Behav Res (2013) 45:999–1010
DOI 10.3758/s13428-013-0328-z

Statistical criteria for parallel tests: A comparison of accuracy and power

Miguel A. García-Pérez

Published online: 15 February 2013
© Psychonomic Society, Inc. 2013

Abstract Parallel tests are needed so that alternate forms can be applied to different groups or on different occasions, but also in the context of split-half reliability estimation for a given test. Statistically, parallelism holds beyond reasonable doubt when the null hypotheses of equality of observed means and variances across the two forms (or halves) are not rejected. Several statistical tests have been proposed for this purpose, but their performance has never been compared. This study assessed the relative performance (type I error rate and power) of the Student–Pitman–Morgan, Bradley–Blackwood, and Wilks tests of equality of means and variances in the typical conditions surrounding studies of parallelism—namely, integer-valued and bounded test scores with distributions that may not be bivariate normal. The results advise against the use of the Wilks test and support the use of the Bradley–Blackwood test because of its simplicity and its minimally better performance in comparison with the more cumbersome Student–Pitman–Morgan test.

Keywords Classical test theory · Parallel tests · Reliability · Simulation · Statistical tests

Parallel tests are sometimes needed so that alternate forms can be administered either to different groups of respondents or to the same group on separate occasions (e.g., Kronmüller et al., 2008; Werheid et al., 2002). The need for parallel tests arises also on estimation of reliability with the split-half method, which implies partitioning the test at hand into two parallel halves (e.g., Altin & Gençöz, 2009; Joyce et al., 2010; Schmidtke & Metternich, 2009; Woods et al., 2008). Cronbach's alpha is also used rather often for the latter purpose (Hogan, Benjamin, & Brezinski, 2000), but both approaches to reliability estimation have pros and cons (Charter, 2001; Sijtsma, 2009; Thompson, Green, & Yang, 2010). Acknowledging this controversy but not aiming to pursue the discussion further, the present article focuses on comparing statistical criteria to check out parallelism, whether in the context of construction of alternate forms or in the context of split-half reliability estimation.

Raykov, Patelis, and Marcoulides (2011) proposed a latent variable approach that can be used only in quests for the parallelism of at least three measures, and they also discussed the methodological limitations of attempts to check for the parallelism of only two measures. This should not be a problem in the case of reliability estimation, because a three-part (instead of a two-part) partition of any given test is always feasible. But constructing a third form of a test is not always easy. A similar problem arises in the assessment of reliability with the test–retest method, which requires checking out that the two applications of the test were parallel. A third retest would also be necessary to use the method of Raykov et al. in this context. A statistical criterion for examining the parallelism of two measures is thus necessary, despite the difficulties discussed by Raykov et al.: If the two measures turned out to be nonparallel, the correlation between them would not be an estimate of the reliability of the test. It is also remarkable that reliability in empirical studies is routinely reported as the measured correlation between two presumed parallel tests (parallel-form reliability), two presumed parallel halves (split-half reliability), or two presumed parallel applications (test–retest reliability) with no accompanying evidence that parallelism actually held.

Gulliksen (1950, pp. 207–210) described some strategies that increase the likelihood that a partition of a set of items

M. A. García-Pérez (✉)
Departamento de Metodología, Facultad de Psicología,
Universidad Complutense, Campus de Somosaguas,
28223 Madrid, Spain
e-mail: miguel@psi.ucm.es

renders parallel tests, whether this is done to construct two parallel forms of a test or simply to define two halves for split-half reliability estimation. He also discussed statistical criteria for parallel tests. In general, k ($k \geq 2$) tests or partitions are parallel if the true scores are the same on all tests and the error variances are identical. Neither of these conditions can be directly tested empirically (Raykov et al., 2011), but they have implications that permit an indirect examination of parallelism. A testable consequence of parallelism is that the observed scores on the various tests must have means, variances, and intercorrelations that can be regarded as samples from a single multivariate population in which the means, variances, and intercorrelations are identical.

Gulliksen (1950, chap. 14) presented an indirect statistical criterion due to Wilks (1946) that examines this necessary condition for parallelism by simultaneously testing the hypotheses of equal means, equal variances, and equal intercorrelations for $k \geq 2$, but he also pointed out that the $k = 2$ case—in which a single correlation is involved—can be treated with a simple check of equality of means and variances. The $k = 2$ case is the most frequent, if only because of the prevalence of split-half methods in reliability estimation and the rare interest in developing more than two forms of a given test. It should be noted that some computer programs (e.g., TAP; Brooks & Johanson, 2003) report split-half reliability estimates by partitioning a test into even and odd halves or into first and second halves, carrying out no statistical test of parallelism. Then it is the user's responsibility to arrange the items in the data file so that one of these partitioning methods renders parallel halves.

Besides application of the Wilks test with $k = 2$, other tests of equality of two means and variances with related samples have been proposed. For instance, the usual paired-samples t-tests for equality of means and for equality of variances could be applied jointly with a Bonferroni correction to test the compound hypothesis of equality of means and variances. Alternatively, E. L. Bradley and Blackwood (1989) have proposed a rather simple statistic that simultaneously tests for equality of means and variances with paired samples. Given this diversity of approaches, the choice of a statistical test may seem a question of personal preference or convenience, but it is unlikely that all choices are equivalent in terms of accuracy, power, and robustness. Indeed, statistical tests of any kind are derived under usually strong distributional assumptions, and their theoretical behavior is characterized in asymptotic situations. In actual practice, however, statistical tests are used with small samples and with variables that may not be distributed as was assumed in the derivation of the test. But not all alternative statistical tests for the same purpose maintain their asymptotic properties in small-sample conditions, nor are they all equally robust to violation of distributional assumptions.

This work investigated the small-sample properties of the three tests just mentioned, taking into account the characteristics of observed test scores. One of these is that observed scores are discrete and bounded; the second is that observed scores may not be normally distributed and, even if they are, the measurement model does not ensure that observed scores on two actually parallel tests will have a bivariate normal distribution. These two characteristics are quite apparent in empirical distributions of observed test scores (see, e.g., Arostegui, Núñez-Antón, & Quintana, 2007; Arostegui, Padierna, & Quintana, 2010; Torrance et al., 2009) and pose a potential threat to the accuracy of statistical tests designed for use with continuous, unbounded, and normally distributed variables. This study used simulation methods to assess the accuracy (type I error rates) and power (one's complement of type II error rates) of each of the statistical tests described next.

Statistical tests

Let X and Y be normally distributed random variables with means μ_x and μ_y and variances σ_x^2 and σ_y^2, respectively. Also, let ρ_{xy} be the correlation between X and Y, and note that at this point, no assumption needs yet to be made about the form of the bivariate distribution of X and Y. In practice, statistical inference about the means and variances of X and Y requires drawing a sample of n paired observations and computing at least the sample means $\overline{X}$ and $\overline{Y}$, the sample variances s_x^2 and s_y^2,[1] and the sample product–moment correlation r_{xy}.

Separate statistical tests for the equality of the means or the variances of two variables are well known, and using both of them for a simultaneous test of both null hypotheses requires the use of a Bonferroni correction. Thus, for a size-α test, the joint hypothesis of equal means and variances is rejected if at least one of the two tests is rejected at an $\alpha^* = \alpha/2$ level. The null hypothesis of equality of means is tested through the well-known statistic

$$T_m = \frac{\overline{X} - \overline{Y}}{\sqrt{s_x^2 + s_y^2 - 2r_{xy}s_x s_y}/\sqrt{n-1}}, \tag{1}$$

which has a t distribution with $n - 1$ degrees of freedom if the null hypothesis of equality of means is true. On the other hand, the null hypothesis of equality of variances is tested through the well-known statistic (Morgan, 1939; Pitman, 1939)

$$T_v = \frac{\sqrt{n-2}\left(s_x^2 - s_y^2\right)}{2s_x s_y\sqrt{1 - r_{xy}^2}}, \tag{2}$$

[1] Sample variances are assumed in this article to be uncorrected for bias (i.e., computed with a denominator of n rather than $n - 1$).

which has a t distribution with $n-2$ degrees of freedom if the null is true.

An alternative to application of these two tests with a Bonferroni correction (which we will refer to as the Student–Pitman–Morgan test) is the Bradley–Blackwood test (E. L. Bradley & Blackwood, 1989), which stems from the observation that testing simultaneously for equality of the means and variances of X and Y is equivalent to testing for null slope and intercept in the regression of $D = X - Y$ on $S = X + Y$. If X and Y have a bivariate normal distribution, this is achieved simultaneously through the statistic

$$F = \frac{\left(\sum_{i=1}^{n} D_i^2 - SSE\right)/2}{SSE/(n-2)}, \tag{3}$$

where SSE is the residual sum of squares from the regression of D on S. If $\mu_x = \mu_y$ and $\sigma_x^2 = \sigma_y^2$, this test statistic is distributed F with 2 and $n-2$ degrees of freedom. It should be noted at this point that the so-called "Bland–Altman method" in which $X - Y$ is plotted against $(X + Y)/2$ to measure the agreement between X and Y finds justification in Morgan's (1939) and Pitman's (1939) demonstration that the covariance of $X - Y$ and $X + Y$ is $\sigma_x^2 - \sigma_y^2$.

Finally, the Wilks test for $k = 2$ as discussed by Gulliksen (1950) tests equality of means and variances through the statistic

$$L_{mvc}\frac{s_x^2\, s_y^2\left(1-r_{xy}^2\right)}{s^2\left(1+r_{xy}\right)\left(s^2\left(1-r_{xy}\right)+v\right)}, \tag{4}$$

where $s^2 = \left(s_x^2 + s_y^2\right)/2$ and $v = \left(\overline{X} - \overline{Y}\right)^2/2$. Then, $-n\log(L_{mvc})$ has an approximate chi-square distribution with 2 degrees of freedom, and Wilks (1946, p. 272) noted that the approximation should be adequate for $k \le 5$ and $n \ge 50$.

Method

Following the classical measurement model, a sample of n true scores T was drawn from a population with mean μ_t and variance σ_t^2, and observed scores X and Y were obtained from them through addition of independent and normally distributed random errors E_x and E_y such that $X = T + E_x$ and $Y = T + E_y$. In all cases, E_x had mean 0 and variance σ_e^2, and it is useful to recall at this point that the reliability of X is $\rho_{xx'} = \sigma_t^2/\left(\sigma_t^2 + \sigma_e^2\right)$. In studies of type I error rates, E_y had the same mean and variance as E_x on a condition by condition basis. In studies of power, E_y had mean $a\sigma_e\sqrt{2}$ and variance $\sigma_e^2\,(b - \rho_{xx'})/(1 - \rho_{xx'})$, with $a > 0$ and $b > 1$. These relations were chosen so that a and b respectively represent effect sizes for mean differences and variance ratios, as is discussed next.

The effect size for the mean difference between X and Y is $|\mu_x - \mu_y|/\sigma_{x-y}$ (see, e.g., Faul, Erdfelder, Lang, & Buchner, 2007, Table 3). When $b = 1$, effect size equals a because $\mu_x = \mu_t$, $\mu_y = \mu_t + a\sigma_e\sqrt{2}$, and $\sigma_{x-y} = \sigma_e\sqrt{2}$. On the other hand, a definition of effect size for tests of related variances does not seem to exist, but consider for convenience the definition used for the case of independent variances—namely, σ_y^2/σ_x^2 (Faul et al., 2007, Table 9). Then, regardless of the value of a, it can easily be seen that effect size is given by b because $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$ and $\sigma_y^2 = \sigma_t^2 + \sigma_e^2\,(b - \rho_{xx'})/(1 - \rho_{xx'})$. Note, then, that $a = 0$ and $b = 1$ render identically distributed E_x and E_y (i.e., strict parallelism), that $a = 0$ with $b > 1$ implies equal means for X and Y but different variances (i.e., τ-equivalence), and that $a \neq 0$ with $b = 1$ implies equal variances for X and Y but different means.

Normal and uniform deviates required to generate values for T, E_x, and E_y with prescribed parameters were drawn through NAG subroutines G05DDF and G05DAF (Numerical Algorithms Group, 1999), respectively. The set of simulation conditions resulted from the factorial combination of several forms of the distribution of true scores, several reliability levels, four options regarding rounding and bounding of observed scores, and several sample sizes. The particular levels that were used along these dimensions are described next.

True scores T were generated to have symmetrical or asymmetrical distributions. Symmetrically distributed true scores were drawn either from normal or from uniform distributions with parameters that yielded the same mean and variance irrespective of the form of the distribution. For a test yielding observed scores in the range between 0 and X_{max}, normally distributed true scores were generated to have mean $\mu_t = X_{max}/2$ and variance $\sigma_t^2 = (X_{max}/6)^2$, so that the range $[0, X_{max}]$ encroaches six standard deviations and, thus, the distribution of observed scores does not show the strong ceiling or floor effects that would arise from poorly constructed tests. A matching condition was defined with true scores that are uniformly distributed in the range between $X_{max}(3 - \sqrt{3})/6$ and $X_{max}(3 + \sqrt{3})/6$, which also renders true scores with mean $X_{max}/2$ and variance $(X_{max}/6)^2$ (Evans, Hastings, & Peacock, 2000, p. 171). The set of conditions in this respect covered values of X_{max} from 15 (implying a relatively short test of 30 items in the context of split-half reliability studies) to 45 (a moderately long test) items in steps of 5 items, which thus represent conditions in which mean true score varies from 7.5 to 22.5 and true score variance varies from 6.25 to 56.25. Asymmetrically distributed true scores were drawn from folded normal distributions (Leone, Nelson, & Nottingham, 1961), whose form resembles the empirical distribution of observed scores from tests showing moderate ceiling or floor effects. The parameters of folded normal distributions varied in preliminary

simulation studies so as to cover a relatively broad range of realistic distributions, but the results did not differ meaningfully. For this reason, results will be presented here only for folded distributions with $\mu = 7X_{max}/20$ and $\sigma^2 = (11X_{max}/60)^2$ (see Fig. 4c below for an illustration) and involving the same range of X_{max} described above for symmetric distributions.

Error variance σ_e^2 varied so that the reliability of observed X scores ranged from .65 to .95 in steps of .03 units. Hence, $\sigma_e^2 = \sigma_t^2(1 - \rho_{xx'})/\rho_{xx'}$ varied between $\sigma_t^2/19$ (when $\rho_{xx'} = .95$) and $7\,\sigma_t^2/13$ (when $\rho_{xx'} = .65$). In studies of type I error rates, the variance of E_y matched the variance of E_x on a condition by condition basis. In studies of power, the variance of E_y varied as described earlier, with a ranging between 0 and 1.5 and b ranging between 1 and 2.5. When $b = 1$, the reliability of observed Y scores was identical to that of observed X scores despite variations in the means of X and Y produced by a; when $b > 1$, the reliability of Y scores varied inversely with b, as is evident by the role played by b in the expression presented above (i.e., the variance of E_y increases with b).

Observed scores X and Y generated by this process are continuous, real-valued, and unbounded even with uniformly distributed true scores because these are then corrupted by normally distributed errors. In actual practice, observed scores are discrete and integer-valued and have hard bounds at the minimum and maximum scores attainable in the test. Therefore, all analyses were carried out on the untouched samples of observed scores initially generated (which are thus close to meeting the distributional assumptions of the test statistics) and also on the samples that resulted from rounding each observed score to the nearest integer and/or replacing it with boundary values (i.e., either by 0 or by X_{max}) if they were off bounds.

Finally, the size n of the sample of examinees varied between 50 and 350 in steps of 50, since samples smaller than 50 units are never used in studies of parallelism.

A simulation condition was defined as a particular combination for the options described—for example, a sample of $n = 150$ examinees responding to two parallel forms (i.e., the mean and variance of E_x and E_y are the same) of a test in which observed scores range between 0 and $X_{max} = 30$ (thus, $\mu_t = 15$ and $\sigma_t^2 = 25$), with normally distributed true scores, when reliability is .8 (thus, $\sigma_e^2 = 6.25$), and where observed scores are rounded and bounded. For each simulation condition, 100,000 replicates were drawn, and all statistical tests described in the preceding section were applied to the data from each replicate. Then the empirical accuracy (or power, as applicable) of a given statistical test was defined as the proportion of cases in which the null hypothesis of equality of means and variances was rejected. Two-tailed tests were considered for the statistics in Eqs. 1 and 2; the tests based on the statistics in Eqs. 3 and 4 are right-tailed by definition. Accuracy was evaluated at nominal test sizes $\alpha \in \{.10, .05, .01\}$; power was evaluated at $\alpha = .05$. Because error rates computed from 100,000 replicates can be taken at face value,[2] no subsequent statistical analyses were performed, and all comparisons were carried out in terms of discrepancies between empirical and nominal error rates.

Results

Before describing the results in detail, it is useful to look at the sampling distributions of all four test statistics when the null hypothesis is true and all assumptions hold or almost hold. Figure 1 shows empirical sampling distributions (histograms) of the test statistics in Eqs. 1–4 along with their asymptotic distributions (curves) for the case of samples of $n = 50$ examinees from a population in which true scores are normally distributed with $\mu_t = 15$ and $\sigma_t^2 = 25$ (i.e., $X_{max} = 30$), reliability is .65 (i.e., $\sigma_e^2 = 13.46$), and observed scores are neither rounded nor bounded. Also printed in each panel are the empirical percentages of cases (across all 100,000 replicates) in which the statistics exceeded the critical limits for size-.1, size-.05, and size-.01 tests (two-sided or one-sided, as applicable). Thus, for the test statistic in Eq. 1 (i.e., a t-test for equality of means), the critical limits for two-tailed tests render empirical type I error rates of 10.07 % (5.00 + 5.07), 5.09 % (2.53 + 2.56), and 0.99 % (0.50 + 0.49), respectively, for nominal test levels of 10 %, 5 %, and 1 % (see Fig. 1a). Two-tailed tests for equality of variances through the statistic in Eq. 2 render similarly accurate type I error rates of 10.01 %, 5.01 %, and 1 % (see Fig. 1b). The empirical type I error rates of the Student–Pitman–Morgan test (i.e., the previous two tests applied together with a Bonferroni correction) naturally remain also accurate at 9.84 %, 4.99 %, and 1.00 %, respectively, for nominal significance levels of 10 %, 5 %, and 1 % (results not shown in Fig. 1). On the other hand, right-tailed tests of the null hypothesis of equality of means and variances through the statistic in Eq. 3 are also quite accurate (see Fig. 1c). In contrast, the Wilks test (Fig. 1d) appears much too conservative, with empirical type I error rates that are nearly half as large as they should be.

Figure 2 shows that accuracy deteriorates under rounding and bounding—that is, when simulated observed scores are adjusted to meet the empirical constraint that they take integer values between 0 and X_{max}, as would actually be the case on administration of actual tests to real examinees.

[2] With this number of replications, error rates are approximately estimated to within $\pm 1.96\sqrt{\alpha(1-\alpha)/10^5}$. Use of the more adequate Score confidence interval (García-Pérez, 2005) renders a similar range at sample sizes as large as this.

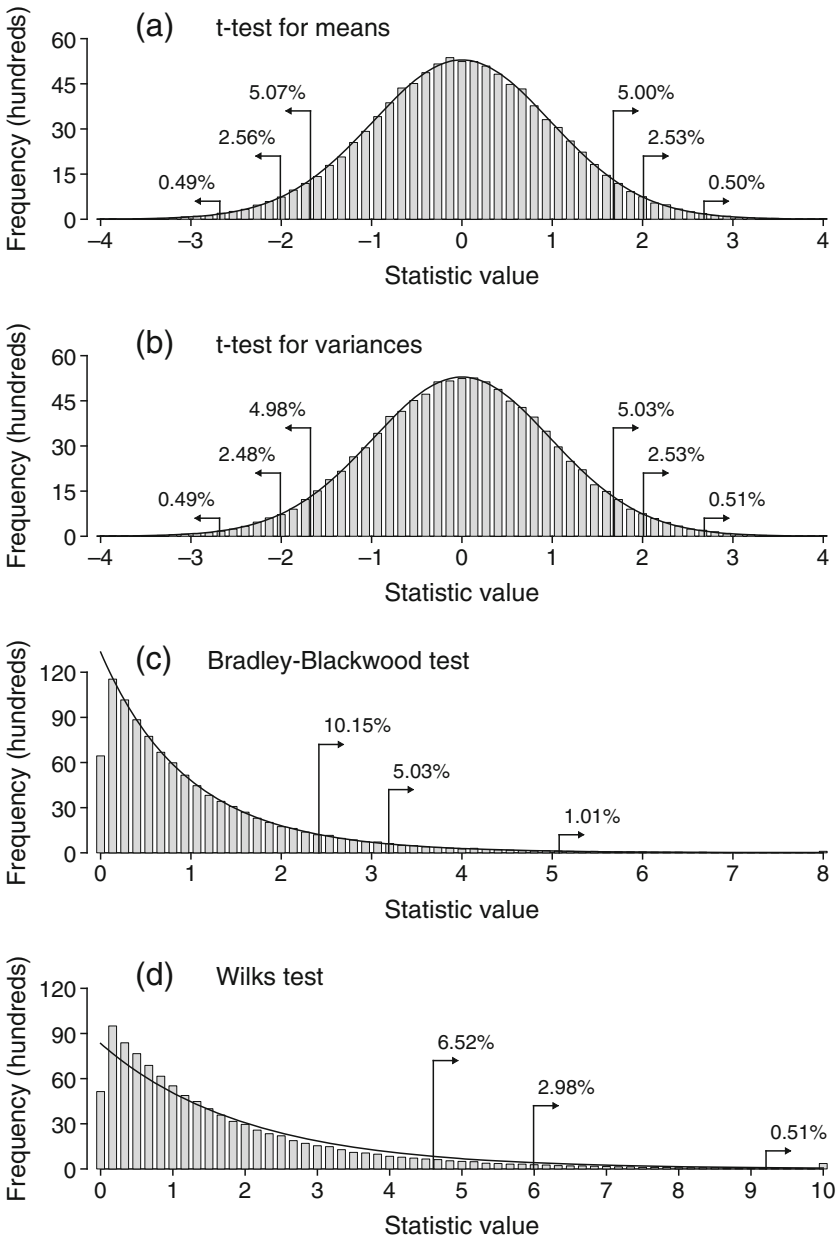

Fig. 1 Sampling distributions and empirical rejection rates of four test statistics when the null hypothesis is true and scores are continuous, real-valued, and unbounded

In these conditions, only the *t*-test for equality of means remains accurate when considered in isolation (see Fig. 2a); all the remaining statistical tests turn somewhere between meaningfully to overly conservative (see Fig. 2b–d). As a consequence of the inaccuracy of the *t*-test for equality of variances (see Fig. 2b), the Student–Pitman–Morgan test has deflated empirical type I error rates at 8.70 %, 4.34 %, and 0.80 %, respectively, for nominal significance levels of 10 %, 5 %, and 1 %. It should nevertheless be pointed out that test sizes that are within 20 % of the nominal size (i.e., actual sizes no lower than 8 %, 4 %, or 0.8 %, respectively, for nominal test sizes of 10 %, 5 %, or 1 %) are slightly above J. V. Bradley's (1978) "fairly stringent" criterion for robustness (±10 %), but they are acceptable by Robey and Barcikowski's (1992) "intermediate criterion" (±25 %). From this perspective, the Bradley–Blackwood and Student–Pitman–Morgan tests, unlike the Wilks test, are robust, although they are certainly subject to improvement.

Type I error rates

Figure 3 gives a broader picture of how empirical type I error rate varies for each test statistic with examinee sample size, $X_{\max}$ (and, hence, score mean and variance), form of the distribution of true scores, and rounding and bounding of observed scores. All results plotted in Fig. 3 represent conditions in which reliability was $\rho_{xx'} = \rho_{yy'} = .65$ (the lowest value included in our simulations). Consider first the left panel of Fig. 3a, for normally distributed true scores with mean 7.5 and variance 6.25 (given that $X_{\max} = 15$) and

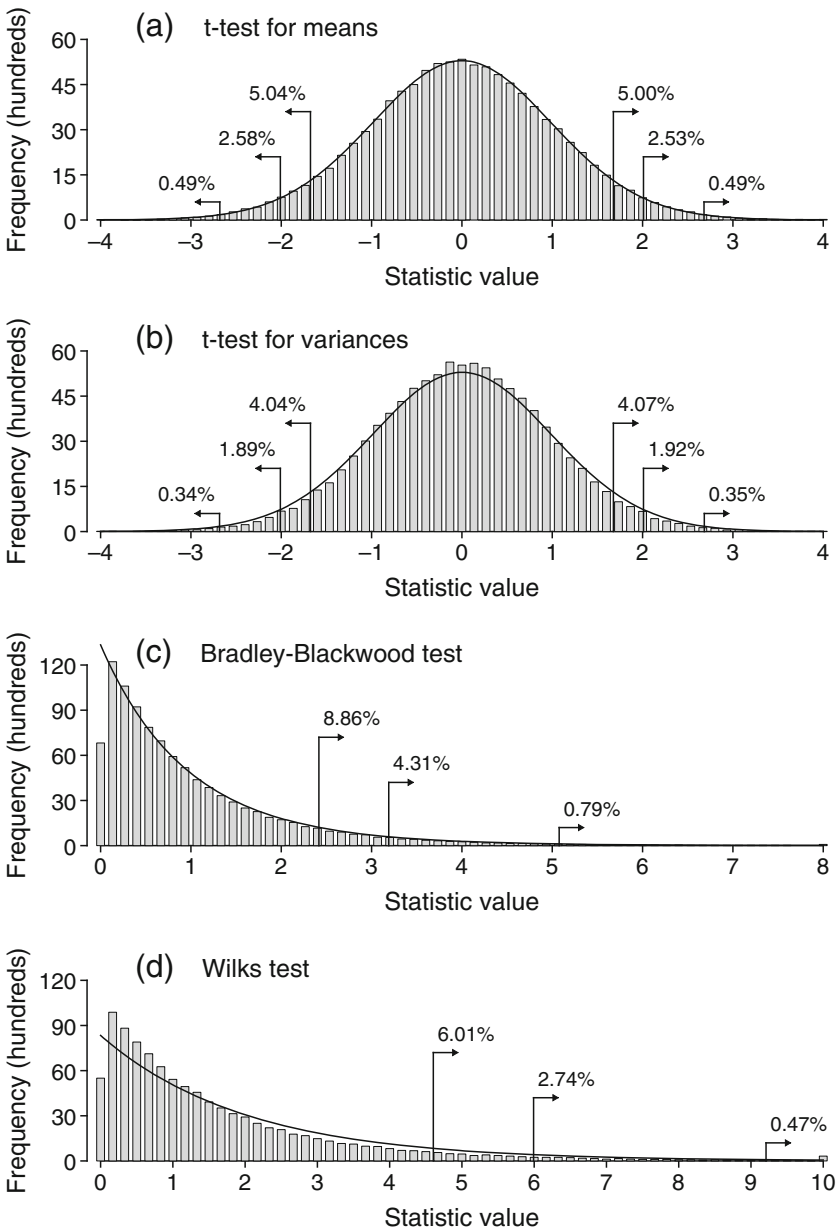

Fig. 2 Sampling distributions and empirical rejection rates of four test statistics when the null hypothesis is true and scores are discrete, integer-valued, and bounded

untouched (i.e., continuous and real-valued) observed scores. It is quite apparent that the Bradley–Blackwood test (rightward-pointing triangles) renders the most accurate error rates at any nominal size (solid symbols, .1 size; gray symbols, .05 size; open symbols, .01 size), although the Student–Pitman–Morgan test (circles) is also very accurate. On the other hand, the Wilks test (leftward-pointing triangles) is very conservative, yielding empirical type I error rates that are about half their nominal rate and, therefore, outside the 20 % tolerance limit. This pattern of results is identical in the right panel of Fig. 3a (for $X_{max} = 45$ so that true scores have a mean of 22.5 and a variance of 56.25), and also for uniformly distributed true scores (Fig. 3b) and folded normal true scores (Fig. 3c) with either score range. Results for intermediate values of X_{max} were also identical and are not shown graphically. Interestingly, then, the form of the distribution of true scores (even if it shows common asymmetries in empirical score distributions) does not have any implication for the accuracy of these statistical tests.

Figures 3d–f show analogous results for conditions differing only in that observed scores were rounded and bounded. Quite clearly, all tests turn conservative in these circumstances (but still within the 20 % tolerance limit, except for the Wilks test), and somewhat more when the distribution of test scores is normal or folded normal (Fig. 3d, f) than when it is uniform (Fig. 3e). In either case, whether the score range is narrow ($X_{max} = 15$; left column) or broad ($X_{max} = 45$; right column) is again inconsequential. Simulations in which observed scores were either only rounded or only bounded revealed that the actual cause of

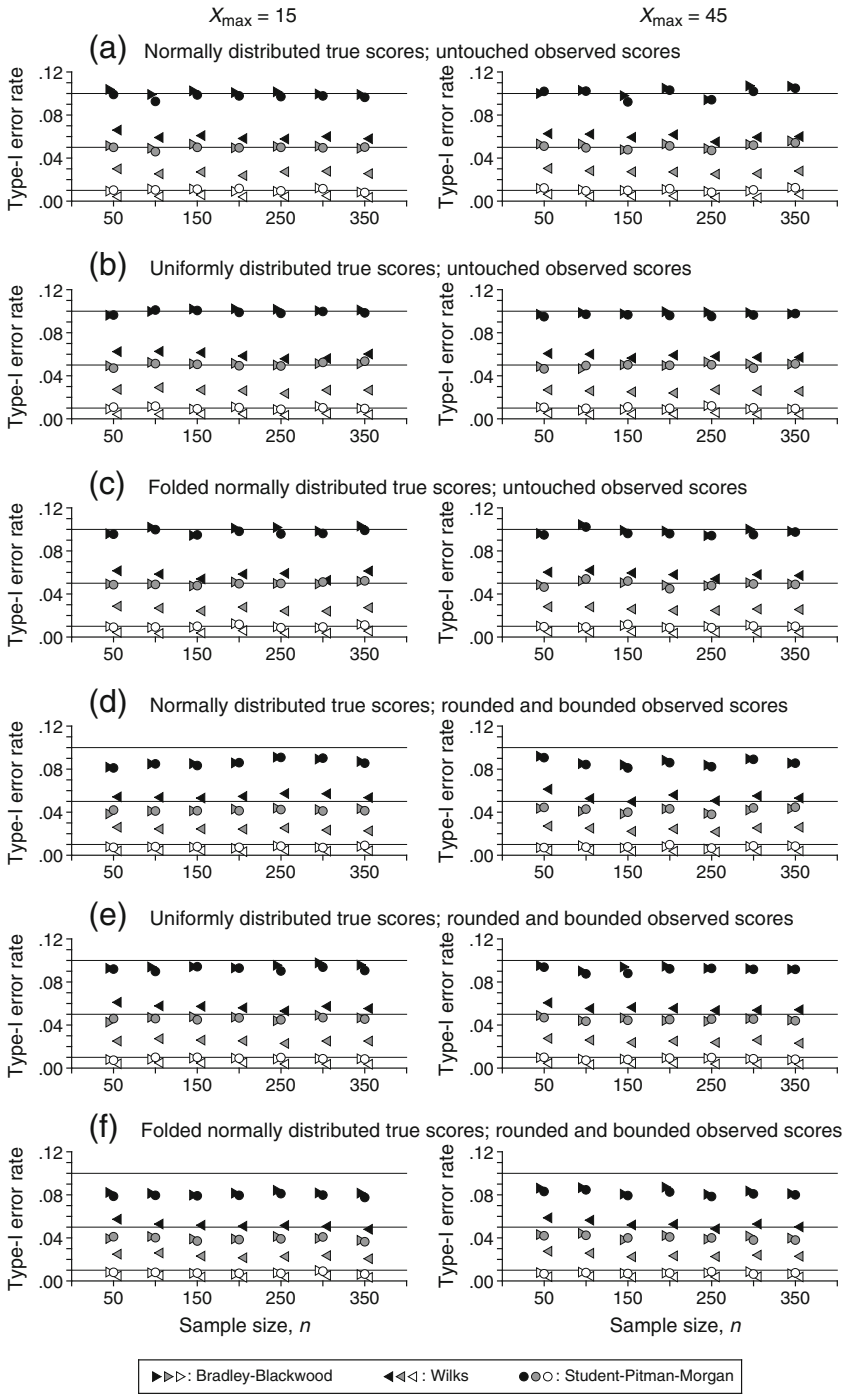

Fig. 3 Type I error rate of three test statistics for equality of means and variances as a function of sample size for different score ranges (columns) and distributions of scores (rows)

the deflated type I error rates was bounding. Bounding alters observed scores more often when true scores are normally (or folded normally) distributed than when they are uniformly distributed (see Fig. 4), and this is the reason for the increased deflation of type I error rates with normal or folded normal distributions of true scores, as compared with uniform distributions (Fig. 3d, f, as compared with Fig. 3e).

Results presented in Fig. 3 arise when reliability is .65. Figure 5 shows results as a function of reliability when X_{max} = 45 and n = 100; results for other score ranges and sample sizes were indistinguishable, as was the case in Fig. 3. When observed scores are untouched, the type I error rates of Bradley–Blackwood tests (rightward-pointing triangles) and Student–Pitman–Morgan tests (circles) do not change in any meaningful respect as reliability increases, whether true scores are normally (Fig. 5a), uniformly (Fig. 5b), or folded normally (Fig. 5c) distributed. In contrast, the inaccuracy of the Wilks test (leftward-pointing triangles) further deteriorates as reliability increases. For rounded and bounded observed scores

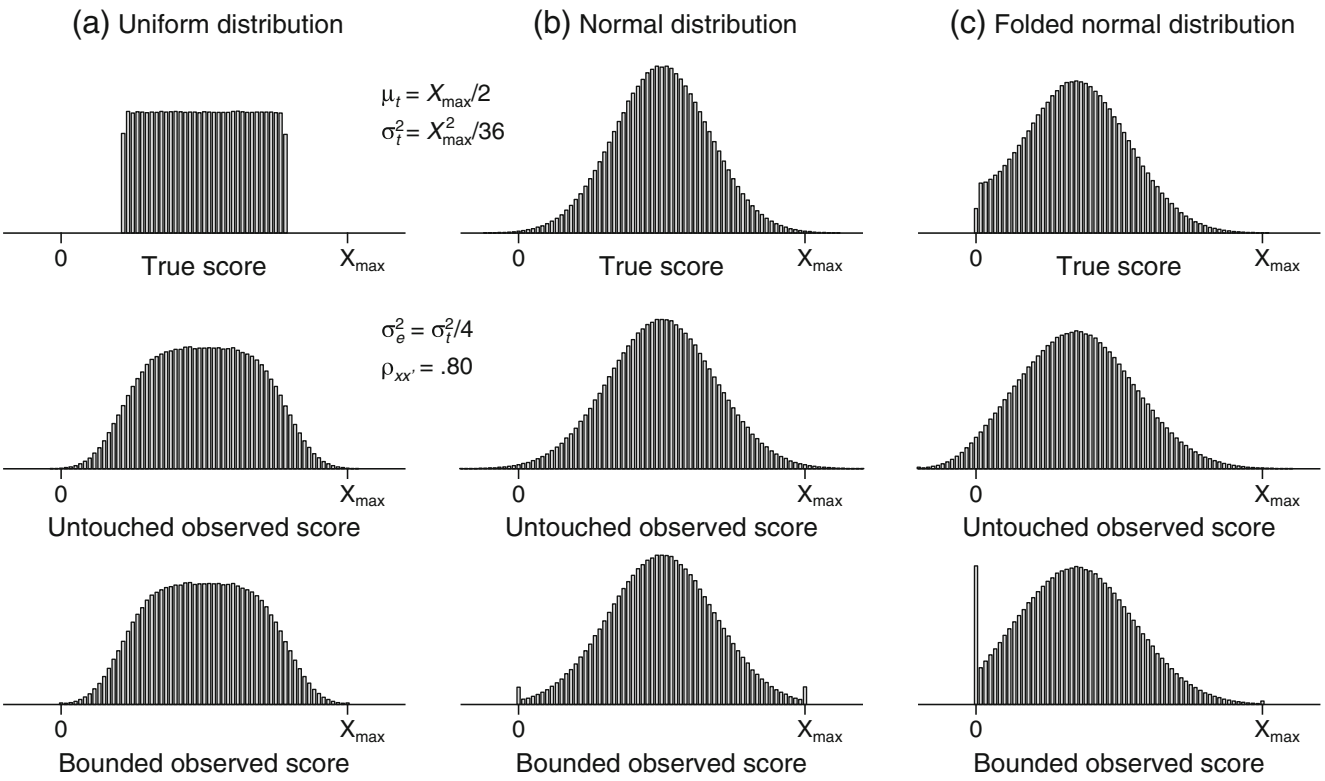

Fig. 4 Distributions of true scores (top row), real-valued and unbounded observed scores (center row), and integer-valued and bounded observed scores (bottom row)

(Fig. 5d, f), the accuracy of Bradley–Blackwood tests and Student–Pitman–Morgan tests improves as reliability increases. This is a natural consequence of the fact that the original distributions of (untouched) observed scores are narrower as reliability increases, so that bounding changes scores less often. Under conditions of rounding and bounding, the accuracy of the Wilks test also deteriorates as reliability increases.

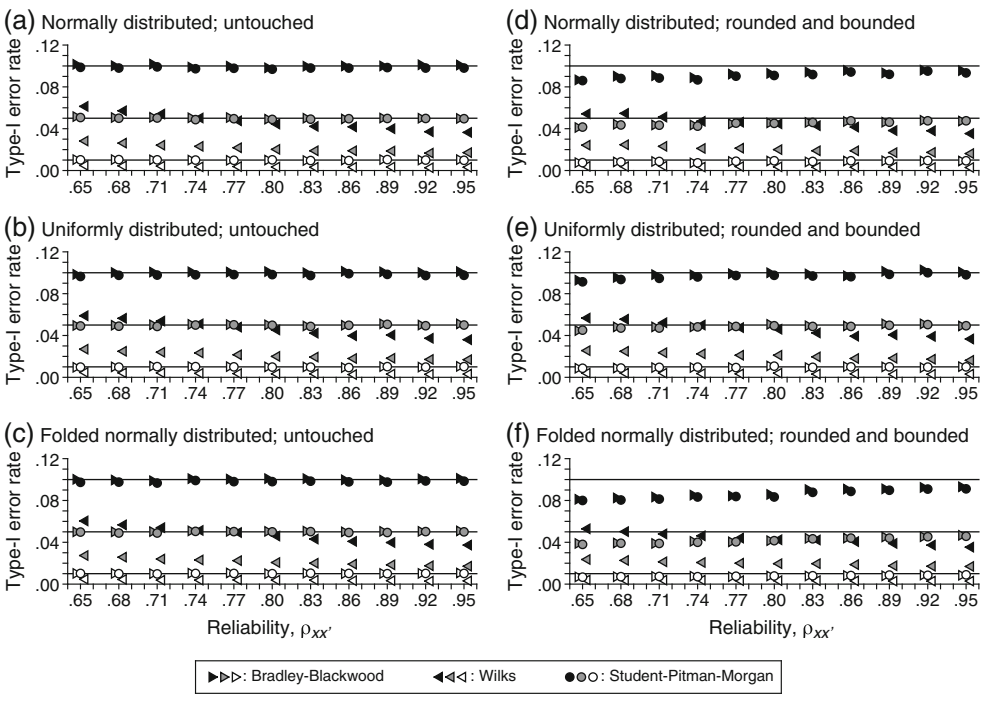

Fig. 5 Type I error rate of three test statistics for equality of means and variances as a function of reliability for different distributions of scores (rows)

Power

A comparison of power curves for different score ranges (X_{max} between 15 and 45) and form of distribution of true scores (normal, uniform, or folded normal) revealed no meaningful differences, and therefore, results will be presented only for X_{max} = 15 and normal distributions of true scores.

Figure 6 shows power curves as a function of effect size a when b = 1—that is, when observed scores X and Y differ in mean (with $\mu_y > \mu_x$) but not in variance. Each panel displays results for a given reliability level (columns) and sample size (rows) for untouched test scores (solid symbols) and rounded and bounded test scores (open symbols). Despite expected variations in power with sample size, two characteristics stand out. One is the lack of differences between Bradley–Blackwood (rightward-pointing triangles) and Student–Pitman–Morgan (circles) tests, which shows in that data points representing these tests are superimposed; the power of the Wilks test (leftward-pointing triangles) is somewhat lower throughout. The second characteristic is the minimal effect that rounding and bounding has, which shows in that solid and open symbols are generally superimposed, although power is slightly lower with rounding and bounding when reliability is very high (rightmost column).

It may seem surprising that power curves differ across columns despite the fact that the horizontal axis represents a standardized measure of effect size that should equalize results across changes in the variance of X and Y. Although this is true in general, it should be remembered that observed scores X and Y do not have a bivariate distribution with a fixed correlation across the changes in variance represented in different columns of Fig. 6. Indeed, and as a result of the classical measurement model, the correlation between X and Y is given by the attenuation formula $\rho_{xy} = \sqrt{\rho_{xx'}}\sqrt{\rho_{yy'}}$. Since $\rho_{xx'} = \rho_{yy'}$ in Fig. 6 and they vary across columns as indicated at the top, ρ_{xy} varies across columns just as reliability does. Variations in power with reliability must, then, be understood as a consequence of variations in the correlation between observed X and Y scores.

Figure 7 shows power curves as a function of the binary logarithm of effect size b for a = 0—that is, when observed scores X and Y differ in variance (with $\sigma_y^2 > \sigma_x^2$) but not in

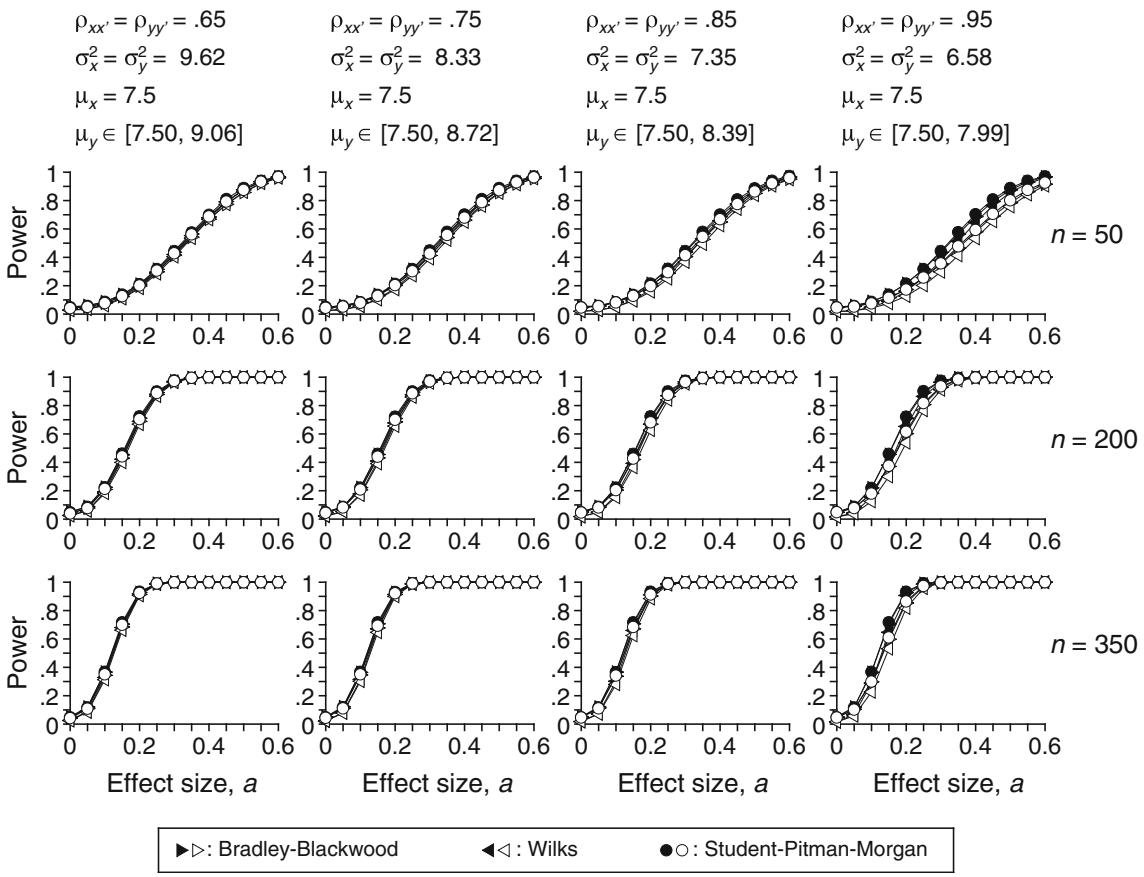

Fig. 6 Power of three test statistics for equality of means and variances as a function of effect size (relative difference between means) for different reliabilities (columns) and sample sizes (rows)

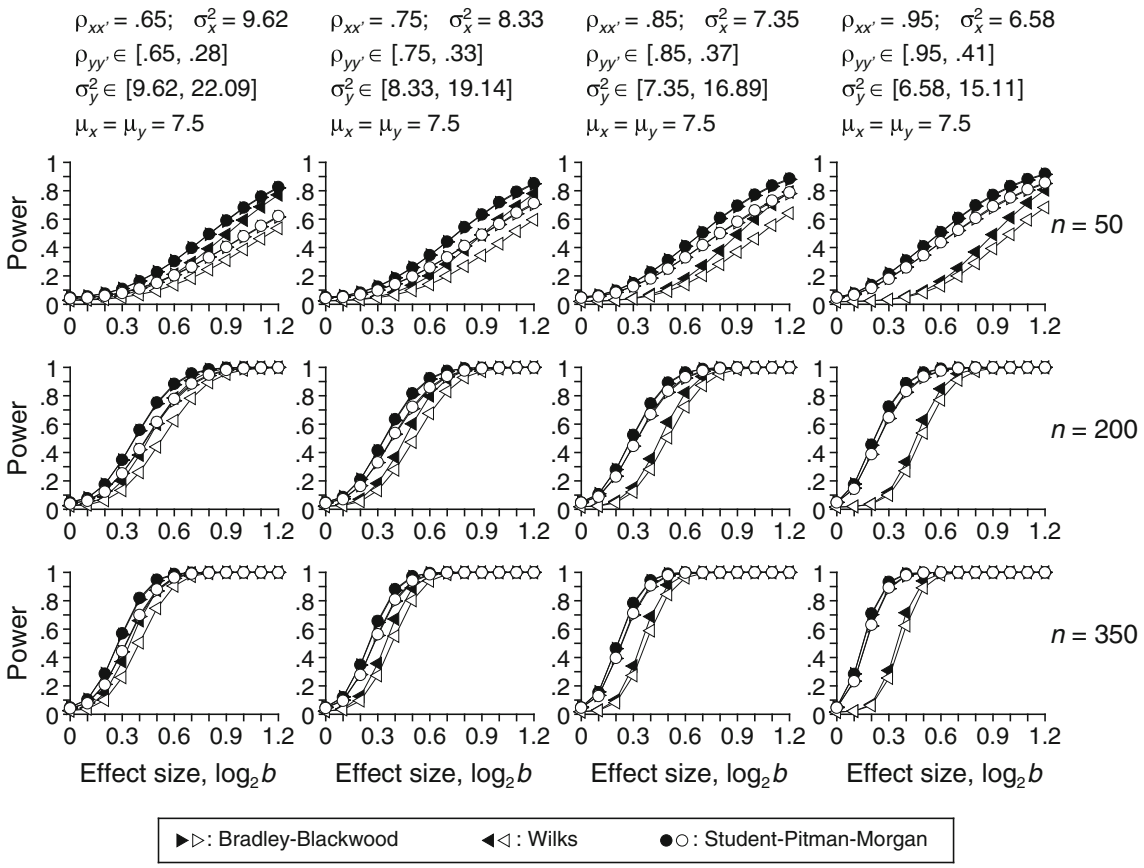

Fig. 7 Power of three test statistics for equality of means and variances as a function of effect size (ratio of variances) for different reliabilities (columns) and sample sizes (rows)

mean. As compared with the previous case, the power of the Wilks test (leftward-pointing triangles) is noticeably lower in these conditions, while Bradley–Blackwood and Student–Pitman–Morgan tests have nearly identical power. Also, power deteriorates meaningfully when observed scores are rounded and bounded (open symbols). This is only a natural consequence of the fact that the reliability of Y scores decreases (and, hence, their variance increases) as b increases, so that rounding and bounding produces increasingly stronger spikes at the boundaries of the score range such as those shown in the bottom panel of Fig. 4b.

Variations in power curves across columns in this case are, in principle, a consequence of the fact that the horizontal axis does not represent a standardized measure of effect size. But variations in correlation between X and Y are also involved here. For instance, from values given at the top of each column in Fig. 7, the rightmost data points (for $\log_2 b = 1.2$) in the panels on the left column imply $\rho_{xy} = \sqrt{.65}\,\sqrt{.28} = .427$; in contrast, the rightmost data points in the panels on the right column imply $\rho_{xy} = \sqrt{.95}\,\sqrt{.41} = .624$. The effects of correlation on power can be best appreciated in the top row of Fig. 7 and suggest that a standardized measure of effect size should take correlation into account.

Discussion

Results presented in Figs. 3, 5, 6 and 7 and discussed in the preceding section can be summarized as follows. The Wilks test is not advisable under any circumstances when $k = 2$, whereas the accuracy and power of Bradley–Blackwood and Student–Pitman–Morgan tests are virtually identical, with only a minimal advantage of the former as regards type I error rates. If test scores were real-valued and unbounded, Bradley-Blackwood and Student–Pitman–Morgan tests would maintain their nominal sizes for $n > 50$ (the smallest sample size used in our simulations, which is usually exceeded in studies of parallelism), whether scores were symmetrically distributed (from normal or uniform distributions) or asymmetrically distributed (from folded normal distributions) within the limits typically observed in empirical test score distributions. With (inescapably) bounded test scores, both tests become slightly and equally conservative, although their actual test size remains well within 20 % of the nominal size, something that is usually regarded as acceptable (García-Pérez & Núñez-Antón, 2009; Robey & Barcikowski, 1992; Serlin, 2000; Serlin & Harwell, 2004). Because the conservatism of both tests depends on the extent to which the distribution of observed scores is curtailed by the hard bounds of minimal and maximal scores, cautious

users should seek evidence of this characteristic in a scatterplot of observed scores, if only to gain subjective confidence on the results of these statistical tests.

The Pitman test for equality of related variances (which is one of the components of the Student–Pitman–Morgan test) has been reported to be nonrobust to certain deviations from normality (McCulloch, 1987; Wilcox, 1990), but our results show that it is more robust to nonnormality (at least for uniform distributions and folded normal distributions) than it is to the bounding of otherwise continuous and normally distributed values, even when bounding affects score distributions as little as Fig. 4 indicates. The same holds for the Bradley–Blackwood test, whereas the *t*-test for equality of means (the other component of the Student–Pitman–Morgan test) is certainly much more robust, perhaps as a consequence of the applicability of the central limit theorem (which certainly does not help to free tests for equality of variances from their strong distributional assumptions).

A perfect solution to the problem of statistically assessing parallelism with prescribed accuracy does not seem to exist, but Bradley–Blackwood and Student–Pitman–Morgan tests are satisfactory solutions, and both of them have adequate power to detect differences in means or variances. Both statistical tests are recommended for the assessment of whether two tests (parallel forms), two halves (split-half), or two applications of a test (test–retest) are parallel.

The *t*-test for related means can be computed with any statistical software package, but neither of the two other tests (i.e., the Morgan–Pitman *t*-test for related variances, which is needed for a full Student–Pitman–Morgan test, or the Bradley–Blackwood test) appears to be available in any of the software packages more widely used. Nevertheless, the statistics that need to be introduced in Eq. 2 for the Morgan–Pitman *t*-test or in Eq. 3 for the Bradley–Blackwood test can be easily obtained with any software package. For instance, the standard deviations, variances, and correlation needed for Eq. 2 are straightforwardly obtained, and note that Eq. 2 has the same form whether variances and standard deviations are or are not corrected for bias. On the other hand, the value of the residual sum of squares needed for Eq. 3 is reported in the summary ANOVA table for the regression of *D* on *S*, whereas the value of $\sum_{i=1}^{n} D_i^2$ can easily be obtained from the variance of *D* reported by the software that was used. Typically, variances reported by statistical software are corrected for bias, and hence, recovering $\sum_{i=1}^{n} D_i^2$ from them amounts to computing $(n-1)\widetilde{s}_d^2 + n\overline{D}^2$, where $\overline{D}$ is the reported sample mean and $\widetilde{s}_d^2$ is the reported (unbiased) sample variance; if the reported variance is not corrected for bias, $\sum_{i=1}^{n} D_i^2$ is recovered as $n\left(s_d^2 + \overline{D}^2\right)$ instead.

Acknowledgements This research was supported by grant PSI2009-08800 from Ministerio de Ciencia e Innovación (Spain).

References

Altin, M., & Gençöz, T. (2009). Psychopathological correlates and psychometric properties of the White Bear Suppression Inventory in a Turkish sample. *European Journal of Psychological Assessment, 25*(1), 23–29. doi:10.1027/1015-5759.25.1.23

Arostegui, I., Núñez-Antón, V., & Quintana, J. M. (2007). Analysis of the short form-36 (SF-36): The beta-binomial distribution approach. *Statistics in Medicine, 26*(6), 1318–1342. doi:10.1002/sim.2612

Arostegui, I., Padierna, A., & Quintana, J. M. (2010). Assessment of HRQoL in patients with eating disorders by the beta-binomial regression approach. *International Journal of Eating Disorders, 43*(5), 455–463. doi:10.1002/eat.20713

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x

Bradley, E. L., & Blackwood, L. G. (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician, 43*(4), 234–235. doi:10.1080/00031305.1989.10475665

Brooks, G. P., & Johanson, G. A. (2003). TAP: Test Analysis Program. *Applied Psychological Measurement, 27*(4), 303–304. doi:10.1177/0146621603027004007

Charter, R. A. (2001). It is time to bury the Spearman–Brown "prophecy" formula for some common applications. *Educational and Psychological Measurement, 61*(4), 690–696. doi:10.1177/00131640121971446

Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical distributions* (3rd ed.). New York: Wiley.

Faul, F., Erdfelder, E., Lang, A.-G., &Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. doi:10.3758/BF03193146

García-Pérez, M. A. (2005). On the confidence interval for the binomial parameter. *Quality and Quantity, 39*(4), 467–481. doi:10.1007/s11135-005-0233-3

García-Pérez, M. A., & Núñez-Antón, V. (2009). Statistical inference involving binomial and negative binomial parameters. *Spanish Journal of Psychology, 12*(1), 288–307.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*(4), 523–531. doi:10.1177/00131640021970691

Joyce, A. S., Adair, C. E., Wild, T. C., McDougall, G. M., Gordon, A., Costigan, N., et al. (2010). Continuity of care: Validation of a self-report measure to assess client perceptions of mental health service delivery. *Community Mental Health Journal, 46*(2), 192–208. doi:10.1007/s10597-009-9215-6

Kronmüller, K.-T., Saha, R., Kratz, B., Karr, M., Hunt, A., Mundt, C., et al. (2008). Reliability and validity of the Knowledge about Depression and Mania Inventory. *Psychopathology, 41*(2), 69–76. doi:10.1159/000111550

Leone, F. C., Nelson, L. S., & Nottingham, R. B. (1961). The folded normal distribution. *Technometrics, 3*(4), 543–550. doi:10.1080/00401706.1961.10489974

McCulloch, C. E. (1987). Tests for equality of variances with paired data. *Communications in Statistics—Theory and Methods, 16*(5), 1377–1391. doi:10.1080/03610928708829445

Morgan, W. A. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika, 31*(1/2), 13–19. doi:10.1093/biomet/31.1-2.13

Numerical Algorithms Group. (1999). *NAG Fortran library manual, Mark 19*. Oxford, UK: Author.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika, 31*(1/2), 9–12. doi:10.1093/biomet/31.1-2.9

Raykov, T., Patelis, T., & Marcoulides, G. A. (2011). Examining parallelism of sets of psychometric measures using latent variable modeling. *Educational and Psychological Measurement, 71*(6), 1047–1064. doi:10.1177/0013164410391250

Robey, R. R., & Barcikowski, R. S. (1992). Type-I error and the number of iterations in Monte-Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 43*(1), 113–130. doi:10.1111/j.2044-8317.1992.tb00993.x

Schmidtke, K., & Metternich, B. (2009). Validation of two inventories for the diagnosis and monitoring of functional memory disorder. *Journal of Psychosomatic Research, 67*(3), 245–251. doi:10.1016/j.jpsychores.2009.04.005

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods, 5*(2), 230–240. doi:10.1037/1082-989X.5.2.230

Serlin, R. C., & Harwell, M. R. (2004). More powerful tests of predictor subsets in regression analysis under nonnormality. *Psychological Methods, 9*(4), 492–509. doi:10.1037/1082-989X.9.4.492

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. doi:10.1007/s11336-008-9101-0

Thompson, B. L., Green, S. B., & Yang, Y. (2010). Assessment of the maximal split-half coefficient to estimate reliability. *Educational and Psychological Measurement, 70*(2), 232–251. doi:10.1177/0013164409355688

Torrance, N., Smith, B. H., Lee, A. J., Aucott, L., Cardy, A., & Bennett, M. I. (2009). Analysing the SF-36 in population-based research. A comparison of methods of statistical approaches using chronic pain as an example. *Journal of Evaluation in Clinical Practice, 15*(2), 328–334. doi:10.1111/j.1365-2753.2008.01006.x

Werheid, K., Hoppe, C., Thöne, A., Müller, U., Müngersdorf, M., & von Cramon, D. Y. (2002). The Adaptive Digit Ordering Test: Clinical application, reliability, and validity of a verbal working memory test. *Archives of Clinical Neuropsychology, 17*(6), 547–565. doi:10.1093/arclin/17.6.547

Wilcox, R. R. (1990). Comparing the variances of two dependent groups. *Journal of Educational Statistics, 15*(3), 237–247. doi:10.3102/10769986015003237

Wilks, S. S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *Annals of Mathematical Statistics, 17*(3), 257–281. doi:10.1214/aoms/1177730940

Woods, S. P., Moran, L. M., Dawson, M. S., Carey, C. L., Grant, I., & HNRC Group. (2008). Psychometric characteristics of the Memory for Intentions Screening Test. *The Clinical Neuropsychologist, 22*(5), 864–878. doi:10.1080/13854040701595999