European Journal of Psychology Open

The Bayesian One-Sample t-Test Supersedes Correlation Analysis as a Test of Validity

Phivos Phylactou, Marietta Papadatou-Pastou, and Nikos Konstantinou

Online First Publication, December 17, 2024. https://dx.doi.org/10.1024/2673-8627/a000069

Original Article

The Bayesian One-Sample *t*-Test Supersedes Correlation Analysis as a Test of Validity

Phivos Phylactou[1,2], Marietta Papadatou-Pastou[3,4], and Nikos Konstantinou[5]

[1]School of Physical Therapy, University of Western Ontario, London, Ontario, Canada
[2]The Gray Centre for Mobility and Activity, Parkwood Institute, London, Ontario, Canada
[3]Biomedical Research Foundation of the Academy of Athens, Greece
[4]Department of Special Education and Psychology, National and Kapodistrian University of Athens, Greece
[5]Department of Rehabilitation Sciences, Cyprus University of Technology, Limassol, Cyprus

Abstract: *Introduction:* The validity of measurement, which refers to how accurately tools measure what they are intended to measure, is essential in science. Researchers rely on statistical approaches to test the validity of their measures. One such approach is correlation analysis. Even though correlation analysis can capture high nonsystematic errors between measures, it can often lead to misleading conclusions when observations are measured with systematic errors. *Methods:* We used Monte Carlo simulations with 10,000 iterations to generate the data in each simulation. *Results:* We demonstrate how correlation analysis is commonly used to test for validity and how this method can fail with systematic error. We further propose an alternative to correlation analysis – the Bayesian one-sample *t*-test – for cases where using a simple statistical test can be justified. We provide additional simulations as well as an application to real data, showcasing the implementation of the Bayesian one-sample *t*-test and how to use it to address the limitations of correlation analysis. *Discussion:* We suggest using the Bayesian one-sample *t*-test to identify both systematic and nonsystematic error and moreover to provide evidence for the null hypothesis of no differences between two measures. *Conclusion:* As a test of validity, the Bayesian one-sample *t*-test supersedes correlation analysis.

Keywords: correlation, validity, psychometrics, Bayesian, *t*-test

Introduction

Validity, which refers to the accuracy of a tool to measure what it is intended to measure, is a vital element of science – be it an instrument or a method. Ideally, the tools employed for producing rigorous science should be perfectly valid, allowing for errorless measurements. In the psychological sciences, however, such errorless measurement is often impossible (Schmidt & Hunter, 1996), given the multifaceted aspects of the human psyche and the available tools often utilized to measure it (e.g., self-report questionnaires). These error-prone approaches are partially responsible for the criticisms that psychology, as a scientific field, has received for its less-than-optimal reproducibility over the past decade (e.g., Derksen, 2019; see also Scheel et al., 2021).

Fortunately, this so-called *reproducibility (aka replication) crisis* in psychology has facilitated promising advancements in the field, such as the introduction of preregistered reports (Chambers & Tzavella, 2022; Nelson et al., 2018; Nosek et al., 2018) and the promotion of Bayesian statistics (Derksen, 2019; Dienes, 2014, 2019, 2021a, 2021b; Dienes & Mclatchie, 2018; Scheel et al., 2021). One way in which these advancements promote reproducibility is by fostering validity assurances, which necessitate statistical analyses that provide evidence regarding the validity of the employed tools. To comply with these assurances, psychology researchers commonly tend to prefer simple statistical approaches.

Even though sophisticated validity analyses (e.g., specificity and sensitivity analyses; Marchevsky et al., 2020; Stites & Wilen, 2020) should often be used, relying on the simplest statistical model may be satisfactory. A simple statistical model can be easy to implement, save time, offer a straightforward interpretation, and avoid overfitting and consequently misinterpreting the data (Blanchard et al., 2018; Myung & Pitt, 1997). One such simple statistical model, frequently utilized to test validity, is *correlation*.

As an illustration, correlation was recently employed to test the validity of psychometric tools (e.g., Yaşar et al., 2022), computerized tools (Drevon et al., 2017), and even meta-analytic evidence (Phylactou et al., 2022).

Although, under some circumstances, correlations can be an adequate validity test, in many cases, they might be insufficient to capture important quantitative differences between measures, such as systematic error (e.g., a consistent offset of the measure from the true value). In this simulation study, we argue that, by applying a Bayesian approach, we can replace correlations with another simple statistical test, the Bayesian one-sample *t*-test, which can serve as a more sensitive validity test. As described in detail later (see Simulation 3: Bayesian One-Sample *t*-Test as a Measure of Validity), the Bayesian *t*-test can test for either the presence or absence of quantitative differences between two measurements (Phylactou, Chen, et al., 2024; Rouder et al., 2009; Wagenmakers et al., 2010). Further, using simulated data, we provide three examples to indicate (1) how correlations are commonly used as a test of validity, (2) how correlations can fail as a measure of validity, and (3) how the Bayesian one-sample *t*-test can function as a validity test. Finally, we showcase the application of the Bayesian one-sample *t*-test on real data from earlier work (Charalambous, Phylactou, Kountouri, et al., 2022).

Simulation 1: Correlations as a Measure of Validity

Often, one can use a correlation to test the validity of a tool against another tool of a similar construct. For example, Charalambous, Phylactou, Kountouri, and colleagues (2022) used correlation analysis to test the validity of a psychometric tool (Aphasia Impact Questionnaire-21 Greek Version; AIQ) compared to another tool considered the "gold standard" in the field: Stroke and Aphasia Quality of Life Scale-39 (SAQOL; Hilari et al., 2003). Researchers can distinguish between low, moderate, high, and extreme validity using predefined cut-offs. Commonly, a correlation of $\rho > 0.9$ or $\rho < -0.9$ is considered an acceptable validity cut-off in psychometry, while a correlation between $0.5 > \rho > -0.5$ typically indicates very poor validity (Taherdoost, 2016; see also Charalambous, Phylactou, Elriz, et al., 2022; Charalambous, Phylactou, Kountouri, et al., 2022; Hilari et al., 2003, 2018).

To illustrate how to implement a correlation to test the validity in such a manner, consider a hypothetical scenario of comparing two weighing scales. One is a perfect scale that always provides values corresponding to one's true weight. Let's assume that we weighted 100 1-year-old children on this perfect scale, whose weights ($\text{Weights}_{\text{true}}$) stem from a normal distribution, with a mean $m = 11.5$ kg and a standard deviation $sd = 1.8$, as shown in Equation 1 (data based on Fryar et al., 2021).

$$Weight_{true} \sim N(11.5, 1.8) \quad (1)$$

The second scale slightly deviates by up to 0.5 kg above or below the actual weight, which we consider a tolerable measurement error in this example, where we are weighing humans. If we assume that any deviation between −0.5 kg and 0.5 kg in this low error scale has equal probability (i.e., is nonsystematic), then we can express this low error using a uniform distribution, as shown in Equation 2.

$$Error_{low} \sim U(-0.5, 0.5) \quad (2)$$

Therefore, for each child, we have a value corresponding to their true weight, drawn by the normal distribution shown in Equation 1, and a value corresponding to their slightly inaccurate weight, their true weight plus some nonsystematic error drawn by the distribution in Equation 2. We used Monte Carlo simulations to simulate these distributions. Specifically, we generated 10,000 simulations from each distribution for 100 observations. All simulations presented in this study were generated in Python (v3.9.13), using the scipy (v1.12.0) and numpy (v1.26.4) packages, while we conducted the statistical analyses using the pingouin package (v0.5.2). To create the distribution used in Simulation 1, we drew one of the 10,000 values for each observation at random. Figure 1A illustrates the simulated distributions of this example. Given such low error in the second scale (i.e., 0.5 kg), a correlation analysis between the weight from the perfect scale and the weight from the low error scale (see Figure 1B) indicates an almost perfect correlation (Pearson's $\rho = .989$, $BF_{10} = 6.80 \times 10^{78}$, $p < .001$). From this, we infer that the two scales provide almost identical values: The validity is extremely high.

Now consider a third scale, with a higher deviation of up to 5 kg. As in the previous example, we assume that any deviation between −5 kg and 5 kg has equal probability (i.e., is nonsystematic), and we express this as a uniform distribution, as shown in Equation 3.

$$Error_{high} \sim U(-5, 5) \quad (3)$$

We used the same Monte Carlo approach described above to generate the distribution of the high-error scale. The distributions of the simulated weights of the perfect and the high-nonsystematic-error scales are presented in Figure 1C. In this case, we can still find a correlation between the two scales (Figure 1D), although this correlation is far from perfect (Pearson's $\rho = .477$, $BF_{10} = 31621.97$, $p < .001$), with the correlation index signifying very poor reliability (i.e., $-0.5 < \rho < 0.5$; see Taherdoost, 2016).

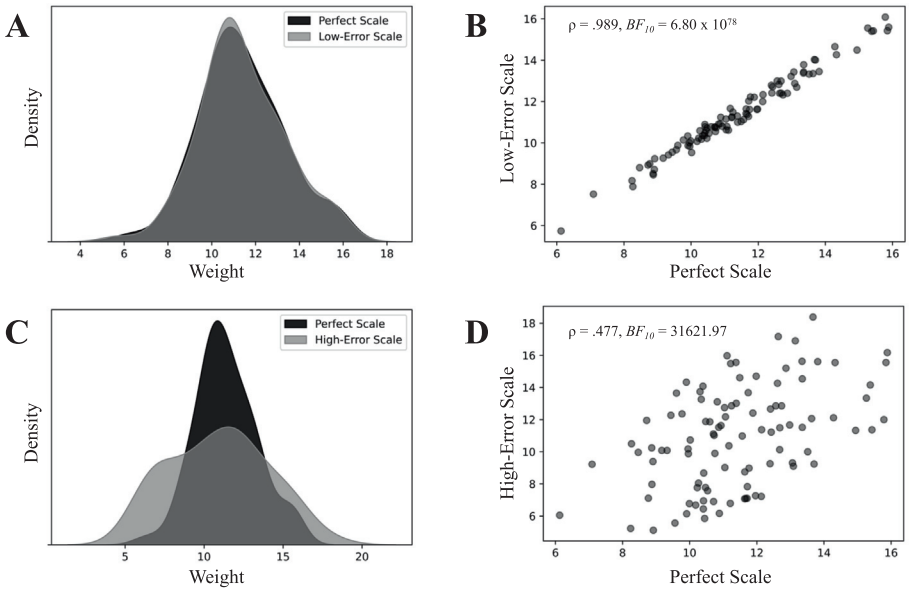

Figure 1. Distributions and scatter plots of the simulated data from a perfect scale, a scale with low deviation, and a scale with high deviation. (A) The distribution of the weights as observed by the perfect scale (black) and the low-error scale (gray). Given the small measurement error (up to 0.5 kg), the two distributions show high overlap. (B) A scatter plot showing the values obtained by the perfect scale on the y-axis and the values obtained by the low-error scale. An almost perfect linear relationship is evident. (C) The distribution of the weights observed by the perfect (black) and high-error (gray) scales. Because of the high error (up to 5 kg), there is very little overlap between the distributions. (D) A scatter plot of the values observed by the perfect (*x*-axis) and the high-error (*y*-axis) scales. A linear relationship still exists between the two observations but with high variability.

This example demonstrates how to implement correlations as a simple statistic to test the validity of a measure or a tool, as compared to a gold standard. Next, we turn to demonstrating how this analysis can falsely indicate high validity when, in reality, there is none.

Simulation 2: How Correlations Can Fail as a Validity Test

As demonstrated above, correlations can offer a simple and straightforward indication of validity. However, because of their simplicity as a statistical model, they often produce misleading results and consequently fail as a test of validity. Once again, we turn to the weighing scale example and the normally distributed weights from Equation 1. This time, however, we consider a faulty scale that deviates systematically between 4.9 kg and 5.1 kg above the actual weight. If we assume equal probability of this error, we can express the deviation as the uniform distribution shown in Equation 4.

$$Error_{systematic} \sim U(4.9, 5.1) \quad (4)$$

The distribution of the 100 simulated observations resulted using the Monte Carlo simulation described in the previous section (see Simulation 1: Correlations as a Measure of Validity). Figure 2A shows the distributions of the simulated weights deriving from the perfect scale and the scale with the systematic error above the actual weight. In this scenario, a correlation analysis between the two scales (Figure 2B) indicates an almost perfect linear relationship (Pearson's $\rho = .999$, $BF_{10} = 7.30 \times 10^{143}$, $p < .001$). If we consider how correlations were previously used to test validity (e.g., Charalambous, Phylactou, Elriz, et al., 2022; Charalambous, Phylactou, Kountouri, et al., 2022; Drevon et al., 2017; Hilari et al., 2018; Phylactou et al., 2022; Yaşar et al., 2022), we should assume that the scale with the systematic error is a valid measure, since the correlation is almost perfect, although in every measure by the faulty scale there is a constant difference from the actual weight between 4.9 kg and 5.1 kg. In other words, the results from this correlation analysis are misleading and can be misinterpreted as evidence of high validity, even though the two tools provide different observations.

Here, note that further statistical analyses can be implemented to identify whether correlations correspond to high validity between the tools or whether the tools differ between them. Such analyses concern the calculation of differences between the slopes and/or the intercepts of the observed correlations (see Phylactou et al., 2022, for an example), compared to the expected (e.g., perfect) correlation (Figure 2C). However, such analyses are considered

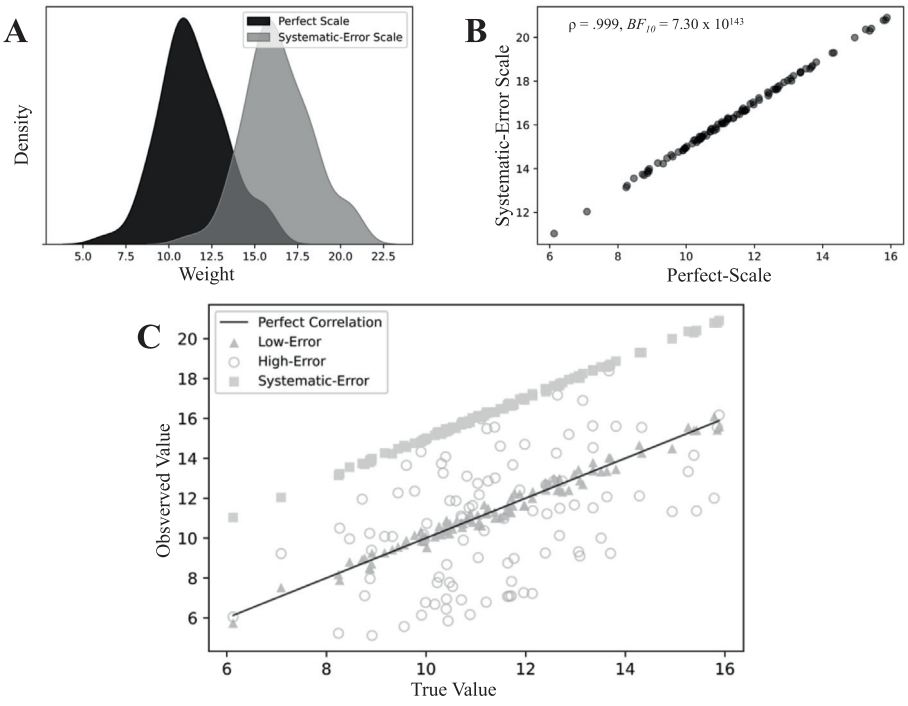

Figure 2. Distributions and scatter plots of the simulated data from a perfect scale, a scale with low deviation, and a scale with high deviation. (A) The distribution of the weights as observed by the perfect scale (black) and the scale with a systematic error (gray). Given the systematic error above the true weight (between 4.9 and 5.1 kg above), the mean of the systematic error scale is shifted, creating only minor overlap between the tails of the distributions. (B) A scatter plot of the perfect (*x*-axis) and systematic-error (*y*-axis) scales, which indicates a nearly perfect correlation. (C) Scatterplots generated by plotting the true values observed by the perfect scales with the low-error (triangle), high-error (circle), and systematic-error (square) scales, as compared to a perfect linear relationship (black line).

exploratory and are often omitted (see Charalambous, Phylactou, Elriz, et al., 2022; Charalambous, Phylactou, Kountouri, et al., 2022; Drevon et al., 2017; Hilari et al., 2018; Yaşar et al., 2022), since they do not offer the benefits of applying a simple statistical test as described above (e.g., Blanchard et al., 2018; Myung & Pitt, 1997; see also Wagenmakers et al., 2010).

Simulation 2 demonstrates how the use of correlation analysis to test validity can be misleading. Considering both Simulations 1 and 2, we conclude that a correlation analysis can capture random (nonsystematic) error between two measures; however, if the error is systematic, then the results might easily be misinterpreted. In terms of psychometrics, we can associate this example with a situation where one examiner tends to systematically provide lower or higher scores in a particular test battery because of stricter or more flexible definitions of the psychological characteristic under test. However, this does not necessarily signify that scientists should turn to complex statistical models to account for all possible variability that the utilized psychometric tools could expect. In the following section, we propose using a one-sample Bayesian *t*-test, which can be employed as a simple but more sensitive alternative to correlation analysis.

Simulation 3: Bayesian One-Sample *t*-Test as a Measure of Validity

Why Bayesian? one might think. As described in detail in the following paragraphs, advocates in favor of Bayesian statistics (e.g., Derksen, 2019; Dienes, 2014, 2019, 2021a, 2021b; Dienes & Mclatchie, 2018; Scheel et al., 2021) argue that, compared to the traditional Neyman-Pearson frequentist approach (Neyman & Pearson, 1933), a Bayesian approach offers numerous advantages. One such advantage relates to calculating the Bayes Factor (BF). The BF has the advantage of quantifying evidence in favor of either of two competing hypotheses, such as the alternative hypothesis or the null hypothesis, as opposed to the frequentist approach of using a *p*-value, which can only inform about the rejection (or failure of the rejection) of the null hypothesis (Dienes, 2014, 2021a; Johansson, 2011; Wagenmakers, 2007; but see Lakens et al., 2020). In addition to the property of the BF to test evidence in favor of the null hypothesis, the Bayesian approach offers additional advantages over the frequentist approach. For example, within the Bayesian framework, the resulting posterior probabilities – and, respectively, the BFs – are informed by both prior probabilities and the observed

data. As new data are observed, these probabilities can be updated. Hence, this continuous nature of Bayesian probabilities, in contrast to *p*-values, renders the Bayesian framework coherent, consistent, and complete (Jaynes, 2003; Wagenmakers et al., 2018).

In the context of validity testing, we are interested in the differences (or lack thereof) between the two tools for each set of observations. As an indication of validity, no differences should exist between the two tools, assuming a small margin of measurement error tolerance. To statistically analyze differences (e.g., differences between two means), the frequentist *t*-test is commonly utilized (Delacre et al., 2017), which, through the conventional null-hypothesis significance testing, provides a *p*-value that indicates whether to reject the null hypothesis (i.e., no difference) or not (for a critical review of this approach, see Scheel et al., 2021). In other words, using the conventional frequentist approach to obtaining a *p*-value, we can only infer the proposed alternative hypothesis; but when it comes to the null, the null-hypothesis significance testing approach does not allow us to discriminate between the absence of evidence and evidence of absence (Dienes, 2014, 2021a, 2021b; Dienes & Mclatchie, 2018). However, to use a *t*-test as a validity test, we need to be able to provide evidence in favor of no difference (i.e., the *null* hypothesis). Put simply, we need to be able to statistically support a null effect. This complication could explain why correlation analysis has been preferred as a validity measure. Specifically, if researchers have to rely on significant *p*-values to support a hypothesis, they are limited in choosing a statistical test that can only reject a null hypothesis (Scheel, 2022). For example, under the frequentist approach, a *t*-test cannot provide any inferences regarding the validity of tools, because the *p*-value can only provide an indication of whether a difference exists but cannot inform us about the null hypothesis (i.e., no difference; see also Dienes, 2014, 2019; Rouder et al., 2009). Therefore, to provide statistical evidence supporting validity (i.e., significant *p*-value), researchers utilize correlation analysis, so that they can reject the null hypothesis (i.e., no linear relationship between the two tools), because, with the use of the frequentist *t*-test, they cannot provide any statistical support for the null hypothesis (i.e., no difference) but can only fail to reject it.

However, the alternative – using a Bayesian approach – does allow us to infer the null hypothesis (Dienes, 2014, 2019, 2021a, 2021b). Within the Bayesian framework, the statistical models that are implemented provide a ratio, called the BF. The BF indicates how likely it is for the observed data to stem from one of two competing theories, most commonly the alternative and the null hypothesis (denoted "BF_{10}" and, respectively, "BF_{01}" to signify the ratio of the null hypothesis over the alternative hypothesis; Derksen, 2019; Dienes, 2014, 2019, 2021a, 2021b; Dienes & Mclatchie, 2018; Rouder et al., 2009; Phylactou & Konstantinou, 2022). For example, a $BF_{10} = 5$ indicates that the data are 5 times more likely to have been observed under the alternative hypothesis, while a $BF_{10} = 0.2$ shows that the data are 5 times likely to have been observed under the null hypothesis.

Table 1. Bayes Factor evidence threshold heuristics (adapted from Jeffreys, 1998, and Lee & Wagenmakers, 2014)

Bayes factor (BF)	Interpretation
$1 < BF < 3$	Anecdotal evidence
$3 \leq BF < 10$	Moderate evidence
$10 \leq BF < 30$	Strong evidence
$30 \leq BF < 100$	Very strong evidence
$100 \leq BF$	Extreme evidence

Because the BF can quantify evidence in favor of either the alternative or the null hypothesis, it can be used to decide which hypothesis best describes the observed data. For statistical inference within psychology, BF heuristics have been proposed (Jeffreys, 1998; Lee & Wagenmakers, 2014) to help researchers set thresholds for deciding whether to accept the evidence in favor of one hypothesis or the other. Table 1 shows the proposed interpretations of various BF values. Even though the practice of setting specific thresholds for accepting one hypothesis over another opposes the continuous and updatable nature of the BF, it may serve as an important avenue enabling scientists to define and test dichotomous claims, which some argue are crucial for science (Uygun Tunç et al., 2023). Of note, a recent simulation of 200 million BFs showed that for the Bayesian *t*-test, a $BF \geq 3$, should be considered adequate for psychological research, especially when gathering evidence in favor of the null hypothesis (Phylactou, Chen, et al., 2024). As such, we similarly propose that, when using the Bayesian one-sample *t*-test as a test of validity, a $BF \geq 3$ should be employed as the decision threshold.

We suggest using a *t*-test rather than correlation analysis under the Bayesian approach because correlation analysis cannot distinguish systematic from nonsystematic error (unless more sophisticated analyses are performed) under both the frequentist and Bayesian approaches, as illustrated in Simulations 1 and 2. In contrast, a Bayesian one-sample *t*-test can be used to set a tolerable margin of measurement error (by adjusting the *t*-test-test value; see below) and provide evidence supporting either a difference, or the lack thereof, when comparing two tools. Notably, more sophisticated Bayesian validity tests were previously proposed (Schluter, 2009); however, such modeling approaches can be computationally demanding and go beyond a simple statistical approach (when this can be deemed satisfactory), which is the focus of the current work.

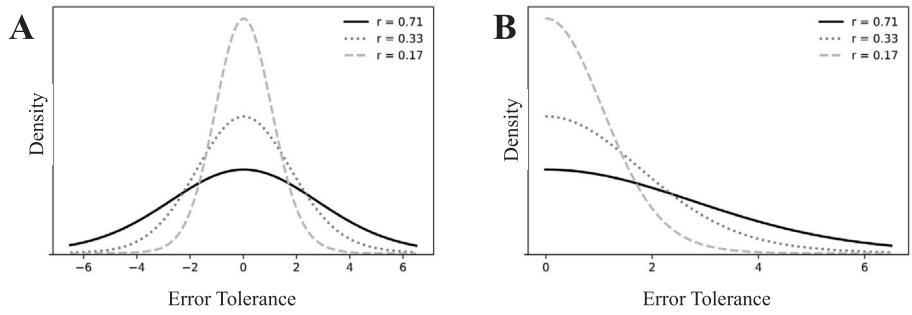

Figure 3. Example of Cauchy distributions to reflect different magnitudes of expected error. Example of Cauchy prior distributions, with different scales to inform the one-sample Bayesian *t*-test. (A) The Cauchy distribution is commonly used to test bidirectional effects, while (B) the half Cauchy can be used to restrict the expected values to positive or negative values and thus test a directional hypothesis.

For the commonly used Bayesian *t*-test, a prior distribution is assigned to the alternative hypothesis, which describes the expected effect size (i.e., the difference; δ) under the alternative hypothesis. Most commonly in psychological research, this prior distribution is expressed as a Cauchy distribution (Rouder et al., 2009). Once data have been observed, a posterior distribution can be computed, which represents the uncertainty about δ (i.e., the difference between the observed data and the test value of interest). Based on the merit that most comparisons within psychology concern nested models (i.e., the null hypothesis contains all parameters for the model of the *t*-test; Heck & Bockting, 2023; Wagenmakers et al., 2010), a BF can be computed using the Savage-Dickey density ratio (Dickey, 1971; see also Heck & Bockting, 2023; Wagenmakers et al., 2010). Put simply, the Savage-Dickey density ratio conveniently enables the computation of BFs for nested models by dividing the height of the posterior distribution by the height of the prior distribution at the test value (δ) of interest. Thorough details regarding the Bayesian *t*-test as well as mathematical proof are described in previous research (Fu et al., 2021; Kruschke, 2013; Rouder et al., 2009).

Note that, to apply a Bayesian approach, researchers must incorporate some prior assumptions regarding the data they are expecting to observe. These prior assumptions are expressed as probability distributions, usually informed by the previous literature or, in some cases, by intuition (e.g., based on the researcher's expertise; Wagenmakers et al., 2010). For the Bayesian *t*-test, these assumptions are mainly reflected in the prior distribution of the alternative hypothesis (Rouder et al., 2009), which demonstrates the magnitude (and sometimes the direction, as discussed below; see also Dienes, 2021b) of the expected effect size. Rouder and colleagues (2009) proposed a prior distribution to be used as a default for the *t*-test in psychological research, expressed as a Cauchy distribution centered on 0 with a scale of approximately $r \approx 0.71$ (see Figure 3A). Others have argued against the use of default priors (often referred to as *objective priors*; see Bandyopadhyay & Brittan, 2010; Gelman & Shalizi, 2013; Świaôtkowski & Carrier, 2020) and suggest using, for each specific contrast we are comparing, different prior distributions to reflect any prior assumptions (Dienes, 2019; 2021b).

In the case of validity testing, one can adjust the prior distribution to express how much variability can be tolerated between the compared measures. The Cauchy is considered an appropriate distribution to express this variability because it resembles a normal distribution. But it has fatter tails, which decay much slower than in the case of the normal distribution. This means that, under a Cauchy distribution, values closer to the center of the distribution have a much greater probability than extreme values, even though extreme values are still probable. Put simply, in terms of validity testing, a Cauchy prior distribution can be set accordingly to reflect the expected magnitude of error between the two measures being tested. For example, the scale of a Cauchy distribution centered on 0 can be adjusted to represent the variance, in terms of the error magnitude, that we can tolerate (see Figure 3A). Specifically, the Cauchy proposed by Rouder and colleagues (2009), with a scale $r = 0.71$, creates a cumulative probability under which approximately 80% of the expected observed values lay between 2 and −2, thus differ by at least a magnitude of 2 (compared to 0). Respectively, a Cauchy with a scale of $r = 0.33$ reflects a cumulative probability, where approximately 80% of the expected values lie within a magnitude of 1; a Cauchy scaled at $r = 0.17$ creates an 80% cumulative probability of the expected values lying within a magnitude of 0.5. Table 2 summarizes different Cauchy scales and cumulative probabilities, which can be used as an approximation for different levels of magnitude when comparing the expected error between different measures. Reducing the scale (r) of the Cauchy distribution informs the analysis that less error is tolerated, whereas increasing the cumulative probability percentage updates

Table 2. Scale of the *t*-test Cauchy prior distribution for different standard deviation thresholds when comparing *z*-scores across various distribution cumulative probabilities

	Cauchy scale (*r*)		
	Cumulative probability		
Magnitude	80%	90%	95%
0.5	0.17	0.08	0.04
1	0.33	0.16	0.08
1.5	0.49	0.25	0.12
2	0.71	0.32	0.16
2.5	0.82	0.4	0.2
3	0.98	0.48	0.25
5	1.7	0.8	0.4

the analysis to create a stricter threshold as evidence for no difference (but see Dienes, 2019, for an argument about reporting BF for multiple priors).

Given that, under the Bayesian framework, one can estimate the likelihood of both the null and the alternative hypothesis, a Bayesian one-sample *t*-test can, therefore, offer a simple approach to test whether the values derived from two different measures are the same or different from pre-established tolerance magnitudes (reflected in the one-sample *t*-test test value). As a validity test, evidence favoring the null hypothesis (i.e., no difference) indicates no differences between the two measures, i.e., a $BF_{10} < 1$ should be expected. Numerous thresholds have been proposed concerning the BF value that should be considered substantial, with various work suggesting at least a BF = 3 in favor of any of the two competing hypotheses (Derksen, 2019; Dienes, 2014, 2021a, 2021b). Note that previous work discussed that obtaining evidence for the null requires greater power than evidence in favor of the alternative, which is why some argue favor flexible thresholds (e.g., Dienes 2021a). Thus, when it comes to validity, researchers should be cautious when deciding on the validity of their measures based on the BF. Considering the above, a $BF_{10} < 1/3$ should serve as the bare minimum threshold for evidence of validity.

Because the above can reflect some assumptions regarding validity (see Table 2), and with the BF threshold defined, we now turn to an illustration on using a Bayesian *t*-test in the context of validity analysis. When employing a *t*-test as a validity test, we are interested in the absolute error when comparing one measure to another. Consequently, we should conduct analyses of the absolute values of the difference between the two measures (i.e., removing the sign after subtracting each individual value between two measures; see Equation 5). The transformation to the absolute values is an important step, because if the error is accumulated symmetrically around the true value, then the measures may average around the tolerable error – and thus the *t*-test fails to capture the actual difference between the two measures being tested. Following the transformation to the absolute values, we must compare the differences against the magnitude of the error we are willing to tolerate. For example, if we are unwilling to tolerate any differences between the two measures, then we should compare the absolute differences against 0. In other words, the closer the test value is to 0, the less error we are willing to tolerate. Notably, the specific choice of the test value for the one-sample *t*-test depends upon the outcome test measure and the amount of error we are willing to tolerate. Because of transforming the values to the absolute differences, the *t*-test can now become directional, considering only differences above the accepted error tolerance value. Within the Bayesian framework, this is simply reflected by halving the Cauchy prior to include only positive values (see Figure 3B). Thus, from here on out, we conduct our Bayesian *t*-tests using a half Cauchy prior centered on 0, with a scale $r = 0.17$ (see Table 2), which corresponds to a distribution that allocates an 80% cumulative probability for values between $0 \leq 0.5$. Put simply, the small Cauchy scale ($r = 0.17$) increases the density around the test value, which represents the null hypothesis, thus applying a very strong assumption for the test, making it harder to accumulate evidence in favor of the null hypothesis (Phylactou, Chen, et al., 2024). We choose this approach, considering that it is ideal to increase confidence in the resulting evidence, within the context of validity.

We now turn back to the weighing scale examples for Simulation 3. In these examples, our prior choice reflects the assumption that we can tolerate approximately up to 0.5 kg of scale deviance. By adjusting our prior and the test value of the one-sample *t*-test, we can reduce or increase this tolerance accordingly. Once again, we assume that we weighted 100 individuals on a perfect scale, and we have noted that their weights are described by a normal distribution with mean $M = 11.5$ ($SD = 1.8$), as shown in Equation 1. As before, we are interested in comparing the measurements acquired by the perfect scale ($Weight_{true}$) to those obtained by the low-error scale ($Weight_{low\text{-}error}$), which deviates by up to 0.5 kg (see Equation 2, something we consider a tolerable error. The first thing we need to do to apply the Bayesian *t*-test is to calculate the absolute difference between each measurement using the formula in Equation 5

$$\delta^i = \left| Weight^i_{true} - Weight^i_{low-error} \right|, \quad (5)$$

where *i* denotes each child weighed. Next, we employ a one-sample Bayesian *t*-test comparing against the value 0.5, reflecting the tolerable error, and test for evidence in favor of the null hypothesis (i.e., H_0: $\delta \leq 0.5$). This comparison yields a very high ratio in favor of the null hypothesis ($|\text{mean difference}_{true\text{-}low}| = 0.24$, $BF_{10} = 1.37 \times 10^{-28}$, $t_{(99)} = -17.12$, $p = 1$), indicating that, given the observed data,

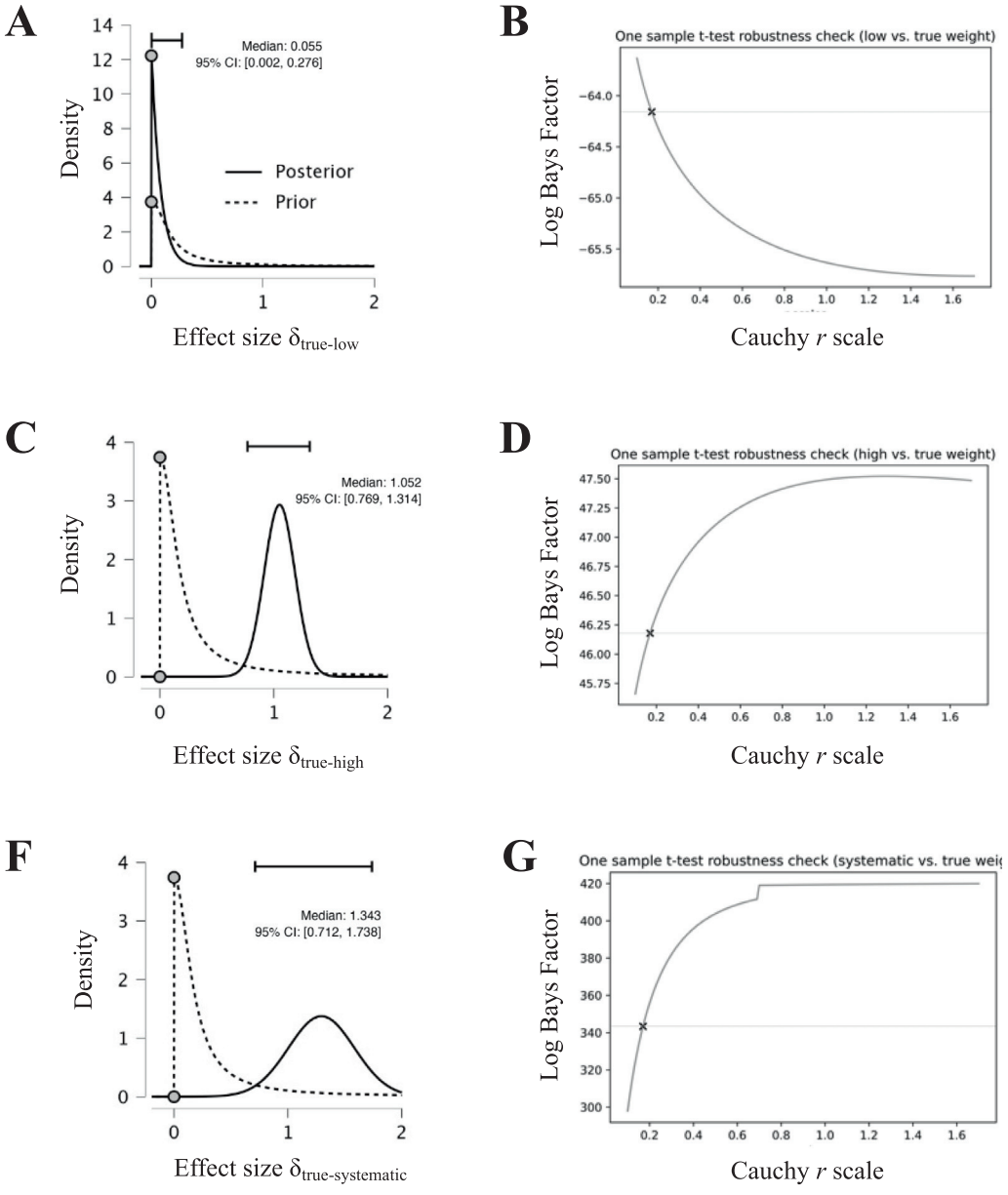

Figure 4. Posterior distributions and Bayes factor robustness analysis of the Bayesian one-sample *t*-tests comparing the difference between the simulated perfect scale with the low-deviating scale, the high-deviating scale, and the systematically deviating scale. Posterior distributions and robustness analyses for the one-sample *t*-test (one-sided, test value: $\delta > 0.5$) performed on the difference between the perfect scale with the low-deviating scale (A, B), the high-deviating scale (C, D), and the systematically deviating scale (F, G). The posteriors and reported Bayes factors (in text) were calculated using a Cauchy scale of $r = 0.17$, marked with an "x" in the robustness plots. For illustration purposes, the robustness analyses plot the log Bayes factor, where positive values indicate support for the alternative hypothesis, while negative values indicate support for the null hypothesis.

it is 6.14 × 1026 times more likely that the error is equal to or less than 0.5 kg (H_0: $\delta \leq 0.5$). As reflected by the posterior distribution, the effect δ was estimated as median $\delta_{\text{true-low}} = 0.055$, (95% CI = [0.002, 0.276]; Figure 4A). A BF robustness analysis shows that evidence favoring H0 remains for various assumptions on the prior distribution ranging from $r = 0.1$ to $r = 1.7$ (Figure 4B). This result provides evidence of the validity of the small error scale, compared to the perfect scale, for precision up to 0.5 kg. On the contrary, if we apply the same test on the largely deviating scale, which deviates up to 5 kg according to the uniform distribution of Equation 3, the t-test results in evidence supporting the alternative hypothesis ($|\text{mean difference}_{\text{true-high}}| = 2.41$, $BF_{10} = 1.14 \times 10^{20}$, $t_{(99)} = 13.11$, $p < .001$), and thus that the error from the highly deviating scale is larger than the tolerable 0.5 kg (i.e., H_1: $\delta > 0.5$). The posterior distribution estimated an effect δ of median. $\delta_{\text{true-high}} = 1.052$, (95% CI = [0.769, 1.314]; Figure 4C), while the robustness analysis indicates

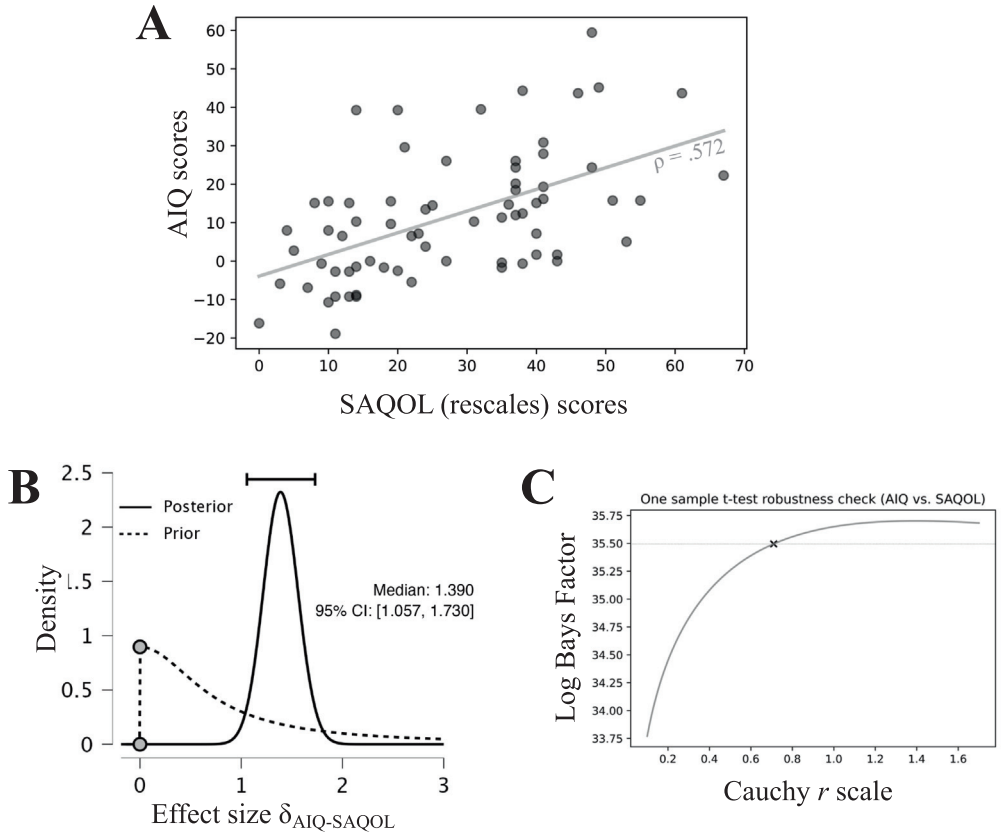

Figure 5. Scores, posterior distributions, and Bayes factor robustness analysis of the Bayesian one-sample *t*-tests comparing the difference between scores on the AIQ and SAQOL from Charalambous, Phylactou, Kountouri, and colleagues (2022). (A) A scatter plot of individual AIQ and SAQOL scores from 69 participants. (B) A posterior distribution for the one-sample *t*-test (one-sided, test value: $\delta > 2$) performed on the difference between the AIQ and SAQOL. (C) Robustness analysis of the Bayesian one-sample *t*-test. The posterior and reported Bayes factor (in text) were calculated using a Cauchy scale of $r = 0.71$, marked with an "x" in the robustness plot. For illustration purposes, the robustness analysis plots the log Bayes factor, where positive values indicate support for the alternative hypothesis, while negative values indicate support for the null hypothesis. AIQ, Aphasia impact questionnaire; SAQOL, Stroke and Aphasia Quality of Life Scale.

consistent support for H_1 under various prior distribution assumptions (Figure 4D). Moreover, applying the Bayesian one-sample *t*-test on the systematic error scale, which consistently deviates above the actual weight according to Equation 4, the *t*-test successfully identifies the error by providing evidence in favor of the alternative hypothesis ($|\text{mean difference}_{\text{true-systematic}}| = 5, BF_{10} = 1.33 \times 10^{149}, t_{(99)} = 731.51, p < .001$). The effect δ was estimated as median $\delta_{\text{true-systematic}} = 1.343$, (95% CI = [0.712, 1.738]; Figure 4F), with consistent evidence in favor of H1 under various prior distribution assumptions (Figure 4G). These findings indicate that, contrary to the correlational analysis, the Bayesian one-sample t-test can successfully identify differences in nonsystematic and systematic error cases.

Simulation 3 indicates how the Bayesian one-sample *t*-test can be applied to test for the validity between two measures when one serves as the gold standard. It also shows how the Bayesian one-sample *t*-test approach can supersede correlation analysis by successfully identifying systematic error, which, as shown in Simulation 2, correlation analysis fails to detect.

Real Data Application

We applied our proposed approach to data from previously published work to further illustrate the application of the Bayesian one-sample *t*-test as a validity test. This earlier work opted to estimate the psychometric properties of the AIQ (Charalambous, Phylactou, Kountouri, et al., 2022), which is a tool that assesses the impact of aphasia on the quality of life of people with aphasia, in a sample of 69 participants. As a test of validity, the authors in this study performed a correlation between the AIQ scores and the scores of a similar ("gold standard") test, the SAQOL (Hilari et al., 2003). The authors reported a correlation of $\rho = -0.572$, considering it as adequate validity for the AIQ.

Note that the negative correlation between the AIQ and the SAQOL was expected, since high scores in the AIQ indicate poorer quality of life, whereas high scores in the SAQOL indicate greater quality of life. To apply our Bayesian one-sample *t*-test approach, we inversed the scores of the SAQOL (by subtracting each individual score from the maximum score value), so that both tools' scoring follows

the same direction (i.e., high scores indicate poorer quality of life for both tools). Consequently, the correlation remains identical regarding magnitude, but the direction of the sign is reversed, such that it became ρ = 0.572. Further, since each tool is scored on a different scale (AIQ: from 0 to 84; SAQOL: from 1 to 5), we rescaled each individual SAQOL score *i* to match the scoring of AIQ, using linear transformation, as shown in Equation 6:

$$SAQOL^{i}_{rescaled} = \frac{(SAQOL^{i} - 1)}{4} \times 84 \qquad (6)$$

As expected, this rescaling maintained the estimated correlation between the raw scores of the two measures (ρ = 0.572). A scatter plot of the AIQ scores and the rescaled SAQOL scores is presented in Figure 5A.

Similar to Equation 5, we estimated the absolute difference between the individual AIQ and SAQOL scores (i.e., $\delta^{i}_{\text{AIQ-SAQOL}} = |AIQ^{i} - SAQOL^{i}|$, where *i* denotes each participant) to conduct the Bayesian one-sample *t*-test. For our validity test, we defined a difference of up to 2 score points as a tolerable margin of error between the two scales. Thus, we set 2 as the test value for our Bayesian *t*-test, which was informed by a half Cauchy prior centered on 0, with a scale of $r = 0.71$ (see Table 2). The results of the *t*-test provided evidence in favor of the alternative hypothesis ($|\text{mean difference}_{\text{AIQ-SAQOL}}| = 18.736$, $BF_{10} = 2603 \times 10^{12}$, $t(68) = 11.786$, $p < .001$), indicating that there are quantitative differences between the two scales which are larger than the tolerable error margin (i.e., H_1: $\delta > 2$). The effect δ, represented by the posterior distribution, was estimated as median $\delta_{\text{AIQ-SAQOL}} = 1.390$, (95% CI = [1.057, 1.730]; Figure 5B), with consistent evidence in favor of H_1 as illustrated in the robustness plot (Figure 5C). As such, contrary to the authors' conclusion, our Bayesian one-sample *t*-test illustrates that quantitative differences are evident in the scores given by the sample in the AIQ and the SAQOL, thus challenging the potential validity of the tool(s).

Concluding Remarks

Researchers are often required to rely on simple statistical approaches to save time, resources, and help avoid overfitting data (Blanchard et al., 2018; Myung & Pitt, 1997). Here, we make the case that relying on a simple correlation analysis to test the validity of measures can be a questionable approach. In Simulation 1, we illustrate how correlation analysis was previously implemented as a test of validity. However, through Simulation 2, we provide an example of how this correlational approach can fail when examining validity in the case of systematic error. To overcome this limitation, we propose the alternative of using a Bayesian one-sample *t*-test, which, as demonstrated through Simulation 3, can supersede correlation analysis in identifying both cases of nonsystematic and systematic error. It further provides evidence for the null hypothesis of no differences between two tools.

As these simulations show, the proposed Bayesian one-sample *t*-test offers a better validity estimate than correlation analysis within the frequentist approach. However, this Bayesian *t*-test is not meant to replace any existing sophisticated or established validity tests (e.g., specificity and sensitivity analyses; Marchevsky et al., 2020; Stites & Wilen, 2020). Rather, we suggest that simple reliability tests, such as the proposed Bayesian one-sample *t*-test, should be implemented only when relying on the simplest statistical approach can be sufficiently justified (Blanchard et al., 2018; see also Wagenmakers et al., 2010).

Note that additional validity tests do exist as alternatives to the simple correlation analysis. For example, the intraclass correlation coefficient (Shrout & Fleiss, 1979) and the weighted kappa (Cohen, 1960, 1968; Fleiss, 1971) serve as more sensitive validity and reliability measures to correlation. Further, specific measures, such as Krippendorff's alpha, were developed specifically to estimate systematic error (Krippendorff, 1970). One benefit of our proposed Bayesian one-sample *t*-test over these alternatives is that the *t*-test provides a measure of evidence (i.e., BF), allowing us to compare how strongly the evidence supports the existence of a difference between the measurements under test. Subsequent work could investigate how the Bayesian one-sample *t*-test compares to other validity (or reliability) measures such as ICC, weighted kappa, and Krippendorff's alpha.

Here, we proposed the Bayesian one-sample *t*-test based on simulations and an application on a single dataset; this limitation should be discussed. Implementing the Bayesian *t*-test as a validity test in a real-world context might reveal further validity issues not thoroughly discussed here. For example, rescaling or standardization might be required when applying the approach to real data (see Real Data Application). In some cases, this approach might result in similar issues as the ones discussed against using correlation analysis (i.e., failing to identify systematic error). Specifically, some standardization approaches (e.g., *z*-scores) might remove any possible systematic error between the two measurements at-test, thus making the Bayesian one-sample *t*-test approach prone to the same limitations of correlation analysis. Future work could focus on testing the proposed Bayesian one-sample *t*-test in additional real-world applications using different standardization or rescaling approaches to discover possible limitations and strengths.

In conclusion, common practices used in psychological research, such as correlation analysis as a test of validity, can often be misleading. We show here that correlation

analysis provides misleading conclusions regarding validity when there is systematic measurement error. As an alternative, we propose implementing the Bayesian one-sample *t*-test, which supersedes correlation analysis as a test of validity, since it can successfully identify both systematic and nonsystematic error and further provides evidence for the null hypothesis of no difference between two tools.

References

Bandyopadhyay, P. S., & Brittan, G. (2010). Two dogmas of strong objective Bayesianism. *International Studies in the Philosophy of Science, 24*(1), 45–65. https://doi.org/10.1080/02698590903467119

Blanchard, T., Lombrozo, T., & Nichols, S. (2018). Bayesian Occam's razor is a razor of the people. *Cognitive Science, 42*(4), 1345–1359. https://doi.org/10.1111/cogs.12573

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour, 6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Charalambous, M., Phylactou, P., Elriz, T., Psychogios, L., Annoni, J. M., & Kambanaros, M. (2022). Adaptation of the Scenario Test for Greek-speaking people with aphasia: A reliability and validity study. *International Journal of Language & Communication Disorders, 57*(4), 865–880. https://doi.org/10.1111/1460-6984.12727

Charalambous, M., Phylactou, P., Kountouri, A., Serafeim, M., Psychogios, L., Annoni, J. M., & Kambanaros, M. (2022). Adaptation of the Aphasia Impact Questionnaire – 21 into Greek: A reliability and validity study. *Clinical and Translational Neuroscience, 6*(4), Article 24. https://doi.org/10.3390/ctn6040024

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. https://doi.org/10.1037/h0026256

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology, 30*(1), 92–101. http://doi.org/10.5334/irsp.82

Derksen, M. (2019). Putting popper to work. *Theory & Psychology, 29*, 449–465. https://doi.org/10.1177/0959354319838343

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics, 42*(1), 204–223. https://www.jstor.org/stable/2958475

Dienes, Z. (2021a). Obtaining evidence for no effect. *Collabra: Psychology, 7*(1), Article 28202. https://doi.org/10.1525/collabra.28202

Dienes, Z. (2021b). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice, 8*(1), 9–26. https://doi.org/10.1037/cns0000258

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science, 2*(4), 364–377. https://doi.org/10.1177/2515245919876960

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, Article 781. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review, 25*(1), 207–218. https://doi.org/10.3758/s13423-017-1266-z

Drevon, D., Fursa, S. R., & Malcolm, A. L. (2017). Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. *Behavior Modification, 41*(2), 323–339. https://doi.org/10.1177/0145445516673998

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

Fryar, C. D., Carroll, M. D., Gu, Q., Afful, J., & Ogden, C. L. (2021). *Anthropometric reference data for children and adults: United States, 2015–2018*. https://stacks.cdc.gov/view/cdc/100478

Fu, Q., Hoijtink, H., & Moerbeek, M. (2021). Sample-size determination for the Bayesian t-test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods, 53*(1), 139–152. https://doi.org/10.3758/s13428-020-01408-1

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*(1), 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Heck, D. W., & Bockting, F. (2023). Benefits of Bayesian model averaging for mixed-effects modeling. *Computational Brain & Behavior, 6*(1), 35–49. https://doi.org/10.1007/s42113-021-00118-x

Hilari, K., Byng, S., Lamping, D. L., & Smith, S. C. (2003). Stroke and Aphasia Quality of Life Scale – 39 (SAQOL-39) evaluation of acceptability, reliability, and validity. *Stroke, 34*(8), 1944–1950. https://doi.org/10.1161/01.STR.0000081987.46660.ED

Hilari, K., Galante, L., Huck, A., Pritchard, M., Allen, L., & Dipper, L. (2018). Cultural adaptation and psychometric testing of the Scenario Test UK for people with aphasia. *International Journal of Language & Communication Disorders, 53*(4), 748–760. https://doi.org/10.1111/1460-6984.12379

Jaynes, E. T. (2003). *Probability theory: The logic of science* (Vol. 727, . Cambridge University Press.

Jeffreys, H. (1998). *The theory of probability*. Oxford University Press.

Johansson, T. (2011). Hail the impossible: p-values, evidence, and likelihood. *Scandinavian Journal of Psychology, 52*(2), 113–125. https://doi.org/10.1111/j.1467-9450.2010.00852.x

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30*(1), 61–70. https://doi.org/10.1177/001316447003000105

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General, 142*(2), 573–603. https://doi.org/10.1037/a0029146

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology: Series B, 75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Marchevsky, A. M., Walts, A. E., Lissenberg-Witte, B. I., & Thunnissen, E. (2020). Pathologists should probably forget about kappa: Percent agreement, diagnostic specificity and related metrics provide more clinically applicable measures of inter-observer variability. *Annals of Diagnostic Pathology, 47*, Article 151561. https://doi.org/10.1016/j.anndiagpath.2020.151561

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*(1), 79–95. https://doi.org/10.3758/BF03210778

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical*

Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231(694–706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Phylactou, P., Chen, S., Seminowicz, D., & Schabrun, S. (2024). Can we find evidence for the null in a Bayesian t-test? Not unless we reconsider Bayes factor thresholds. *PsyArXiv*. https://doi.org/10.31234/osf.io/kytj7

Phylactou, P., & Konstantinou, N. (2022). Bayesian t-test sample size determination: Reference tables for various Bayes factor thresholds, effect sizes, sample sizes, and variance assumptions. *PsyArXiv*. http://doi.org/10.31234/osf.io/jnp8c

Phylactou, P., Papadatou-Pastou, M., & Konstantinou, N. (2024). *Correlations are not valid for validity*. Open dataset. https://osf.io/6mxzc/

Phylactou, P., Traikapi, A., Papadatou-Pastou, M., & Konstantinou, N. (2022). Sensory recruitment in visual short-term memory: A systematic review and meta-analysis of sensory visual cortex interference using transcranial magnetic stimulation. *Psychonomic Bulletin & Review, 29*, 1594–1624. https://doi.org/10.3758/s13423-022-02107-y

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development, 31*(1), Article e2295. https://doi.org/10.1002/icd.2295

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science, 16*(4), 744–755. https://doi.org/10.1177 %2F1745691620966795

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*(2), 199–223. https://doi.org/10.1037/1082-989X.1.2.199

Schluter, P. J. (2009). A multivariate hierarchical Bayesian approach to measuring agreement in repeated measurement method comparison studies. *BMC Medical Research Methodology, 9*, 1–13. https://doi.org/10.1186/1471-2288-9-6

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Stites, E. C., & Wilen, C. B. (2020). The interpretation of SARS-CoV-2 diagnostic tests. *Med, 1*(1), 78–89. https://doi.org/10.1016/j.medj.2020.08.001

Świaôtkowski, W., & Carrier, A. (2020). There is nothing magical about Bayesian statistics: An introduction to epistemic probabilities in data analysis for psychology starters. *Basic and Applied Social Psychology, 42*(6), 387–412. https://doi.org/10.1080/01973533.2020.1792297

Taherdoost, H. (2016). Validity and reliability of the research instrument: How to test the validation of a questionnaire/survey in a research. *International Journal of Academic Research in Management, 5*(3), 28–36. https://doi.org/10.2139/ssrn. 3205040

Uygun Tunç, D., Tunç, M. N., & Lakens, D. (2023). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. *Theory & Psychology, 33*(3), 403–423. https://doi.org/10.1177/09593543231160112

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology, 60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology, Part II: Example applications with JASP. *Psychonomic Bulletin & Review, 25*(1), 58–76. https://doi.org/10.3758/s13423-020-01798-5

Yaşar, E., Günhan Şenol, N. E., Ertürk Zararsız, G., & Birol, N. Y. (2022). Adaptation of the Aphasia Impact Questionnaire – 21 into Turkish: Reliability and validity study. *Neuropsychological Rehabilitation, 32*(7), 1550–1575. https://doi.org/10.1080/09602011.2021.1917427

History
Received October 13, 2023
Accepted November 7, 2024
Published online December 17, 2024

Open Science
The code used to generate the simulated data and the figures can be publicly accessed through the open science framework repository at https://osf.io/6mxzc/.

ORCID
Phivos Phylactou
https://orcid.org/0000-0002-7333-8761
Marietta Papadatou-Pastou
https://orcid.org/0000-0002-2834-4003
Nikos Konstantinou
https://orcid.org/0000-0003-4531-3636

Dr. Phivos Phylactou
The Gray Centre for Mobility and Activity
Parkwood Institute Main Building
550 Wellington Road, office B2–155B
London, ON, N6C 0A7
Canada
pphylact@uwo.ca