to estimate and RLS orders them randomly, there is little chance of getting an ordering representative of the population values. The fact that this ordering cannot be shuffled by the data starts to represent a formidable constraint. Yet, as we demonstrate, under sample and effect sizes common to many areas in psychology, this researcher using RLS will, on average, be more absolutely accurate at estimating the population means than will be a researcher using sample means or, more generally, OLS.

The reader may wonder how, under any circumstances, OLS could incur greater *MSE* than RLS. Linearly constrained estimators, of which RLS is a special case, have some favorable properties. While biased, meaning that their average does not equal the true value of $\boldsymbol{\beta}$, linearly constrained estimators have less variance than does OLS (Toro-Vizcarrondo & Wallace, 1968) and, thus, can outperform OLS given limited sample and effect sizes (e.g., Teräsvirta, 1983). OLS is quite sensitive to the sample on which it is estimated and, for that reason, can be erratic in future samples due to factors such as measurement and prediction error. In other words, OLS often overfits the sample data. Dana (2008) described how improper linear models like RLS approximate shrinkage estimators that are conservatively biased toward no effect. The researcher might rule out the possibility that the true effects are very large a priori; that is, the population value of R^2 is unlikely to be very large. With such priors, bias can lead to more efficient estimation. Indeed, if an upper bound can be placed on the value of R^2, the accuracy of such estimates can always be improved by biasing them toward no effect (Eldar, Ben-Tal, & Nimirovski, 2005).

From a geometric perspective, RLS can be described as first randomly selecting a direction, the $\boldsymbol{a}$ vector. The final RLS estimates will remain on the vector $\boldsymbol{a}$. Data can impact the RLS estimates only via the scalar k. This scalar affects only the length of the final RLS vector, $\boldsymbol{a}k$, which will tend to bias the RLS estimates toward "no effect." Negative values of k are possible, which will flip all of the coefficient signs in $\boldsymbol{a}$ but, because k is a scalar, cannot selectively change them. In this way, we are comparing an estimator, OLS, which estimates *both* the direction and length of $\boldsymbol{\beta}$ against an estimator that can *only* estimate the length of $\boldsymbol{\beta}$. In the next section, we solve for the conditions under which OLS incurs less *MSE* than does RLS via a measure we call v.

The *v* measure

A practical problem for comparing the *MSE*s of OLS and RLS is that *MSE* depends on $\boldsymbol{\beta}$ itself, which is an unknown set of population parameters (indeed, if we had this information, there would be no need for data; we would know the truth). Davis-Stober (2011) provided a solution to this problem by considering all possible values of $\boldsymbol{\beta}$ and deriving the proportion that favor OLS over a least squares projection onto a fixed choice of $\boldsymbol{a}$ in terms of *MSE*. We use those results here to compare OLS with RLS, defining v as the the proportion of population $\boldsymbol{\beta}$ favoring OLS over RLS. Thus, v will range in value from 0 to 1, with values closer to 1 indicating more robust accuracy of OLS. If the researcher is willing to assume, in a Bayesian fashion, that all possible $\boldsymbol{\beta}$ are equally likely, then v is the *probability* that OLS is more accurate than RLS (see Davis-Stober, 2011). We argue that $v > .5$ is a minimum standard for estimation accuracy. If one's estimates are less accurate than our guessing benchmark more than half of the time, there is little point in using them to establish treatment effects. As low as this hurdle may seem, we show that $v < .5$, or even $v = 0$, can happen surprisingly often, particularly when researching effect sizes conventionally categorized as small and medium (Cohen, 1988).

It is helpful to consider v geometrically. Consider the case of multiple regression with two predictors. Suppose the total R^2 for this model can be no larger than .16. In other words, we know, a priori, that our predictors can, at most, provide a "medium" effect by the Cohen (1988) conventions. Since we assumed that all dependent and continuous predictor variables have been standardized (z-transformed), there are algebraic constraints on how large the true regression parameters, β_1 and β_2, can be. If the predictors are uncorrelated, the set of all possible population parameters will form the interior of a circle centered at the origin with radius equal to $R = .4$. For any given pair of true regression parameters (any point within this circle), we can directly compare RLS and OLS in terms of *MSE*.

Suppose, as in Fig. 1, that the random choice of $\boldsymbol{a}$ for RLS yielded $\boldsymbol{a} = (1, 1)$. If the population values of the regression parameters just happened to be nearly equal in magnitude and sign, one would expect this choice of RLS estimator to be more accurate than OLS, depending upon the sample size. On the other hand, if the population values were equal in magnitude but had opposite signs, this choice of RLS estimator would likely be less accurate than OLS. As an illustration, the red region in Fig. 1 displays the set of population $\boldsymbol{\beta}$ such that RLS incurs less *MSE* for this random choice of $\boldsymbol{a}$. The v measure is nothing more than the proportion of the interior of the circle that is blue. As sample size increases, the red region will become narrower; hence, the blue region will become larger. When $v = 0$, the entire disk is red and it doesn't matter what the true population values are, RLS will always be more accurate, in expectation, than OLS. This would occur when sample and or effect sizes are quite small, yet such conditions still yield positive power values. Conversely, when $v = 1$, the entire disk is blue, and OLS will always incur less *MSE* than will RLS. For more than two predictors, v is the volume of the interior of the (hyper)sphere comprising all population values, such that OLS incurs less *MSE* than RLS (see the Appendix and Davis-Stober, 2011, for a more general discussion).

One unintuitive result is helpful for calculating v: When all independent variables are orthogonal, as in a balanced t-test, balanced one-way ANOVA, or multiple regression with uncorrelated predictors, then v does not depend on the outcome of the random choice of $\boldsymbol{a}$ (Davis-Stober, 2011). That is, no matter what random guess goes into RLS, v will always be the same in the orthogonal case. Returning to Fig. 1, different choices of $\boldsymbol{a}$ will orient the red region in a different direction but, ceteris paribus, will leave its area and shape unchanged (Davis-Stober, 2011).

When the independent variables are not orthogonal, good estimates of v can be obtained via a Monte Carlo sampling algorithm. However, the orthogonal result is particularly useful when one considers that OLS estimates will not be affected by inaccuracies due to intercorrelations among predictors (e.g., Kutner, Nachtsheim, Neter, & Li, 2004). In fact, under our assumptions of homoskedasticity and uncorrelated errors, OLS achieves minimum variance among all unbiased estimators by the Gauss–Markov theorem (e.g., Bickel & Doksum, 2001). The orthogonal case is thus an optimistic scenario for OLS in terms of v, and one in which v can be calculated directly:

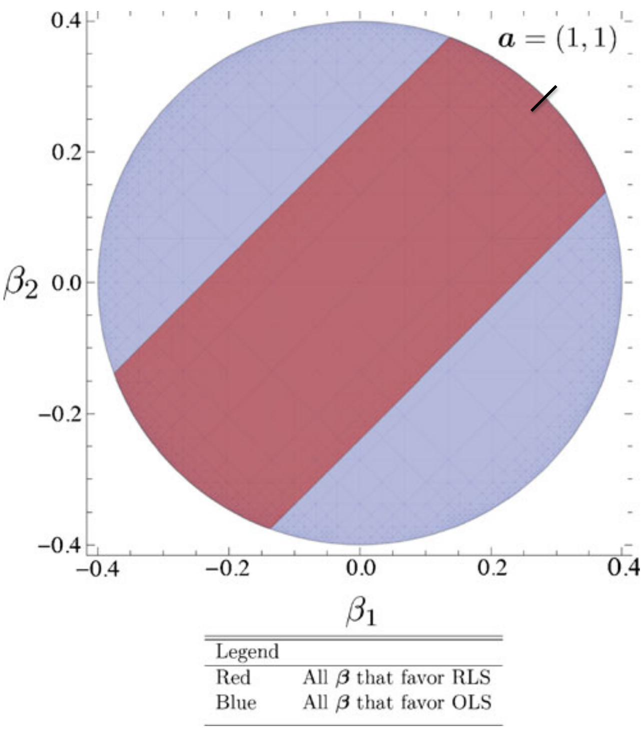

Fig. 1 Illustration of the geometry underlying the v argument. Assume $R^2 \leq .16$ and $p = 2$. The circle represents the set of all possible true $\boldsymbol{\beta}$ values. Let the single random draw of $\boldsymbol{a}$ under the RLS estimator be equal to $\boldsymbol{a} = (1,1)$. Then, assuming a fixed sample size, the red region corresponds to the set of all population $\boldsymbol{\beta}$ in which RLS incurs less *MSE* than does OLS for this particular random choice of $\boldsymbol{a}$. Likewise, the blue region is the set of all $\boldsymbol{\beta}$ in which OLS incurs less *MSE* than does RLS for this choice of $\boldsymbol{a}$. The v measure is the proportion of the circle that is blue

$$v = \frac{2\cos(\alpha)\Gamma\left(\frac{p+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{p+1}{2}\right)}\left[{}_2F_1\left(\frac{1}{2},\frac{1-p}{2},\frac{3}{2},\cos^2(\alpha)\right) - \sin(\alpha)^{p-1}\right], \quad (3)$$

where $\alpha = \cos^{-1}\left(\frac{1-\zeta}{\sqrt{1-2\zeta(1-\zeta)}}\right), \zeta = \frac{\gamma-\sqrt{\gamma-\gamma^2}}{2\gamma-1}, \gamma = \min\left\{\frac{(p-1)\left(1-R^2\right)}{(n-p)R^2}, 1\right\}$, ${}_2F_1(\cdot,\cdot,\cdot,\cdot)$ is the Gaussian hypergeometric function and $\Gamma(\cdot)$ is the gamma function. Equation 3 does not provide much intuition, but it is a closed-form, analytic solution and is easily calculated using R code (R Core Team, 2012) available as supplemental material to this article. Furthermore, *v* is a function of just three quantities with intuitive importance: total sample size (*n*), number of independent variables (*p*), and the population value of R^2. R^2 is a measure of effect size for the entire model—that is, the proportion of variance explained. The *v* measure and, thus, accuracy increase as sample size and effect size increase but decrease as the number of independent variables becomes large.

It is important to point out that *v* requires only standard homoskedasticity assumptions and does not require any assumptions regarding the specific underlying distribution of the error term in the standard linear model. Hence, while an ANOVA requires a normality assumption, this is not required for *v*. Full details of the derivation of *v* can be found in the Appendix.

Using *v* to determine sample size

Assuming a fixed type I error rate, statistical power (Cohen, 1988) is a function of the same quantities as *v* and ranges in value from 0 to 1. Although we assume normality to calculate power for the following graphs, *v* does not require this assumption. To see how the two measures compare, Figs. 2 and 3 plot *v* (the hashed line) against statistical power for an omnibus test with an α-level of .05 (the solid line) at different effect sizes as a function of sample size and numbers of population values to be estimated. The three rows of Fig. 2 represent 3, 6, and 9 population values, and the three columns represent conventionally "small" (R^2=.02), "medium" (R^2=.13), and "large" (R^2=.25) true effect sizes (see Cohen, 1988). Likewise, the three rows of Fig. 3 represent 11, 14, and 18 population values under these same effect sizes.

The reader should note that we are referring to the statistical power of the *F*-test for the full model under consideration. See Maxwell (2004) for alternative definitions of power under the linear model.

Recall that, for ANOVA designs, the common f^2 effect size measure is related to R^2 via the identity $f^2 = \frac{R^2}{1-R^2}$.

Prospectively determining the sample size necessary to produce some *v* given a true R^2 is another way to carry out sample size planning. To see how it compares with traditional power analysis, compare the hashed line with the solid line in Fig. 2. The two curves track each other quite closely, with *v* values generally being smaller than corresponding power values for small samples. As *n* increases, a crossover point occurs at which *v* is larger than power. As effect size decreases and the number of independent variables increases, the curves more greatly diverge, and the crossover point occurs at a larger value of *v*. In these situations, it is possible to have excellent power to reject a null hypothesis under a statistical test using OLS, yet, for nearly all possible true states of nature, OLS is getting no closer to the truth than is RLS, which randomizes information about treatment effects. This is true even for a power of .80 (Fig. 3, lower left-hand graph), a standard benchmark for adequate statistical power. This discrepancy between *v* values and power becomes more pronounced as the number of population values to estimate increases.

As an example, suppose a researcher was planning an experiment with three predictors (*p*=3) and, a priori, expected an overall effect size of R^2=.05. By plugging these values of *p* and R^2 into the *v* equation (using the provided software), it is straightforward to calibrate total sample size to obtain the desired level of *v*. For these values, a sample size of 36 observations per predictor (*n*=108 total) yields a *v* of .51. A sample size of 93 observations per predictor (*n*=279 total) yields a more satisfactory *v* of .80. For this example, a traditional power analysis would recommend 73 observations per predictor for a power of .80. Clearly, rules of thumb such as 20 observations per predictor (e.g., Simmons et al., 2011) will not guarantee a minimally acceptable value of *v*.

Given that *v* is generally less than or equal to power for values<.5, a troubling conclusion emerges. Prior studies have noted that average statistical power in some areas of psychology is at or below .5, with average power from studies with small effects being even lower (Maxwell, 2004; Sedlmeier & Gigerenzer, 1989; Tressoldi, 2012). To the extent that these studies continue to be representative, *v*<.5, or even *v*=0, is a typical condition for published studies in some areas.

To summarize, *v* can be used in conjunction with traditional power analyses to set sample size. It is important to note that *v* and statistical power are conceptually distinct. Statistical power corresponds to the probability of detecting an effect using a null hypothesis test under a specified alpha level. The *v* measure, in contrast, speaks to estimation accuracy—that is, how accurately OLS is estimating the population values, as compared with a benchmark estimator that scrambles the directionality of effects. Again, *v* is distribution free and in no way defined in terms of the null hypothesis testing framework.

Using v in a meta-analysis

We have illustrated how *v* could be applied prospectively to determine sample size. Alternatively, *v* could also be applied

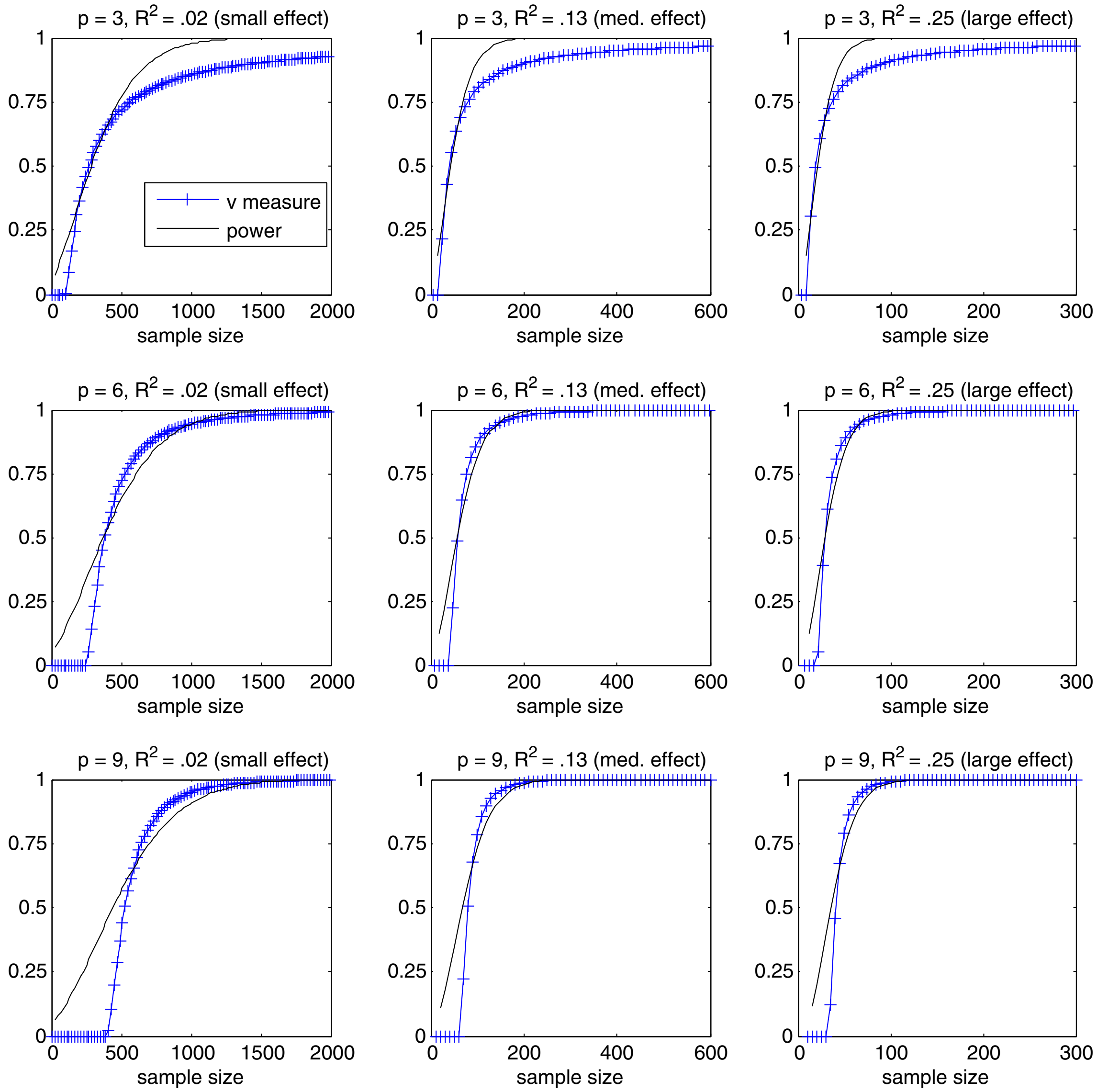

Fig. 2 All nine graphs in this figure plot both v and statistical power (assuming α=.05) as a function of sample size. Graphs 1–3 consider regression models with three parameters under "small" (R^2=.02), "medium" (R^2=.13), and "large" (R^2=.25) effect sizes (Cohen, 1988). Graphs 4–6 consider six parameters under these different effect sizes, with graphs 7–9 considering nine parameters

in a meta-analytic fashion. Meta-analytic techniques are often used to estimate effect sizes for psychological phenomena by pooling similar studies together (Hartung, Knapp, & Sinha, 2008). Given that v requires an estimate of overall effect size, R^2, one could estimate this parameter using traditional meta-analytic techniques. Once a good estimate has been obtained, v could be calculated assuming various levels of n and p to estimate accuracy and plan sample sizes for common experiments.

v as a measure of replicability

The replicability of experimental findings is an important issue that has recently received increased attention after high-profile failures to reproduce influential results (Ritchie, Wiseman, & French, 2012). In areas that rely heavily on null hypothesis testing, replication is typically construed as repeating the result of a hypothesis test on a new set of data. For example, if an experiment produces a significant p-value and another experimenter reruns the experiment but does not produce a significant p-value, we often conclude that the experiment did not replicate. This approach is problematic in some respects. For both the initial result and the replication, it relies on binary decisions about significance that involve luck. Sampling variability leads to both type I and type II errors, which can lead to somewhat perverse conclusions about replicability. For example, if one is replicating an experiment in which there are multiple treatment groups, one could get estimates of the true means that are somewhat close to those

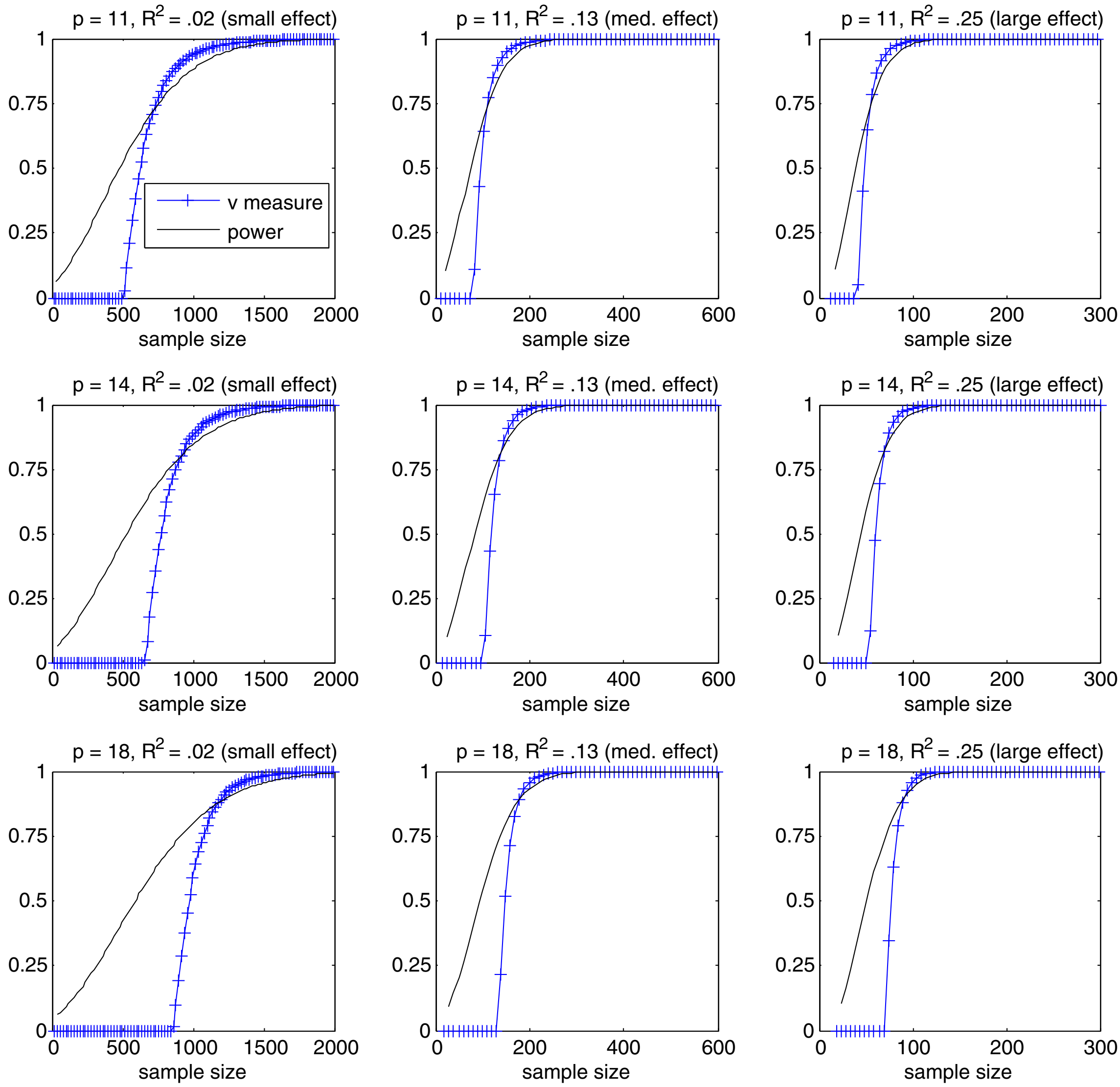

Fig. 3 All nine graphs in this figure plot both *v* and statistical power (assuming α=.05) as a function of sample size. Graphs 1–3 consider regression models with 11 parameters under "small" (R^2=.02), "medium" (R^2=.13), and "large" (R^2=.25) effect sizes (Cohen, 1988). Graphs 4–6 consider 14 parameters under these different effect sizes, with graphs 7–9 considering 18 parameters

in the original experiment, perhaps well within the original confidence intervals, and yet fail to obtain a significant *p*-value. In this way, a null result can be taken as a failure to replicate, and accordingly, the replicating experimenter must make careful statistical power planning decisions to avoid type II errors.

We suggest another way to conceive of replicability that does not rely upon hypothesis testing. A result can be said to be replicable to the extent that the parameter estimates, such as sample means, remain similar across experimental replicates. For example, suppose that an experimenter were to run the same experiment under the same conditions many times, faithfully following all experimental procedures. The estimates will not be exactly the same on every replicate, because of sampling variability. Given the factors of sample size, effect size, and the number of parameters to estimate (e.g., the number of experimental conditions), the amount that the estimates vary across replicates will be greater or smaller. We can quantify the replicability of the experiment simply as the sampling variance of those estimates across replicates.

For unbiased estimators such as sample means (and more generally, OLS), the total variance of the estimates is well-known to be equivalent to the estimator's *MSE*, a metric we have used throughout in our derivation of *v*. Put more formally, for OLS (or any unbiased estimator), *MSE* is simply the trace of the covariance matrix of the OLS estimates. *MSE* is literally equivalent to the sampling variance of the estimator. Furthermore, the sampling variance of the estimator is a quantity that, given some effect size estimate, can be calculated without actually having to run an experiment multiple times. In the simplest case, the effect size from the single experiment, corrected for bias, could be used as an estimate of effect size. Such an approach may overestimate effect sizes, but as is evident in Fig. 2, it is possible for a statistically significant result to have unacceptable *MSE*, as compared with RLS, on its face, given its own effect size, sample size, and number of parameter estimates. Estimates with such large values of *MSE* (total variance) should not be expected to replicate in the first place; that is, we should not expect future estimates to be similar to the original estimates.

To clarify, this perspective on replication is concerned only with the sampling variance of the estimates, and not whether the initial experimental results were obtained through any impropriety on the part of the researcher. For example, Nosek, Spies, and Motyl (2012) described running an exact replication (n=1,300) of one of their own large experiments (total n=1,979) that initially found a hypothesized effect, only to have the effect disappear. The result did not replicate, under the traditional definition, even though no post hoc tampering with the data was used to get the original result and the procedures were carefully reproduced. The culprit was simply sampling variance. See also Miller (2009) and Miller and Schwarz (2011) for additional perspectives on this problem.

To summarize, we propose that several statements are equivalent: The replicability of an experimental result is the replicability of the estimator itself, which is in turn a measure of the expected variability of parameter estimates across experimental replicates, which, for the unbiased estimators that are so ubiquitous in psychology research, is just the *MSE* of the estimator (the expected sum of squared differences between the estimates and their corresponding true values).

Given its precise definition and meaning, as well as its fundamental place in statistics, we argue that *MSE* is underappreciated as a measure of replicability. At the same time, it is not clear that many areas of psychological inquiry are quantified enough to make sense of whether *MSE* values are good or bad. That is, while we can quantify *MSE* and, thus, replicability, we are still left with the difficult problem of deciding whether a given amount of *MSE* is satisfactory in some experimental context. We have argued that v is a solution to such problems because it provides a minimum benchmark for the accuracy of OLS: OLS should be used only under conditions in which it more accurately estimates population parameters than does an estimator that randomly determines the relative values of its coefficients (randomly determines treatment effects). Thus, v can be used as a measure of replicability that avoids some of the problems of traditional approaches. Just as v is a minimum standard for accuracy when determining treatment effects, v also has the interpretation of a standard on whether findings of treatment effects have minimally acceptable replicability—that is, a standard on acceptable sampling variance. If estimates that yield random treatment effects are closer to the true parameter values on average, we should not expect a finding to replicate.

For the biased RLS estimator, *MSE* is a function of both squared bias and variance and, hence, should not be equated with the replicability of RLS. We use the *MSE* of RLS only as a benchmark for determining the acceptability of the *MSE* of OLS, which is, in turn, the replicability of OLS. In this way, v can be taken as a measure of replicability.

General discussion

The v measure evaluates the accuracy of standard estimation techniques, such as sample means and regression coefficients, relative to a benchmark estimator. Our analysis indicates that standard estimation techniques can be extremely inaccurate, particularly for studying small- and medium-sized effects. Current practice in psychology would not preclude publishing findings under conditions when v<.5, or even v=0. As is shown in Figs. 2 and 3, v=0 does not imply nonzero power or even unusually low power for many areas of psychology. Yet, in these situations, it is certain that the ubiquitous OLS estimator will be less accurate, on average, than our RLS estimator, no matter what the true population values are. Furthermore, we stress that these arguments are made under favorable assumptions for standard methods, including that all of their sampling assumptions have been satisfied, while v itself makes no assumptions regarding the specific underlying distributional form. The potentially serious problem of poor v-accuracy comes into play before we even worry about complications such as researcher degrees of freedom (Simmons et al., 2011) and the various types of publication bias (Francis, 2012a, b). Our findings, compounded with human error, suggest that the problem of inaccurate and irreplicable studies may be even graver than previously imagined.

Many readers may have already had the intuition that a significant result from a relatively small sample might not be trustworthy. But how small of a sample is too small, and given what effect size? The v measure gives a principled answer to these questions that is not ad hoc but, rather, derived from a basic argument about what sort of benchmarks OLS estimates should surpass in accuracy. For small to medium effect sizes, the requisite sample size for estimates to be trustworthy can be surprisingly large. In this way, v adds to the literature documenting the statistical challenges of estimating small effects (see Gelman & Weakliem, 2009) and the importance of basing sample size considerations on the accuracy of parameter estimation (Kelley & Maxwell, 2003; Lai & Kelley, 2012).

From a practical standpoint, the v measure can aid in decisions that are inherently subjective in nature. For example, when planning sample size, researchers can calibrate according to how wide they desire their confidence intervals to be (Kelley & Maxwell, 2003). Yet, what one researcher considers to be a narrow confidence interval may be unacceptably wide to someone else. Most areas of psychology are not quantified to the extent that there are standards as to what is or is not a narrow confidence interval. In this case, RLS provides a meaningful benchmark. If we agree that OLS should be more accurate than RLS, then a "narrow" confidence interval is, at a minimum, one in which the OLS estimates are more accurate than RLS—hence, large values of v. In this way, the v measure could be used in conjunction with existing methods to determine sample sizes according to confidence interval width (e.g., Kelley & Maxwell, 2003).

The v measure is defined on the full linear model under consideration, rather than on particular subsets or subtests within the model. For example, if one runs a factorial ANOVA, several partial effects could be measured, and v is not unique to any of them. Rather, it applies to all of them because it measures the accuracy of the OLS estimator on the full ANOVA model. Thus, where we have referred to R^2 or effect size, we mean the proportion of variance explained by the model, not a partial effect size.

In its present formulation, v and RLS are applicable to the linear model. The concept, however, could be generalized to other statistical models. The most obvious future generalization is the multivariate general linear model, in which there are multiple dependent variables. Such an advance would require more mathematical work, but the general ideas of v could be applied. Extending to this more general model would allow for an analysis of the accuracy of estimates obtained from repeated measures designs and growth models.

The v measure is based upon the RLS estimator, which utilizes a least-squares argument subject to uniform random weights. A key feature of this estimator is that no matter how much data are collected, the relative signs, orderings, and magnitudes of the coefficients are always determined randomly; that is, RLS has no convergence properties with regard to the relative values of its coefficients. Future work could explore alternative methods of scaling the random vector $\boldsymbol{a}$ that do not depend upon a least squares argument—that is, other choices of scaling factor k. Indeed, there may exist alternative estimators that yield random relationships among the parameter estimates that are even more accurate than RLS. In this way, our approach to v was conservative in that we did not explicitly optimize the RLS estimator against OLS to make it the most difficult such benchmark possible.

Recent articles in the popular press (e.g., Carey, 2011; Lehrer, 2010) and peer-reviewed journals (e.g., John, Loewenstein, & Prelec, 2012; Simmons et al., 2011) have ignited a discussion within the scientific community about when empirical results should be believed. This debate has focused almost entirely on human error—that is, activities endemic to the researchers themselves, such as publication bias and researcher degrees of freedom. While we agree that these are problems, our results demonstrate a more basic problem. Even if all practices related to data collection and publication were cleaned up, experimental results based on unacceptably inaccurate estimates would remain. To this problem, there may be no low-cost solutions; our statistical techniques may require larger samples and less measurement error to work correctly.

Author Note We are thankful to Jon Baron, David Budescu, Justin Landy, Denis McCarthy, Barbara Mellers, Uri Simonsohn, Jeffrey Rouder, Michael Birnbaum and Job van Wolferen for helpful comments, and the research team at the Max Planck Institute for Research on Collective Goods. Jason Dana is thankful for support from the Spanish Ministry grant FFI2011-28835.

Appendix

In this appendix, we summarize and apply the results of Davis-Stober et al. (2010a) and Davis-Stober (2011) to derive the v measure for orthogonal designs and lay out an algorithm that calculates v for the nonorthogonal case. We refer the reader to the original papers for the proofs of the various results and theorems.

Modeling assumptions As was stated earlier, u operates within the standard linear model, $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$, where $\boldsymbol{X}$ is a $n \times p$ design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of population weights, and $\boldsymbol{y} \sim (\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_{n \times n})$ with i.i.d. sampling. Unless stated otherwise, we assume that $\boldsymbol{X}$ is of full rank. We assume throughout that the length of the population parameter $\boldsymbol{\beta}$ is finite. Let $\|\beta\| \leq b$, where $b \in \mathbb{R}^+$ and where "$\|\cdot\|$" denotes the standard Euclidean norm. We also assume that $\boldsymbol{y}$ is standardized in z-score format. A design matrix, $\boldsymbol{X}$, is said to be *orthogonal* if $\boldsymbol{X'X}$ is a scalar multiple of the identity matrix. For example, under multiple regression, we assume that all predictor variables are standardized and uncorrelated; thus, $\boldsymbol{X}'\boldsymbol{X} = (n-1)\boldsymbol{R_{XX}}$, where $\boldsymbol{R_{xx}}$ is the predictor intercorrelation matrix; under a balanced one-way ANOVA design, we would use the standard $\boldsymbol{X}$ formulation giving $\boldsymbol{X}'\boldsymbol{X} = \frac{n}{p} I_{p \times p}$ (Muller & Fetterman, 2003).

Definition 1 (Davis-Stober et al., 2010a). Let $\boldsymbol{a}$ be a fixed $p \times 1$ vector of weights, with $\|\boldsymbol{a}\| > 0$. Then the *improper least squares* (ILS) estimator is defined as follows:

$$\widehat{\beta}_{\boldsymbol{ILS}} = \boldsymbol{a}\left(\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}\right)^{-1}\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{y},$$

and can be considered as a special case of general constrained least squares estimation (Amemiya, 1985; Chipman & Rao, 1964). To place ILS in competition with OLS, we must first determine the *MSE* of ILS with the following result. Given a choice of ***a*** and values of ***X*** and σ^2, it is routine to show that the *MSE* incurred by the ILS estimator is the following sum:

$$MSE_{ILS} = \beta' \boldsymbol{W} \beta + \frac{\boldsymbol{a}'\boldsymbol{a}\sigma^2}{\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}}, \tag{4}$$

where ***W*** is a symmetric positive semidefinite matrix defined as $\boldsymbol{W} = \boldsymbol{Q}'\boldsymbol{Q} - \boldsymbol{Q} - \boldsymbol{Q}' + \boldsymbol{I}_{p\times p}$ with $\boldsymbol{Q} = \boldsymbol{a}\left(\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}\right)^{-1}\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}$.

Given values of ***a***, σ^2, and ***X***, we can consider the *MSE* of the ILS estimator as a function of $\boldsymbol{\beta}$. This key result allows us to directly compare the *MSE* of ILS with that of OLS, which is well known to be

$$MSE_{OLS} = \sigma^2 tr\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1},$$

where "*tr*" denotes the trace of a square matrix.

Given a bound, *b*, on the length of the population $\boldsymbol{\beta}$ parameter, the set of all possible $\boldsymbol{\beta}$ forms a hypersphere of dimension *p* with radius *b* (Davis-Stober, 2011). Davis-Stober demonstrated how we can subtract the *MSE* of OLS from that of ILS and solve for the set of population $\boldsymbol{\beta}$ within this hypersphere such that $MSE_{ILS} \leq MSE_{OLS}$. This set, denoted *C*, is as follows:

$$C = \left\{\beta \in \mathbb{R}^p : \beta'\boldsymbol{W}\beta \leq \sigma^2 tr\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \frac{\boldsymbol{a}'\boldsymbol{a}\sigma^2}{\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}}, \|\beta\| \leq b\right\}.$$

The set *C* takes the form of a *p*-dimensional *ellipsoidal hypercylinder* bounded by the *p*-dimensional hypersphere of radius *b*. See Davis-Stober (2011) for a full discussion of this geometry.

The key question for our analysis is the following: What proportion of the hypersphere of all possible $\boldsymbol{\beta}$ does the set *C* occupy? Let *V* denote this proportion. Directly solving for *V* for general *C* is analytically intractable (Davis-Stober, 2011). However, Davis-Stober provides analytic upper and lower bounds on this value by constructing two bounding sets C_- and C_+. Both of these sets are *spherical* hypercylinders that satisfy the relation $C_- \subseteq C \subseteq C_+$, where " $\subseteq$ " denotes the (nonstrict) subset relation. The bounding sets, C_- and C_+, are each defined in terms of *p* (number of parameters), *b* (bound on the length of the population $\boldsymbol{\beta}$), and the *primary angle* of the spherical hypercylinder, α. The primary angle α determines the length of all radii in the spherical hypercylinder that are orthogonal to the primary axis, which is always equal to ***a*** (Davis-Stober, 2011). The length of these radii are thus equal to $b\sin(\alpha)$. Please note that the parameter α in this context is unrelated to the usual α parameter in the null hypothesis testing framework. Davis-Stober solved for the α parameters for the bounding sets, C_- and C_+, which are denoted α_1 and α_2, respectively. We briefly restate these results below.

Result 1. (Davis-Stober, 2011). For the lower bounding set C_-, α_1 is as follows:

$$\alpha_1 = \cos^{-1}\left(\frac{1-\xi_1}{\sqrt{1-2\xi_1(1-\xi_1)}}\right),$$

where $\xi_1 = \frac{\delta_1 - \sqrt{\delta_1 - \delta_1^2}}{2\delta_1 - 1}$, $\delta_1 = \min\left\{\frac{\sigma^2\omega_1}{\mathrm{b}^2}, 1\right\}$, and $\omega_1 = \left(\frac{tr\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\left(\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}\right)^2 - \|\boldsymbol{a}\|^2\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}}{\|\boldsymbol{a}\|^2\boldsymbol{a}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^2\boldsymbol{a}}\right)$.

Result 2. (Davis-Stober, 2011). For the upper bounding set C_+, α_2, is as follows:

$$\alpha_2 = \cos^{-1}\left(\frac{1-\xi_2}{\sqrt{1-2\xi_2(1-\xi_2)}}\right),$$

where $\xi_2 = \frac{\delta_2 - \sqrt{\delta_2 - \delta_2^2}}{2\delta_2 - 1}$, $\delta_2 = \min\left\{\frac{\sigma^2\omega_2}{\mathrm{b}^2}, 1\right\}$, and $\omega_2 = \left(\frac{\left(\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}\right)tr\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) - \|\boldsymbol{a}\|^2}{\left(\boldsymbol{a}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a}\right)}\right)$.

To place upper and lower bounds on *V*, we must calculate the proportion of the hypersphere of all possible βs that C_- and C_+ occupy. The following theorem provides exactly this. Specifically, it allows us to calculate the relative volume of any (full-dimensional) spherical hypercylinder bounded by a (full-dimensional) hypersphere as a function of *p*, *b*, and α.

Theorem 1. (Davis-Stober, 2011). *Let* $V_{\alpha,p}$ *be the volume of a spherical hype-cylinder with radius length equal to* $b\sin(\alpha)$ *and center axis equal to* ***a****, bounded by the p-dimensional hypersphere of radius b divided by the total volume of the p-dimensional hypersphere of radius b. Then* $V_{\alpha,p}$ *is the following real-valued function of alpha and p:*

$$V_{\alpha,p} = 1 - \frac{2\cos(\alpha)\Gamma\left(\frac{p+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{p+1}{2}\right)} \times \left[{}_2F_1\left(\frac{1}{2},\frac{1-p}{2},\frac{3}{2},\cos^2(\alpha)\right) - \sin(\alpha)^{p-1}\right], \tag{5}$$

where $0 \leq \alpha \leq \frac{\pi}{2}$, ${}_2F_1(\cdot,\cdot,\cdot,\cdot)$ *is the Gaussian hypergeometric function and* $\Gamma(\cdot)$ *is the gamma function.*

To apply these results, we must first provide a value for *b*, the upper bound on the length of $\boldsymbol{\beta}$. Generally speaking, we can apply the fundamental regression equation, $R^2 = \beta' R_{XX}\beta$. Since $\boldsymbol{R_{XX}}$ is positive definite, we have $\|\beta\|^2 \leq \frac{R^2}{\lambda_{min}}$, where $\lambda_{\min}$ is a minimal eigenvalue of $\boldsymbol{R_{XX}}$, and thus we let $b^2 = \frac{R^2}{\lambda_{min}}$. Under the orthogonal regression case, this simplifies to $b^2 = R^2$. For orthogonal ANOVA designs, we can use the equation $R^2 = \frac{\beta' X' X\beta}{n-1}$ and thus obtain $R^2 \frac{(n-1)p}{n} = \|\beta\|^2$ and set $b^2 = R^2 \frac{(n-1)p}{n}$. Either operation provides exactly the same results in the following derivations.

By applying Theorem 1 to the bounding sets, C_- and C_+, we obtain, respectively, lower and upper bounds on V, the proportion of population $\boldsymbol{\beta}$ such that $MSE_{ILS} \leq MSE_{OLS}$. These bounds are denoted V_- and V_+. The following theorem establishes when $C_- = C = C_+$ and, therefore, $V_-=V=V_+$.

Theorem 2. (Davis-Stober, 2011). *Assume that **a** is an eigenvector of the matrix **X′X**. Then* $C_- = C_+ = C$ *and* $V_- = V_+ = V$.

For the orthogonal case, ***X′X*** is a scalar multiple of the identity matrix, $\boldsymbol{I}_{p\times p}$, and, hence, *all* possible ***a***s are eigenvectors of the ***X′X*** matrix. Therefore α_1=α_2 for all ***a***s. Let $\alpha_* = \alpha_1 = \alpha_2$. Then, under the orthogonal case, α_* is as follows:

$$\alpha_* = \cos^{-1}\left(\frac{1-\zeta}{\sqrt{1-2\zeta(1-\zeta)}}\right),$$

$\zeta = \frac{\gamma-\sqrt{\gamma-\gamma^2}}{2\gamma-1}, \gamma = \min\left\{\frac{(p-1)\left(1-R^2\right)}{(n-p)R^2}, 1\right\}$. It is important to note that the ***a*** terms drop out when solving for α_* under the orthogonal case. In other words, V is exact and invariant under any choice of ***a*** and is calculated via α_* and Equation 5. This leads to our main result.

Main result on the RLS estimator Assume an orthogonal design. Let ***a*** be a p-dimensional random vector uniformly distributed over the surface of the unit p-dimensional hypersphere centered at the origin. Let the RLS estimator be defined as the ILS estimator with ***a*** obtained from a single random draw of ***a***. Then, assuming fixed values of R^2, n, and p, the distribution of V values under the RLS estimator is degenerate, with all V values being equal for all possible draws from ***a***. Thus, under the RLS estimator, V is exact and is calculated via α_* and Theorem 1.

It is important to note that it suffices to sample ***a*** uniformly from a unit hypersphere centered at the origin, since the ILS estimator is invariant under scalar multiplication of ***a*** (see Davis-Stober et al., 2010a). Under different choices of ***a***, as in the RLS estimator, the main axis of the spherical hypercylinder C will always be the ***a*** that was sampled from ***a***; however, the shape and relative volume of C will be invariant under any choice of ***a***. Thus, the relative volume of the set C is exact under the RLS estimator and is a function of only R^2, n, and p. Finally, we define v (as defined in the main text) as the complement of V—that is, $v := 1 - V$. This simple transformation orients the results in terms of the proportion of population $\boldsymbol{\beta}$ which favor OLS over RLS in terms of *MSE*.

Nonorthogonal case For the nonorthogonal case, the distribution of V values is *not* degenerate; that is, different sampled choices of ***a*** will yield different V values. We estimate v for the RLS estimator via the following Monte Carlo algorithm. For cases where $\boldsymbol{X'X}^{-1}$ is not full rank, we apply the Moore–Penrose inverse of ***X′X***.

Input p, n, adjusted R^2, and ***X′X***.

1. Uniformly sample k-many ***a*** vectors from the surface of p-dimensional unit hypersphere of unit radius and dimension p (or r for non-full rank design matrices) centered at the origin. Let $\boldsymbol{a}_i$ denote the i^{th} sample.
2. Calculate V_- and V_+ for each $\boldsymbol{a}_i$ as described above. Let $(V_-, V_+)_i$ denote the pair associated with each sampled $\boldsymbol{a}_i$.
3. Calculate $V_i = Mean(V_-, V_+)_i, \forall i \in \{1, 2, \ldots, k\}$, where $Mean(\cdot)$ denotes the mean of a set.
4. Calculate $V_{estimate} = Mean(V_i, \forall i \in \{1, 2, \ldots, k\})$.
5. Return $v = 1 - V_{estimate}$.

This algorithm yields an estimate of the expected proportion of population $\boldsymbol{\beta}$ such that $MSE_{OLS} \leq MSE_{RLS}$. The number of samples, k, necessary to estimate v within appropriate error bounds depends on the dimension of the space, p. We recommend a minimum of k=10,000 samples for relatively large values of p—for example, p=10. We also note that for many ***X*** matrices, $b^2 = \frac{R^2}{\lambda_{min}}$ may be an overly conservative bound on the true length of $\boldsymbol{\beta}$. We suggest that $b^2 = \frac{R^2}{\lambda_*}$, where $\lambda_* = \frac{1}{p}\sum_{i=1}^{p} \lambda_i$, may be more appropriate.

References

Amemiya, T. (1985). *Advanced econometrics*. Harvard University Press.

Bem, D. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100,* 407–425.

Bickel, P. J., & Doksum, K. A. (2001). *Mathematical statistics: Basic ideas and selected topics (Vol. 1)* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Carey, B. (2011). Fraud case seen as a red flag for psychology research. *The New York Times, published November 2.*

Chipman, J. S., & Rao, M. M. (1964). The treatment of linear restrictions in regression analysis. *Econometrica, 32,* 198–209.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dana, J., & Dawes, R. (2004). The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics, 29,* 317–331.

Dana, J. (2008). What makes improper linear models tick? In J. Krueger (Ed.), *Rationality and social responsibility: Essays in honor of Robyn Mason Dawes*. Mahwah, NJ: Lawrence Erlbaum Associates.

Davis-Stober, C. P. (2011). A geometric analysis of when fixed weighting schemes will outperform ordinary least squares. *Psychometrika, 76,* 650–669.

Davis-Stober, C. P., Dana, J., & Budescu, D. (2010a). A constrained linear estimator for multiple regression. *Psychometrika, 75,* 521–541.

Davis-Stober, C. P., Dana, J., & Budescu, D. (2010b). Why recognition is rational: Optimality results on single-variable decision rules. *Judgment and Decision Making, 5,* 216–229.

Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist, 34,* 571–582.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95–106.

Eldar, Y. C., Ben-Tal, A., & Nemirovski, A. (2005). Robust mean-squared error estimation in the presence of bounded data uncertainties. *IEEE Transactions on Signal Processing, 53,* 168–181.

Francis, G. (2012a). Replication initiative: Beware misinterpretation. *Science, 336,* 802.

Francis, G. (2012b). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin and Review, 19,* 151–156.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex, and power: Statistical challenges in estimating small effects. *American Scientist, 97,* 310–316.

Hartung, J., Knapp, G., & Sinha, B. K. (2008). *Statistical meta-analysis with applications*. Hoboken, NJ: John Wiley & Sons.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* 696–701.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19,* 640–648.

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.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8,* 305–321.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2004). *Applied linear statistical models*. McGraw-Hill.

Lai, K., & Kelley, K. (2012). Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: Sample size planning via narrow confidence intervals. *British Journal of Mathematical and Statistical Psychology, 65,* 350–370.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer.

Lehrer, J. (2010). The truth wears off. *The New Yorker, published December 13.*

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods, 5,* 434–458.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

Maxwell, S. E., & Delaney, H. D. (2003). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin and Review, 16,* 617–640.

Miller, J., & Schwarz, W. (2011). Aggregate and individual replication probability within an explicit model of the research process. *Psychological Methods, 16,* 337–360.

Muller, K. E., & Fetterman, B. A. (2003). *Regression and ANOVA: An integrated approach using SAS software*. Cary, NC: John WIley & Sons Inc., SAS Institute Inc.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615–631.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence. *Perspectives on Psychological Science, 7,* 528–530.

R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ritchie, S. J., Wiseman, R., French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS ONE, 7,* doi:10.1371/journal.pone.0033423

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

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.

Teräsvirta, T. (1983). Restricted superiority of linear homogeneous estimators over ordinary least squares. *Scandinavian Journal of Statistics, 10,* 27–33.

Toro-Vizcarrondo, C., & Wallace, T. D. (1968). A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association, 63,* 558–572.

Tressoldi, P. E. (2012). Replication unreliability in psychology: Elusive phenomena or "elusive" statistical power? *Frontiers in Psychology, 3,* 218. doi:10.3389/fpsyg.2012.00218

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213–217.