Psychological Methods

# Outlier Removal, Sum Scores, and the Inflation of the Type I Error Rate in Independent Samples t Tests: The Power of Alternatives and Recommendations

Marjan Bakker and Jelte M. Wicherts

Online First Publication, April 28, 2014. http://dx.doi.org/10.1037/met0000014

Psychological Methods

© 2014 American Psychological Association
1082-989X/14/$12.00 DOI: 10.1037/met0000014

# Outlier Removal, Sum Scores, and the Inflation of the Type I Error Rate in Independent Samples *t* Tests: The Power of Alternatives and Recommendations

Marjan Bakker
University of Amsterdam

Jelte M. Wicherts
Tilburg University

In psychology, outliers are often excluded before running an independent samples *t* test, and data are often nonnormal because of the use of sum scores based on tests and questionnaires. This article concerns the handling of outliers in the context of independent samples *t* tests applied to nonnormal sum scores. After reviewing common practice, we present results of simulations of artificial and actual psychological data, which show that the removal of outliers based on commonly used *Z* value thresholds severely increases the Type I error rate. We found Type I error rates of above 20% after removing outliers with a threshold value of $Z = 2$ in a short and difficult test. Inflations of Type I error rates are particularly severe when researchers are given the freedom to alter threshold values of *Z* after having seen the effects thereof on outcomes. We recommend the use of nonparametric Mann-Whitney-Wilcoxon tests or robust Yuen-Welch tests without removing outliers. These alternatives to independent samples *t* tests are found to have nominal Type I error rates with a minimal loss of power when no outliers are present in the data and to have nominal Type I error rates and good power when outliers are present.

*Keywords:* outliers, Type I error, power, robust statistics, nonparametric tests

*Supplemental materials:* http://dx.doi.org/10.1037/met0000014.supp

The practical use of statistical tests in psychological research often deviates from how the use of these tests is described in the textbooks (Bakker, Van Dijk, & Wicherts 2012; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, & Van der Maas, 2011; Wicherts, Bakker, & Molenaar, 2011). For instance, Agresti and Franklin (2007) indicated that outliers should be identified to investigate potential recording errors and recommended that a statistical analysis be repeated with and without an outlier to “make sure the results are not overly sensitive to a single observation” (p. 69). Stevens (2001) stipulated that outliers should not be automatically dropped from the analysis (unless the outlier is caused by an error) and also recommended the reporting of analyses with and without the outlier(s). On the other hand, several of the dozen textbooks[1] that we reviewed concerning outliers (e.g., Howell, 2002; Moore, McCabe, & Craig, 2009; Wilson & MacLean, 2011) did not give any advice about what to do with (nonerroneous) outliers in the data. Other textbooks recommended adjustments of the extreme scores (Dancey & Reidy, 2007), large sample sizes and replications (Nolan & Heinzen, 2007), and transformations, nonparametric and bootstrap procedures, preferably conducted or decided on before analyzing the data (Aron, Aron, & Coups, 2009; Field, 2013; Langdridge & Hagger-Johnson, 2009). The textbooks we scrutinized typically did not recommend a thoughtless removal of outliers, as outliers can be extreme yet actual values of the population under investigation (Freedman, Pisani, & Purves, 2007). However, Howitt and Cramer (2011) stated, “Outliers may be the result of a wide range of different factors. One does not have to identify what is causing such big or small values, but it is important to eliminate them because they can be so misleading” (p. 28), although they did mention the option to report the outcome of the analysis both with complete data and with outliers excluded. Unfortunately, exclusion of data is not always reported. Of the 161 psychological researchers submitting information about their studies to PsychDisclosure.org, 11.2% disclosed that they had not fully reported all excluded observations (LeBel et al., 2013). Furthermore, John et al. (2012) recently surveyed more than 2,000 research psychologists about their involvement in questionable research practices (QRPs) and found that 38% admitted to having decided to exclude data after looking at the impact of doing so on the results. Simmons et al. (2011) called these practices *researchers’ degrees of freedom* and argued that their use can result in strongly inflated Type I error rates. Although Simmons et al. stated

Marjan Bakker, Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands; Jelte M. Wicherts, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, the Netherlands.

The preparation of this article was supported by Netherlands Organization for Scientific Research (NWO) Grants 451-07-016 and 400-08-214.

Correspondence concerning this article should be addressed to Marjan Bakker, Department of Psychology, Psychological Methods, University of Amsterdam, Weesperplein 4, 1018 XA Amsterdam, the Netherlands. E-mail: M.Bakker1@uva.nl

[1] Although this is a convenience sample of textbooks, we think it gives a good illustration of how the handling of outliers is (not) discussed in recent statistical textbooks.

that they found inconsistency in and ambiguity about the decision of removing outliers in reaction time data in 30 *Psychological Science* articles, they did not include the subjective removal of data in their simulation study of inflated Type I error rates. Bakker et al. (2012) studied the effect of the subjective removal of outliers and other QRPs and found Type I error rates up to .40, substantial bias in effect-size estimates, and distortions of meta-analytic results.

This article is concerned with the common practice of deleting outliers in the context of the independent samples *t* test (from now on referred to as *t* test) when data are nonnormal due to the use of sum scores based on tests and questionnaires. We first discuss outliers and nonnormality in psychological data. Then, we review common practice and study by means of simulations the potential inflations of the Type I error rate when data are nonnormal. We also study the merits of a nonparametric and a robust statistical alternative to outlier removal and give concrete recommendations to improve the current practice.

## Outliers

Barnett and Lewis (1994) defined an outlier in a set of data as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" (p. 4) and described two types of outliers. The first one is the *contaminant*, which refers to a value that comes from a different distribution and is not necessarily an extreme value. Examples of contaminants are the score of an ill person in a study of a healthy population or a temperature datum recorded in degrees Fahrenheit instead of degrees Celsius. Insofar that contaminants can be detected on the basis of supplementary information, it is sensible both methodologically and substantively to correct them or delete them from further analyses. The second type of outlier is the *extreme observation*, which refers to a value that is either extremely low or extremely high but is still from the same distribution as the other values. As contaminants can be extreme as well, it is often hard to distinguish contaminants from extreme values. Furthermore, while some extreme values are expected in normally distributed data, they are part and parcel of heavy-tailed distributions. Heavy-tailed distributions may look similar to normal distributions and so may be hard to distinguish from them (Wilcox, 1998), especially in small samples that are quite typical of psychological experiments (Bakker et al., 2012). However, the variance of the heavy-tailed distribution will be much larger, as well as its standard error of the mean. As a consequence, the power of parametric tests like *t* tests or analyses of variance (ANOVAs) will decrease dramatically when applied to data from heavy-tailed distributions (Wilcox, 1997).

Despite the doubts that have been raised on the normal distribution as the underlying distribution of much of the data from psychological research (Micceri, 1989; Taleb, 2007), normality-based tests like ANOVAs and *t* tests continue to be the predominant methods to test for differences in means between groups in psychology. For instance, in a sample of 252 articles from *Psychonomic Bulletin & Review* and *Journal of Experimental Psychology: Learning, Memory, and Cognition,* Wetzels et al. (2011) found an average of 3.39 *t* tests per article. Likewise, in our recent study of the prevalence of reporting errors in a fairly representative sample of 281 psychological articles that involved null hypothesis significance testing (Bakker & Wicherts, 2011), we found an average of 4.29 *t* tests per article.

### Dealing With Outliers

Because of their potential effects on parametric statistical techniques, different methods have been developed to detect outliers. A widely used method of outlier detection is based on computing the absolute *Z* value (estimated with $Z = |X_i - \bar{X}|/SD$, where $X_i$ is the observed value, $\bar{X}$ is the sample mean, and *SD* is the sample standard deviation). A threshold *k* is chosen that is associated with low probability in the standard normal distribution. Common values of *k* are 2 and 3. This means that from a fully random sample from a normal distribution, the expected percentage of identified outliers is 4.55% when $k = 2$ and 0.27% when $k = 3$. Other outlier-detection methods are based on the median absolute deviation statistic (MAD), the interquartile range (IQR), and the discordancy or slippage test (see Barnett & Lewis, 1994, for an extensive description of different outlier-detection methods). We restrict our attention in this article to univariate outlier detection. We refer to Wilcox's (2012) description of multivariate methods like minimum volume ellipsoid, minimum covariance determinant, minimum generalized variance, and Mahalanobis distances.

After detecting an outlier, many researchers will be inclined to delete the outlier from the data and continue with their standard parametric analyses. However, removing outliers from the data set and continuing with parametric statistical analyses may be a suboptimal solution. After removing outliers, the observations become dependent because the remaining data are dependent on the order of the data, and therefore, the estimation of the standard error becomes unsound (see Wilcox, 1998, 2012, for an illustration). As a consequence, when applying a *t* test or ANOVA, the variance will be underestimated after deleting outliers, which will inflate Type I error rates of these parametric tests (Grissom, 2000; Huber, 1981; Wilcox, 1998). As outliers (contaminants) in the data cause a deviation from the assumed distribution, nonparametric and robust statistical methods are good alternatives. Nonparametric methods do not assume that the data are from a specific distribution (usually the normal distribution). For example, the Mann-Whitney-Wilcoxon test (MWW; also called Wilcoxon rank-sum test or Mann-Whitney U test; Mann & Whitney, 1947; Wilcoxon, 1945) uses the order of data values instead of their actual values to compare two independent samples. Nevertheless, the MWW still assumes that the distributions of both groups are equal, and therefore, this test is sensitive to heteroscedasticity (unequal variances). If the distributions are different, a wrong standard error is being used, which can lead to a conservative or anticonservative Type I error rate, especially if the group sizes differ as well (Neuhäuser, Lösch, & Jöckel, 2007; Zimmerman, 1998). Two nonparametric tests that control the Type I error rate reasonably well (Neuhäuser et al., 2007) and can take ties into account are the Brunner-Munzel procedure (Brunner & Munzel, 2000) and Cliff's method (Cliff, 1996). On the other hand, robust statistical methods, which are approximate parametric models (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Huber, 1981; Staudte & Sheater, 1990; Wilcox, 1997), can handle both nonnormal distributions and heteroscedasticity and are therefore more useful than nonparametric tests if the data are heteroscedastic and drawn from a nonnormal distribution (Erceg-Hurn & Mirosevich, 2008; Wilcox, 1998; Zimmerman, 1994, 1998). A robust statistical method to compare means in two independent samples is the Yuen-Welch test (Y-W; Welch, 1938; Yuen, 1974). Bootstrap methods do not make assumptions about the underlying distributions and represent another alternative when the presence of outliers results in a nonnormal distribution (Wilcox, 2012). The Welch-Satterthwaite test (Satterthwaite, 1946; Welch, 1938, 1947) is an alternative of the *t* test when the assumption of heteroscedasticity is violated and is presented together with the *t* test in the software package SPSS. However, this Welch-Satterthwaite test is not robust to possible outliers in the data (Zimmerman & Zumbo, 1992).

## When Normality Does Not Apply

In this article, we review the widespread practice of outlier exclusion and study its ramifications for the Type I (and Type II) error rate when data are nonnormal because of the common use of sum scores from psychological tests and questionnaires. Previous simulation studies (MacDonald, 1999; Sawilowsky & Blair, 1992; Zimmerman, 1994, 1998) focused on the robustness of *t* tests against violations of normality, heteroscedasticity, unequal sample sizes, and the presence of outliers. However, we investigate the influence of the debatable removal of extreme scores from the sample on the Type I error rate. More importantly, previous simulation studies did not involve data distributions that arose from the use of sum scores, which is arguably a very common reason for a failure of normality in psychological research. Apart from the possibility that (underlying) latent traits are nonnormal, there are psychometric reasons for expecting nonnormality even if these underlying traits are normal. These reasons include (a) the fact that sum scores from psychological tests and questionnaires are always bounded by the number of scoring options, (b) psychological tests are typically not overly long in experimental work for practical reasons (e.g., because experimental effects are often short lived), and (c) the possibility that the item set is tailored to be most informative for a particular level of the latent trait that does not suit the study sample well. For instance, many measures of negative moods are more informative for clinical samples (e.g., those with depression), leading to negatively skewed distributed sum scores in nonclinical samples (or in samples that have recovered from depression). Similarly, a cognitive test used in an experiment may be deliberately made more difficult to heighten its cognitive demands, which leads to sum scores that are nonnormally distributed. Such sources of nonnormality in experimental work increase the likelihood of asymmetric and heavy-tailed distributions in which large values (although bounded) of *Z* are quite common. Nonetheless, researchers often do not check the assumptions underlying the *t* test (Hoekstra, Kiers, & Johnson, 2012) and often invoke large *Z* values to exclude particular data points before running a *t* test or another statistical analysis.

Below, we present the results of simulations of artificial data and actual psychological data showing that this practice leads to an unacceptable inflation of the Type I error rate. We also show that the problem is aggravated when researchers are given the freedom to choose (after having seen the data) the level of significance (i.e., threshold values of *Z*) at which they consider data points to be outlying. The identification of outliers has been shown to be quite subjective (Collett & Lewis, 1976), but we are not aware of any research on the implications of such subjective rules of outlier detection on the Type I error rate of the *t* test. Furthermore, we also study the nonparametric MWW test (as we do not focus on heteroscedasticity) and the statistically robust Y-W test as alternatives with a better control of the Type I error rate against a minimal loss in power if the data do not contain outliers and an increase in power if the data contain outliers.

# Current Practice

To get an indication of the current practice of outlier handling, we selected six journals for review: *Journal of Experimental Social Psychology*, *Cognitive Development*, *Cognitive Psychology*, *Journal of Applied Developmental Psychology*, *Journal of Experimental Cognitive Psychology*, and *Journal of Personality and Social Psychology*. We selected these journals because they contain mainly experimental research and represent different research fields within psychology. Furthermore, the first five journals are available through ScienceDirect, which enabled an in-text search for relevant studies. *Journal of Personality and Social Psychology* was searched by using Google Scholar.

A total of 5,129 articles were published between 2001 and 2010. The number of articles for each journal separately is specified in Table 1. Subsequently, we selected the 353 (7%) articles that contained the word *outlier* in the text. Note that the actual number of studies that removed data could be larger due to the use of other terms than *outlier*. From each journal, we randomly selected 25 articles that contained the word *outlier* for close examination.[2]

## Results of the Review

The most commonly used method to detect outliers concerned the use of the *Z* score, which was used in 63 articles (46%), where *k* ranged from 1.76 to 10 (median = 3). Various authors used a value of 3.29 with a reference to Tabachnick and Fidell (2001), who recommended this value because it tests whether a value is more extreme than the mean of the sample with $p < .001$. A problem with this recommendation is that in large samples, some extreme cases are to be expected. Moreover, as every value is tested against the mean, as many tests are performed as the number of data points in the sample, therefore involving as many hypothesis tests as there are participants, which creates a multiple testing problem (Benjamini & Hochberg, 1995). The *Z* score method suffers also from masking, which means that the presence of outliers inflates the sample mean and sample variance and therefore can mask the presence of (other) outliers (Wilcox, 2012). Furthermore, in small sample sizes, the *Z* value will never exceed $(n - 1)/\sqrt{n}$ (Shiffler, 1988), which makes the *Z* score especially unsuitable to identify outliers in small sample sizes. For example, a *Z* value of 3.29 cannot be observed in a sample as small as 12 participants.

In only five articles (3.6%) were outliers identified by means of boxplots or the IQR, which is a better method to identify outliers than using *Z* scores as the IQR suffers less from masking (Wilcox, 2012). Therefore, the IQR is generally recommended in textbooks that introduce statistics for psychology students (e.g., Agresti &

[2] *Cognitive Development* contained only 12 articles that used the term *outlier*. All these 12 articles were examined.

Table 1
*Descriptive Statistics of Outlier Handling Methods*

| Journal | Number of articles | Outlier mentioned | Removed | Number used *k* | Average value of *k* (range) | Double analyses |
|---|---|---|---|---|---|---|
| *Journal of Experimental Social Psychology* | 1,063 | 127 (12%) | 21 (84%) | 12 (48%) | 3.00 (2.00–5.05) | 5 (20%) |
| *Cognitive Development* | 400 | 12 (3%) | 9 (75%) | 5 (42%) | 2.61 (1.76–3.29) | 4 (33%) |
| *Cognitive Psychology* | 349 | 32 (9%) | 17 (68%) | 8 (32%) | 2.50 (2.00–3.00) | 2 (8%) |
| *Journal of Applied Developmental Psychology* | 542 | 33 (6%) | 17 (68%) | 9 (36%) | 2.93 (2.00–3.29) | 5 (20%) |
| *Journal of Experimental Cognitive Psychology* | 685 | 63 (9%) | 21 (84%) | 15 (60%) | 2.70 (1.96–3.29) | 0 (0%) |
| *Journal of Personality and Social Psychology* | 2,090 | 86 (4%) | 21 (84%) | 14 (56%) | 3.82 (2.36–10.00) | 8 (32%) |
| Total | 5,129 | 353 (7%) | 106 (77%) | 63 (46%) | 3.01 (1.76–10.00) | 24 (18%) |

Franklin, 2007; Howitt & Cramer, 2011; Moore et al., 2009). The IQR is the range that contains 50% of the observations that are all in the middle of the sample (75th percentile–25th percentile). This is also the box part of the boxplot. Values that are located 1.5 (or 2) IQR outside the lower and upper quartiles are defined as outliers by Tukey (1977). Furthermore, the 90th percentile and 95th percentile were used as an outlier-detection criterion. None of the inspected articles used the MAD-median rule, where *X* is declared an outlier if

$$\frac{\left| X - M \right|}{MADN} > 2.24, \tag{1}$$

where *M* is the median, *MADN* is *MAD*/0.6745, and *MAD* is the median of the absolute difference between every value and the median of these values. This method suffers even less from masking than the IQR (Wilcox, 2012).

In 106 articles (77%), the outliers were removed before starting the actual analyses. Besides the removal of outliers, we came across nine articles (6.6%) where the most extreme values were replaced with less extreme values. Although this so-called Winsorization procedure is a robust method to estimate the mean, applying statistical analysis like a *t* test on this adjusted data set will not result in robust results because the estimation of the standard error is incorrect (Wilcox, 2012). Hence, this practice is suboptimal.

We came across additional outlier-detection and handling methods that were more specific for the analyses of interest. We found the use of specific cutoff criteria in 14 articles (10.2%), especially with reaction time data (e.g., remove responses longer than 6,000 ms). A problem with outliers in reaction time data is that the data are typically positively skewed with a long tail with slow responses. Miller (1991) already showed that the mean will be biased when the *Z* value outlier criterion is applied and sample sizes differ. Van Selst and Jolicoeur (1994) described other outlier removal procedures for reaction time data that are recursive and insensitive to amount of skew and sample size. However, the variances of the sample will still be reduced, and therefore, the problem with the inflation of the Type I error rate continues to apply. Problems with outliers in reaction time data are beyond the scope of this article.

Additional practices of outlier detection in our set of articles involved Mahalanobis distance, Cook's distance, multivariate percentiles, and inspection of the scatterplot. In a classification study, the RESIDAN procedure was used (Bergman, 1988). Finally, in a few articles, outliers were removed until the data satisfied normality assumptions based on, for instance, Shapiro Wilks's test or certain levels of skewness and kurtosis.

In 24 articles (18%), analyses were reported both with and without outliers. This procedure clearly lends support to the robustness of results and is for example recommended by Stevens (2001) and Howitt and Cramer (2011). However, we also came across some indications of subjective identification of outliers. For example, in one article, the authors stated that a nonsignificant effect was found, but after inspection of the data, the value of *k* was changed from 3 to 2, which gave better results. Furthermore, it was generally unclear whether authors selected their threshold value in advance.

Thus, our review of common practice of dealing with outliers showed that (a) the removal of outliers before starting the actual analyses is common practice and (b) a *Z* value criterion (typically with values of 2 or 3) is the most commonly used method to detect outliers. Next, we study the implications of these practices.

## Simulation Study 1: Removing Outliers and Type I Error Rate

In this simulation study, we investigate the Type I error rate of the *t* test when outliers are removed from the data. We start with randomly generated values from a normal distribution. However, as stated before, psychological research data are often not normally distributed (Micceri, 1989). In psychology, variables are often discrete and bounded because they are based on answers to questionnaires or tests. Therefore, we generated sum scores based on a Rasch model, which is comparable to tests with true/false items, and based on a polytomous item response model, which is comparable to tests with polytomous items (e.g., a 5-point Likert-type scale). We made tests that fitted the latent (or underlying) trait of the simulated test-takers and tested different test lengths. A test is often too difficult or too easy for a test-taker; the difficulty of this test may not match the latent trait of the subjects. For example, most persons score quite low on the Symptom Check List 90–Revised (Derogatis, 1994), which is a questionnaire that measures psychological problems and symptoms of psychopathology. Healthy test-takers have a low probability of responding positively to an item with psychopathology symptoms, while persons with severe problems (high latent trait values) will answer more questions with yes (their latent trait matches the difficulty of the test). Therefore, we also simulated data in which tests did not fit the latent trait distribution but were relatively difficult vis-à-vis the latent trait distribution in the sample. Given symmetry, our results apply equally well to instances where tests are too easy for the sample of test-takers.

Besides simulated data, we used two large actual data sets to study Type I error rates. The first data set involved responses to the Raven's Progressive Matrices (Raven, Raven, & Court, 2003; dichotomous data), and the second data set involved responses to the Dutch (shortened) version of the Profile of Mood States (POMS; Lorr, McNair, & Droppleman, 1992; polytomous data). In these real-data simulations, a random variable can be used to make artificial groups in which the null hypothesis is expected to be true.

## Method

To investigate the influence of removing outliers on the Type I error rate, we collected 100,000 *p*-values of a *t* test comparing two samples of the same distributions. In the first simulation, we used scores that are normally distributed ($N{\sim}0{,}1$).

To simulate dichotomous data, we used a Rasch model, which is an item response theory (IRT) model (Embretson & Reise, 2000). The probability of person *j* answering an item *i* correctly, $\Pr(X_{ij} = 1)$, can be calculated based on the difficulty ($\beta$) of the item *i* and the ability ($\theta$) of person *j* with the following equation:

$$\Pr(X_{ij} = 1) = \frac{\exp(\theta_j - \beta_i)}{1 + \exp(\theta_j - \beta_i)}. \tag{2}$$

Instead of sampling scores directly from a distribution (as done in the first part), we sampled theta values from a normal distribution ($N{\sim}0{,}1$). Furthermore, we sampled beta values from the same distribution. This will lead to a test that fits the ability of the participants. We used four different test lengths (5, 10, 20, and 40 items) that are typical for psychological tests (Emons, Sijtsma, & Meijer, 2007). We calculated for each item–person pair the probability of answering the item correctly with Equation 2. If this probability of answering the item correctly was larger than a value sampled from a uniform distribution (value between 0 and 1), the person has answered the item correctly, and a value of 1 is assigned to this person–item pair. Otherwise, a 0 is assigned to the pair. Thereafter, a sum score is calculated for each person by adding all the item scores. Thus, a person with a high theta value has a higher chance of answering items correctly and will therefore have a higher sum score. To make a difficult test, we used a normal distribution of beta with a mean of 3 ($N{\sim}3{,}1$). The distribution of theta remains the same across the cells of the simulation study.

In the third part, we simulated the sum scores based on a polytomous IRT model. Samejima (1997) developed the graded response model (GRM), with which the probability of scoring in a specific category is modeled by the probability of responding in (or above) this category minus the probability of responding in (or above) the next category. Let $C_k$ denote the number of response categories of item *k*; then, there are $C_k - 1$ threshold values between the response options. We chose to use equal discrimination parameters in our simulation study ($a = 1$). Therefore, the threshold values can be calculated with Equation 2 (see Samejima, 1997, for the complete equations). To simulate sum scores based on polytomous items with five answer categories, we generated four beta values (number of answer options minus one), which correspond to the threshold values between different answer categories. These beta values were drawn from a normal distribution ($N{\sim}0{,}1$) and subsequently ordered. For every item–person pair, the four probability threshold values of the item were calculated by using Equation 2 for the four betas. Thereafter, a random value was generated from a uniform distribution (value between 0 and 1). This value was then compared with the threshold values to determine the answer's category. Next, a sum score was calculated for each person by adding all the item scores. We used five different test lengths (2, 5, 10, 20, and 40 items). To make a difficult test, we generated betas from a normal distribution with mean 3 ($N{\sim}3{,}1$).

Furthermore, we used a real data set that consisted of the answers of 2,301 first-year psychology students to the Raven's Progressive Matrices (Raven et al., 2003), administered (with a time limit of 20 minutes) between 2001 and 2009 at the University of Amsterdam (Amsterdam, the Netherlands). This test measures fluid reasoning and consist of 36 multiple-choice items. In each item, the subject is asked to identify the missing element that completes a pattern and can therefore be answered correctly or wrongly. We used the total score (number of items answered correctly) and the sum score of the first 10 items. As the items increase in difficulty, the first 10 items will make an easy test.

In the last part, we used a real data set that consisted of the answers of 5,912 first-year psychology students to the Dutch (shortened) POMS (Lorr et al., 1992), administered between 1989 and 2001 at the University of Amsterdam (Wicherts & Vorst, 2004). This questionnaire consists of five scales (Tension-Anxiety, Depression-Dejection, Anger-Hostility, Vigor-Activity, and Fatigue-Inertia). These scales consist of 6, 8, 7, 5, and 6 items, respectively, and are answered on a 5-point Likert-type scale.

We compared different sample sizes (20, 40, 100, and 500 per cell), which cover common sample sizes in psychology research (Bakker et al., 2012). For each sample size, we randomly drew 100,000 samples from the real data set. From each sample, we removed outliers with an absolute *Z* value larger than *k*. We used different values of *k* (2 to 4 in steps of 0.1), performed for every value of *k* an independent samples *t* test, and collected the *p*-value. After collecting the *p*-values, we calculated the (two-sided) Type I error rate by counting the number of *p*-values below .05 and dividing it by the total number of collected *p*-values. We calculated the Type I error rate of a two-sided *t* test because we did not have specific expectations about directionality. However, approximately the same results will be expected when calculating the Type I error of a one-sided *t* test. In discussing the results, we focus on some common values of *k* (2, 2.5, and 3). Furthermore, we investigated the subjective use of *k*. This means that a comparison is counted as statistically significant if the test showed a statistically significant difference when all values were included in the sample or when the test showed a statistically significant difference when *k* is 3, 2.5, or 2. This is comparable with adapting *k* until a statistically significant *p*-value is found. This reflects a manner in which researchers can chase for significance (Ioannidis, 2005, 2012), which appears to be a common practice in psychology (John et al., 2012).

## Results

**Normally distributed scores.** Results of the normally distributed scores are presented in Figure 1 (the tables of this result and all the following results are given in the online supplemental materials). Under normality, Type I error rates of the *t* test became higher with decreasing *k*. Larger sample sizes resulted in a higher Type I error rate than smaller sample sizes. When *k* equaled 3, the Type I error rate was only somewhat higher (between .051 and .056); when *k* was 2.5, it was already between .062 and .071; and when *k* was 2, the Type I error rate doubled to values between .098 and .108. The effects of subjective use of *k* are depicted in Figure 1 on the left side of the vertical dotted line. Results indicate that subjective use of *k* can result in an even higher Type I error rate (between .108 and .125) in normally distributed data.

**Rasch model–based sum scores.** Next, we consider data based on sum scores simulated with the Rasch model. We start with a test that fitted the latent trait distribution (theta and beta values coming from the same distribution). Again, Type I error rates of the *t* test became larger as values of *k* decreased (see the left panel of Figure 2). When *k* was 3, the Type I error rate remained close to .05. When *k* equaled 2.5, the Type I error rate increased somewhat to values between .057 and .064, with larger values for shorter tests. However, for $k = 2$, we see a sharp rise in the Type I error rate to values between .101 and .167, with larger rates for shorter test and larger samples. Subjective use of *k* can result in even higher Type I error rates (ranging from .104 to .189).

With sum scores based on the Rasch model with difficult items (average $\beta = 3$), the Type I error rate became quite large even with larger values of *k* (cf. the right panel of Figure 2 and in Table 2 in the online supplemental materials). With $k = 3$, Type I error rates already increased to values between .075 and .137, and with $k = 2.5$, they increased to values between .090 and .155. When *k* was 2, the Type I error rates lay between .103 and .173. Larger Type I error rates are found for shorter test and larger sample sizes. We see a somewhat irregular line for a difficult test with five items and 20 subjects. This is a simulation artifact, as the scores in this situation can only range from 0 to 5 (so an increased number of simulations will not smooth this line). Again, subjective use of *k* resulted in even higher Type I error rates that ranged from .125 to .345.

**GRM-based sum scores.** Now, we consider sum scores based on polytomous data simulated under the GRM. With a test that fitted the latent trait (theta and beta values coming from the same distribution), Type I error rates of the *t* test became larger with smaller values of *k* (see left panel of Figure 3). For the different number of items, we see comparable patterns: The Type I error rate was still very close to .05 when *k* equaled 3 (between .049 and .052) or 2.5 (between .050 and 0.059). Thereafter, we see a sharp increase in Type I error rates, which led to Type I error rates between .095 and .144 when *k* was 2. Subjective use of *k* resulted in Type I error rates that ranged from .096 to .148.

With sum scores based on the GRM with difficult items (average $\beta = 3$), the Type I error rates became larger with smaller

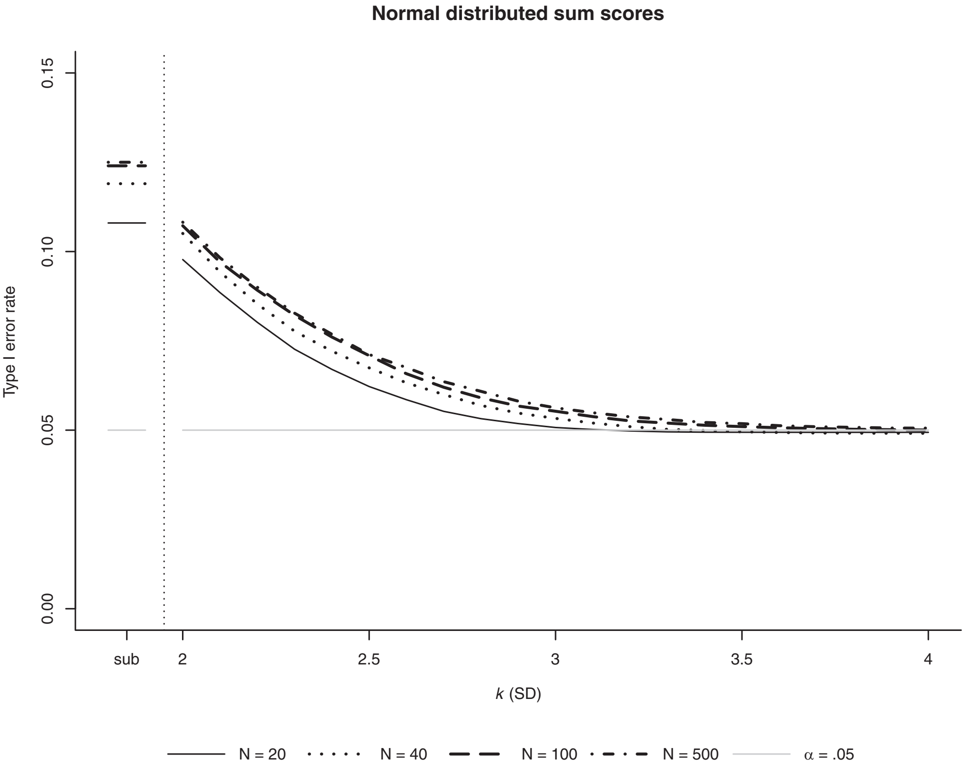

*Figure 1.* Type I error rate of a *t* test of sum scores directly generated from a normal distribution for different values of *k* and different sample sizes, and of subjectively used values of *k* (sub, left side of vertical dotted line). The horizontal gray line denotes the nominal Type I error rate ($\alpha = .05$).

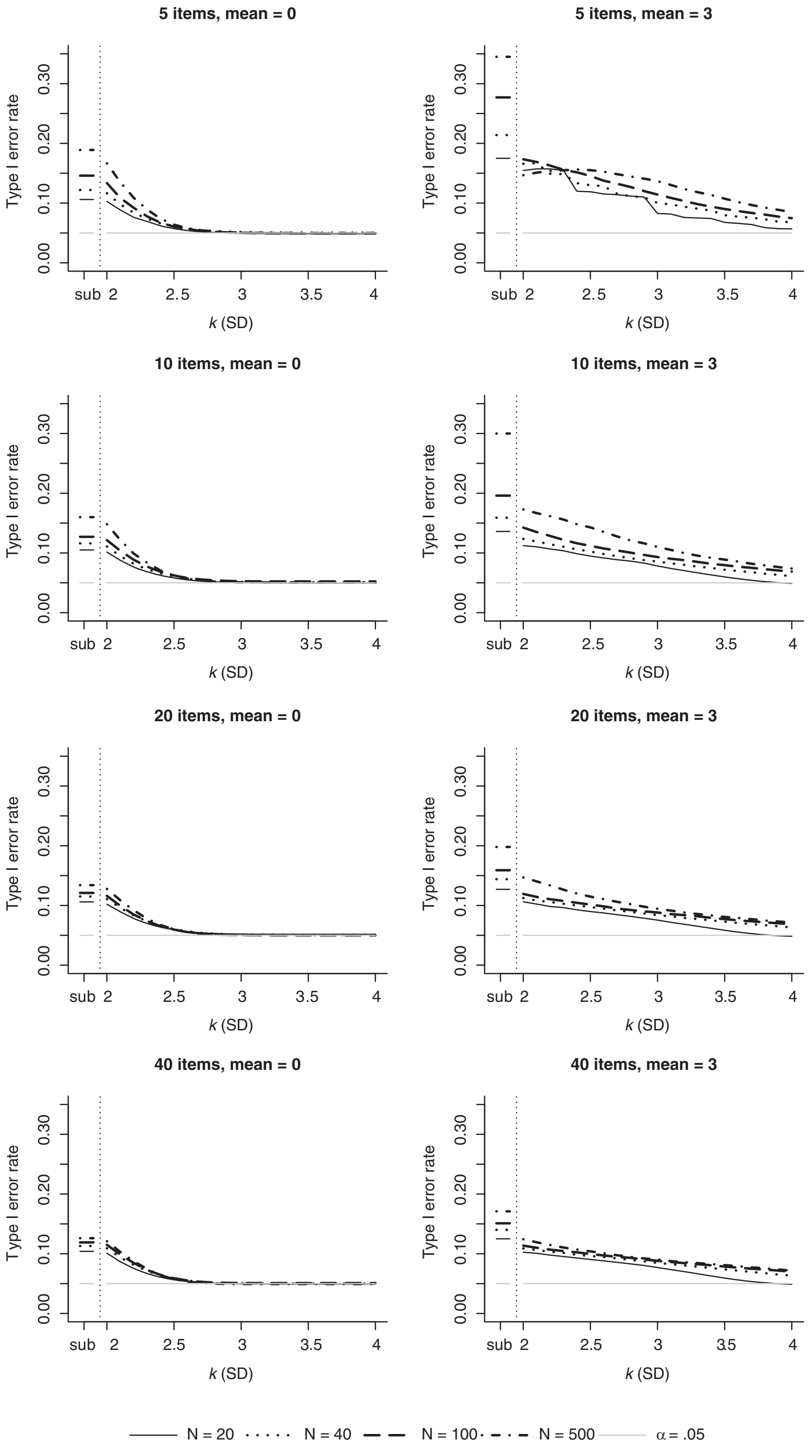

*Figure 2.* Type I error rate of a *t* test of sum scores based on a Rasch model for different values of *k* and different sample sizes and different test lengths for well-fitting items (left column) and difficult items (right column). Type I error rate of subjective use of *k* is presented on the left side of the vertical dotted line in each plot, and the horizontal gray line denotes the nominal Type I error rate (α = .05).

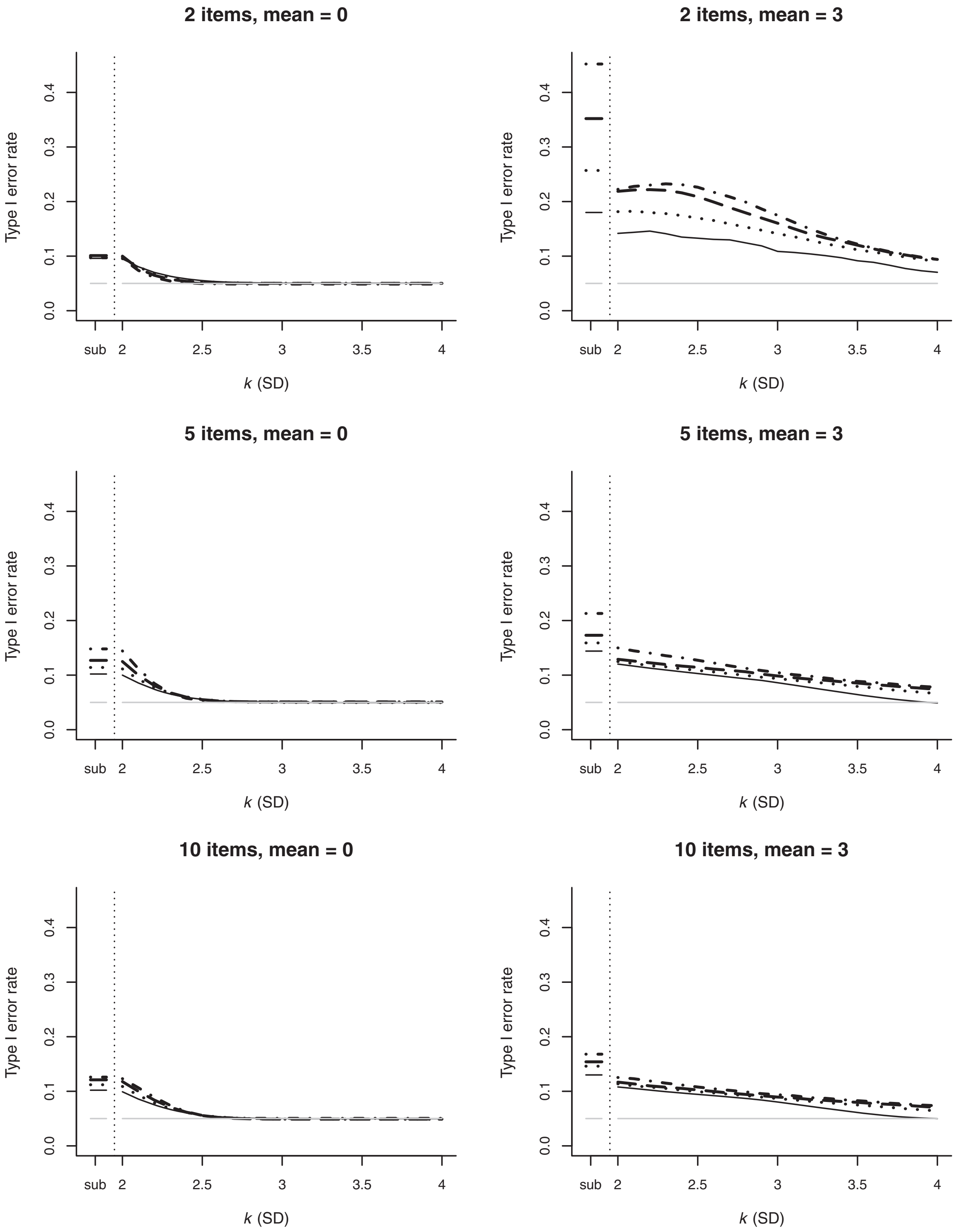

*Figure 3.* Type I error rate of a *t* test of sum scores based on a graded response model for different values of *k* and different sample sizes and different test lengths for well-fitting items (left column) and difficult items (right column). Type I error rate of subjective use of *k* is presented on the left side of the vertical dotted line in each plot, and the horizontal gray line denotes the nominal Type I error rate (α = .05).

(*figure continues*)

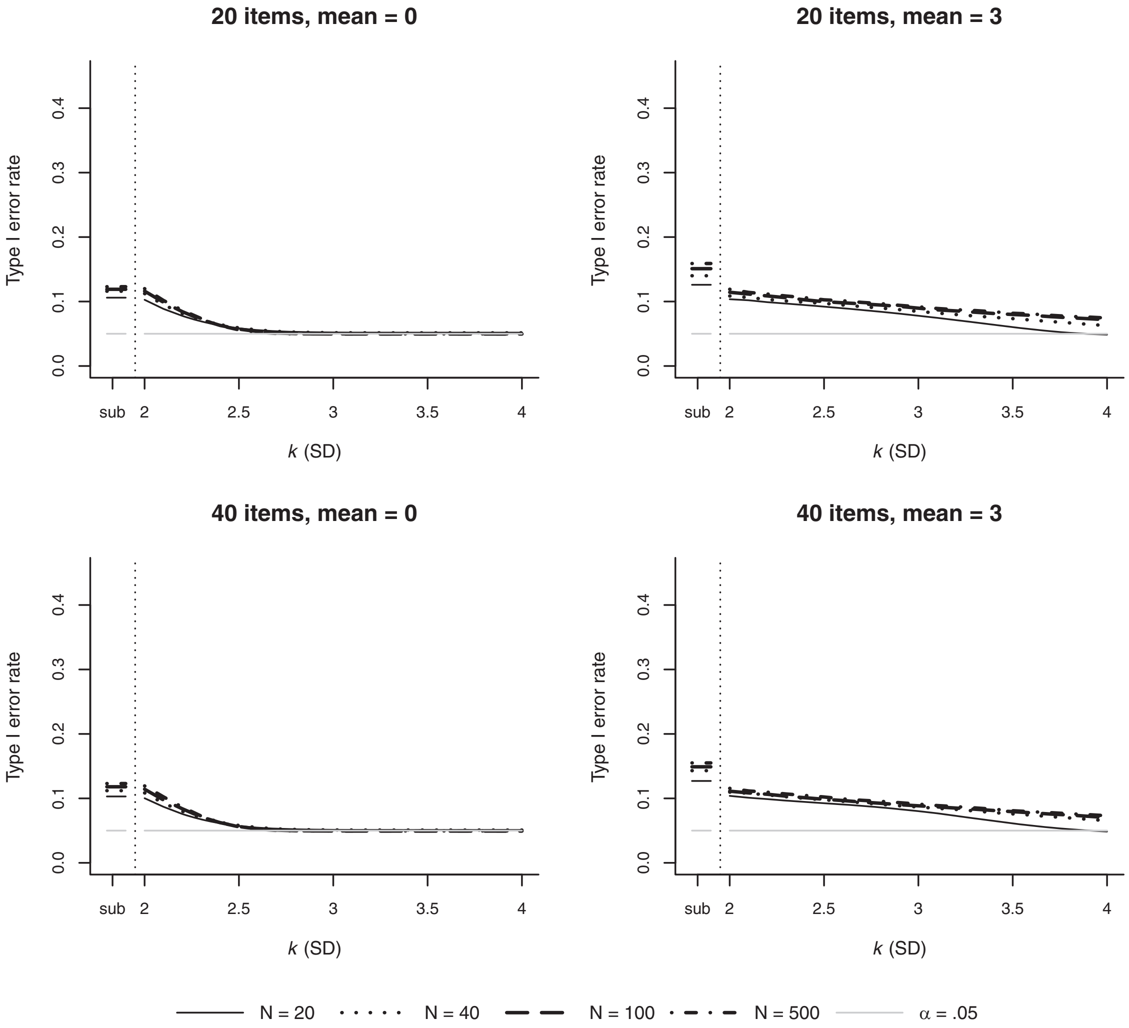

*Figure 3. (continued)*

values of *k*, as can be seen in the right panel of Figure 3. With this difficult test, the Type I error rates were already between .078 and .175 when $k = 3$. When $k = 2.5$, the Type I error rates lay between .092 and .226, and when $k = 2$, they varied from .104 to .222. Subjective use of *k* resulted in even higher Type I error rates that ranged from .126 to .452.

**Real data with dichotomous scores: Raven's Progressive Matrices.** To corroborate the results of the previous simulations with actual data, we used a data set with scores on dichotomous items from Raven's Progressive Matrices. In each iteration, we drew two random samples of the same size from the large data set (without replacement), which enabled us to compute *Z* values and *t* tests in each iteration. Given the randomness of the selection into subgroups, the null hypothesis of the *t* tests can be assumed to be approximately true.

The distributions of the sum scores on the Raven's test are presented in Figure 4, for the entire test (left panel) and the easy test (right panel), which was based on the first 10 items. The total scores showed a moderately skewed distribution (skewness = −0.748), while the sum scores based on the first 10 items showed a skewed distribution (skewness = −2.980).

The results of the real data simulation are presented in Figure 5, where it can be seen that Type I error rates of the *t* test were largest for smaller values of *k*. Some lines in this figure are somewhat irregular because of a simulation artifact due to scores being bounded integers. For the full test scores, removing outliers led to increased Type I error rates between .059 and .065 when *k* was 3, between .067 and .070 when *k* was 2.5, and between .087 and .094 when *k* was 2. Subjective use of *k* resulted in Type I error rates between .107 and .127. Removing outliers from data in which the sum scores were based on the first 10 items led to Type I error rates between .082 and .086 when *k* was 3, between .097 and .127 when *k* was 2.5, and between .104 and .144 when *k* was 2. Again, subjective use of *k* increased the Type I error rates to values between .138 and .198.

**Real data with polytomous scores: Profile of Mood States.** We used the five scales of the POMS as an example of real polytomous data. As can be seen in Figure 6, the distributions of the scale scores are not normal. The scores on the subscale Vigor were closest to normal, with a skewness of −0.119. Depression scores were most skewed (skewness = 1.030). The distributions of subscale scores of Anger, Fatigue, and Tension were moderately

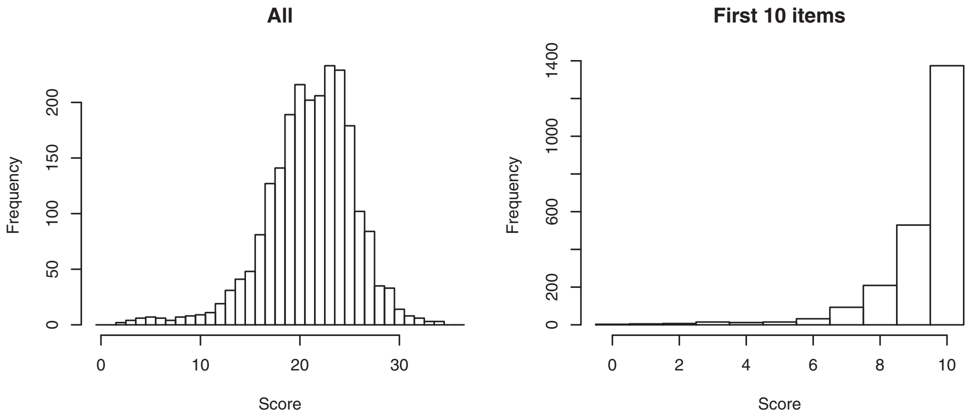

*Figure 4.* Distribution of the sum scores of Raven's Progressive Matrices.

skewed, with skewness values of 0.858, 0.523, and 0.651, respectively. Figure 7 presents the Type I error rates of the different scales. The scale score (Vigor) that aligned most with a normal distribution showed a Type I error rate close to .050 when *k* was 3, but when *k* was 2.5 or 2, the Type I error rate rose to values between .058 and .062 and between .102 and .125, respectively. The other subscale distributions were more skewed, which led the Type I error rates to range from .050 and .087 even when *k* was 3. When *k* was 2.5, the Type I error rate ranged between .069 and .111, and when *k* was 2, the Type I error rate ranged between .100 and .136. Finally, subjective use of *k* resulted in Type I error rates that varied from .107 to .169.

## Simulation Study 2: Type I Error Rate and Power of Y-W and MWW

The simulations above show that removing outliers can severely inflate the Type I error rate of the *t* test, especially when the level of *k* is chosen subjectively. The removal of outliers is therefore not recommended. Nevertheless, if outliers are part of the data, this can have other undesired effects like drops in power (Osborne & Overbay, 2004; Wilcox, 1997). As described above, the Y-W and the MWW tests are less influenced by the presence of outliers and enable a comparison of means in two independent samples without the need to exclude outliers (see the Appendix for a more detailed description of the Y-W and the MWW tests). Here, we investigate whether these tests have a power that can match the power of the *t* test when no outliers are present in the data. The *t* test has more power than both the Y-W and the MWW tests when both samples are from the same normal distribution (Yuen, 1974; Zimmerman & Zumbo, 1992). However, with long-tailed distributions, the power of the Y-W test is superior to the power of the *t* test (Yuen, 1974), and the MWW test has more power than the *t* test when the distribution is skewed, when it has heavy tails, and/or when sample sizes are small (Zimmerman & Zumbo, 1992). Keselman, Othman, Wilcox, and Fradette (2004) found the Y-W test to have good control of Type I error rates even in extreme instances of heterogeneity and nonnormality, especially after transformation for skewness and with bootstrapping.[3] In typical psychological research in which the data are derived from psychological tests, the distribution is different than investigated thus far, as the distribution is bounded and sample sizes are small. Therefore, we study the power of the Y-W and the MWW tests[4] under more common data patterns. Furthermore, we investigate the Type I error rate and the power of the *t* test and compare them to the results of the Y-W and MWW tests when outliers are present in the data. We do not investigate the power of the *t* test after removing outliers based on *k*, as we are interested in a method with both a nominal Type I error rate and good power either with or without outliers. Our first simulation study showed that the Type I error rate of this method increases substantially when no outliers are added to the data.

### Method

We used the same structure as in the first simulation study. We started with normally distributed scores with a sample size of 20, 40, 100, and 500 in each sample. Both samples were drawn from a normal distribution. The first sample was drawn from a standard normal distribution ($N$~0,1), and the second sample had a population mean of 0.0 (no effect), 0.2 (for a small effect), 0.5 (for a medium effect), or 0.8 (for a large effect) and a population standard deviation of 1 (i.e., homoscedascity). We compared the first sample with all the four other samples by means of the *t* test, the Y-W test with 20% trimming, and the MWW test. We used the function `yuen()` from the WRS package for R that can be downloaded from http://r-forge.r-project.org/projects/wrs/ and the function `wilcox.test()` for the MWW test. We did 100,000 comparisons and calculated the proportion of samples that showed an effect at $\alpha = .05$. This gives the Type I error rate if the two samples come from the same distribution (both from $N$~0,1) and the power when samples from different distributions are compared.

[3] Despite small differences in coverage rates between variants of the Y-W test, in Keselman et al.'s (2004) study, the average performance of the standard Y-W test was quite similar to the performance of the bootstrapped and bootstrapped plus transformed variants. Because we do not consider scenarios of nonnormality as extreme scenarios as those by Keselman et al. (in light of our focus on test scores), we restrict the attention to the standard Y-W test here.

[4] We also investigated a permutation test, two variants of the MWW test that take ties into account (Cliff's method and the Brunner-Munzel procedure), and a bootstrap version of the Y-W test (see also Keselman et al., 2004).

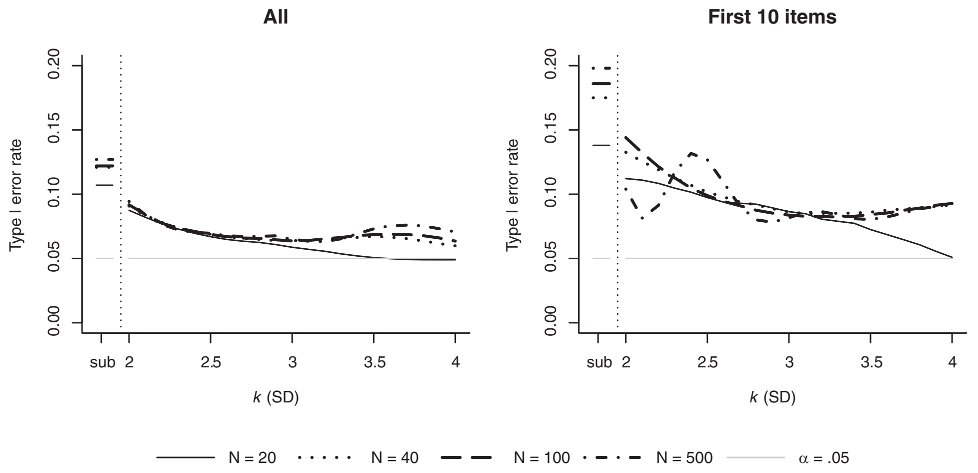

*Figure 5.* Type I error rate of a *t* test of sum scores based on Raven's Progressive Matrices data for different values of *k* and different sample sizes. Type I error rate of subjective use of *k* is presented on the left side of the vertical dotted line in each plot, and the horizontal gray line denotes the nominal Type I error rate (α = .05).

We also generated sum scores based on a Rasch model and based on a GRM. Therefore, we used theta values from *N*~0,1 for the first sample, and from *N*~0,1, *N*~0.2,1, *N*~0.5,1, or *N*~0.8,1 for the second sample. Again, we generated betas from the same distribution of the thetas in the first sample for a more fitting test and generated beta from *N*~3,1 for a difficult test. Sum scores were further generated in the same way as in the first simulation study and based on a model that was measurement invariant across samples (i.e., same item parameters across the two samples). Again, we used test lengths of 5, 10, 20, and 40 items for both the Rasch- and the GRM-based sum scores and used an additional test length of two items for the GRM-based sum scores. The Type I error rate and power of the three different tests were calculated in the same way as for the normally distributed scores.

Furthermore, to investigate the Type I error rate and the power of the *t* test, Y-W test, and MWW test when contaminant outliers are present, we did the same as above but used a mixed-normal distribution, consisting of samples from *N*~0,1 (or *N*~0.2,1, *N*~0.5,1, or *N*~0.8,1, for small, medium, and large effect sizes, respectively) with probability .95, and from *N*~0,400 (or *N*~0.2,400, *N*~0.5,400, or *N*~0.8,400, for small, medium, and large effect sizes, respectively) with probability .05. Researchers have used this distribution widely to simulate heavy-tailed distributions with outlying data (Zimmerman, 1998).

## Results

The plots on the right in Figure 8 show that the Type I error rate (the three solid bars on the left where $d = 0.0$) remained close to .05 for all three tests when no outliers were present in the normal distributed data. The three plots on the right show the power of the different tests for different true effect sizes. When no outliers were present in the samples, the power of the *t* test was somewhat higher than that of the MWW test, and the power of the MWW test was somewhat higher than that of the Y-W test. However, these differences were quite small.

The Rasch- and GRM-based sum scores show the same patterns. As there are many simulation results, we have placed the results in the online supplemental materials and show only one representative example (GRM-based, 40 items, and 40 subjects) in Figure 9. Without outliers, the Type I error rates of the three tests were comparable. Only when samples were small and data were skewed was the Type I error rate of the Y-W test too conservative, with a Type I error rate as low as .015 for a polytomous test with two items and a sample size of 20. The power of the Y-W and the MWW tests was only slightly lower than the power of the *t* test for a test that fitted the samples' latent trait distribution. Furthermore, the power of the MWW test is again slightly superior to the power of the Y-W test. When the sum scores are based on a difficult test (skewed distribution of sum scores) with polytomous items, the power of the MWW test is comparable or even slightly higher than the power of the *t* test.

Results for power and Type I error rate were quite different when outliers were present in the data. The Type I error rate (the three bars with shading lines on the right in Figure 8) of the *t* test was very low when sum scores were directly generated from the mixed-normal distribution, especially with smaller sample sizes. On the other hand, both the Y-W and the MWW tests kept the Type I error rate close to .05. Moreover, the power of the *t* test was also dramatically low, while the power of both the Y-W and the MWW tests remained good (only somewhat lower than without outliers in the data), with a small advantage for the MWW test.

We do not see the same devastating drop in power on the *t* test for Rasch- and GRM-based sum scores when outliers are present in the data (Figure 9). The outliers in the samples from the mixed-normal distribution were extreme and therefore had a profound influence on the performance of the *t* test. In the Rasch- and GRM-based simulations, even if the theta of the outlying case is very high or very low, the sum scores remain bounded by the number of questions and the number of answer options. Still, with test lengths of over 10 items or when the test was difficult (skewed sum scores), the Y-W and the MWW tests outperformed the *t* test. Furthermore, the MWW test performed

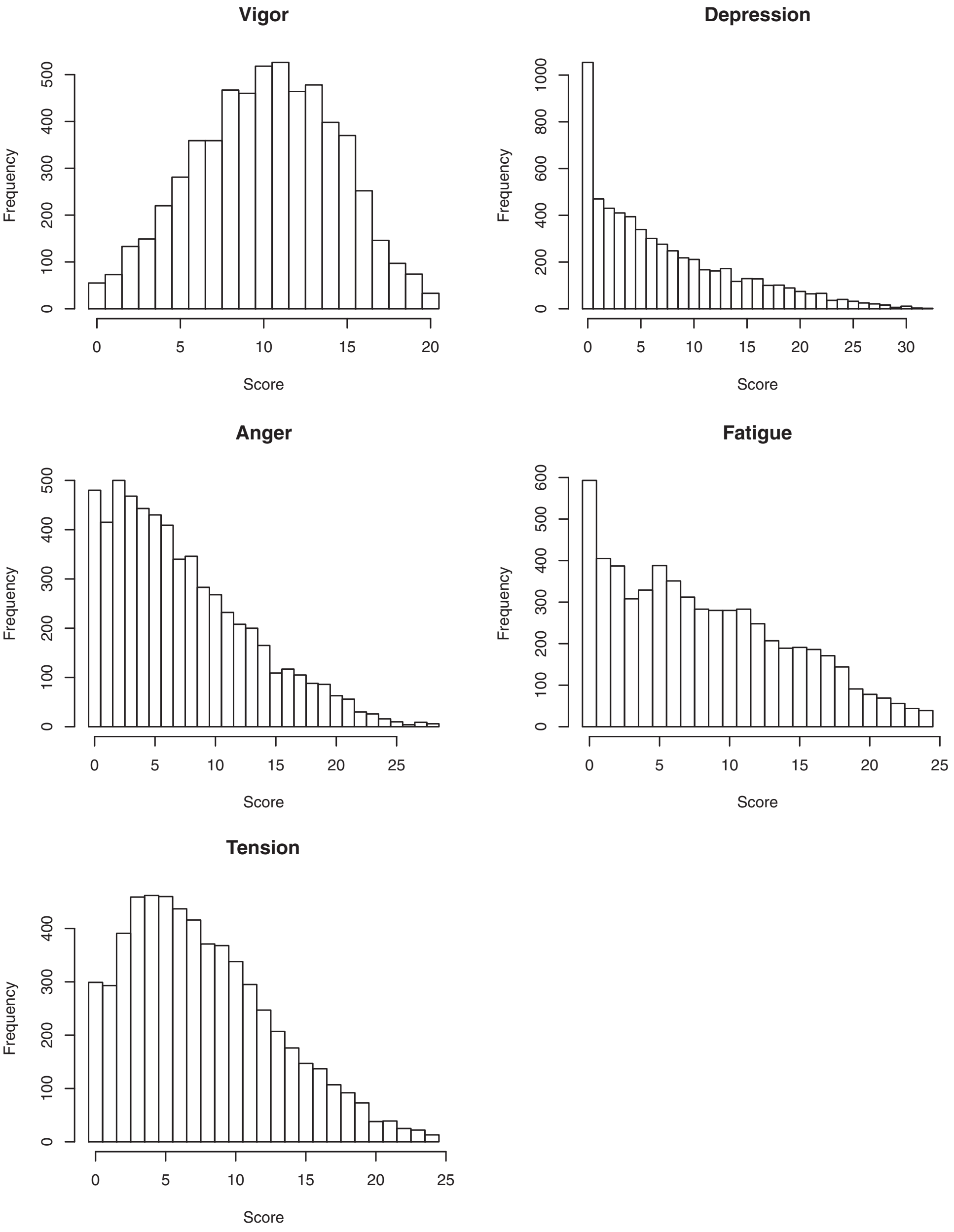

*Figure 6.* Distribution of the sum scores of the Profile of Mood States subscales.

somewhat better than the Y-W test, although the differences are small.[5]

Taken together, the simulations in Study 2 show that both the MWW and Y-W tests perform very well compared to the *t* test under most scenarios. When no outliers were present, the MWW, Y-W, and *t* tests provided similar Type I error rates and power. However, when outliers were present, both the Y-W test and the MWW test outperformed the *t* test in terms of power and Type I error rates.

## Discussion

Removing outliers before starting the actual analyses will result in smaller estimates of the standard error as opposed to not removing them. This, then, leads to an unjust underestimation of the Type I error rate (Wilcox, 2012). Nevertheless, our examination of articles in six psychological journals shows that the removal of outliers before applying the statistical analyses is a common practice, with 77% of reviewed articles that mentioned the word *outlier* doing it. Although various outlier-detection methods are used in practice, the most popular method is based on the

[5] The permutation test performed comparably to the *t* test. Cliff's method and the Brunner-Munzel procedure performed comparably to the MWW test, with a very small advantage for the Brunner-Munzel procedure. The bootstrap Y-W test performed comparably to the Y-W test.

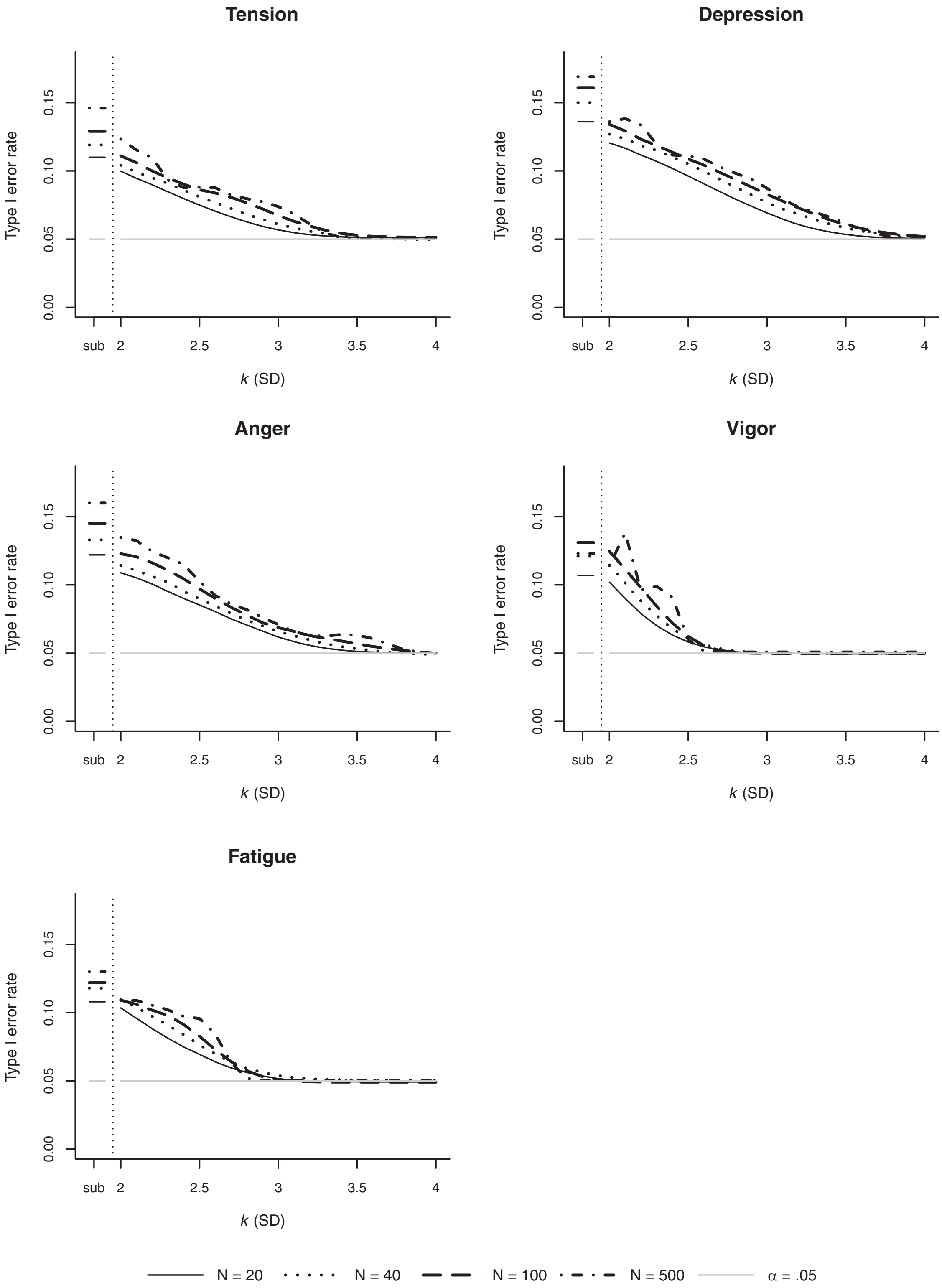

*Figure 7.* Type I error rate of a *t* test of sum scores based on the different Profile of Mood States subscales for different values of *k* and different sample sizes. Type I error rate of subjective use of *k* is presented on the left side of the vertical dotted line in each plot, and the horizontal gray line denotes the nominal Type I error rate (α = .05).

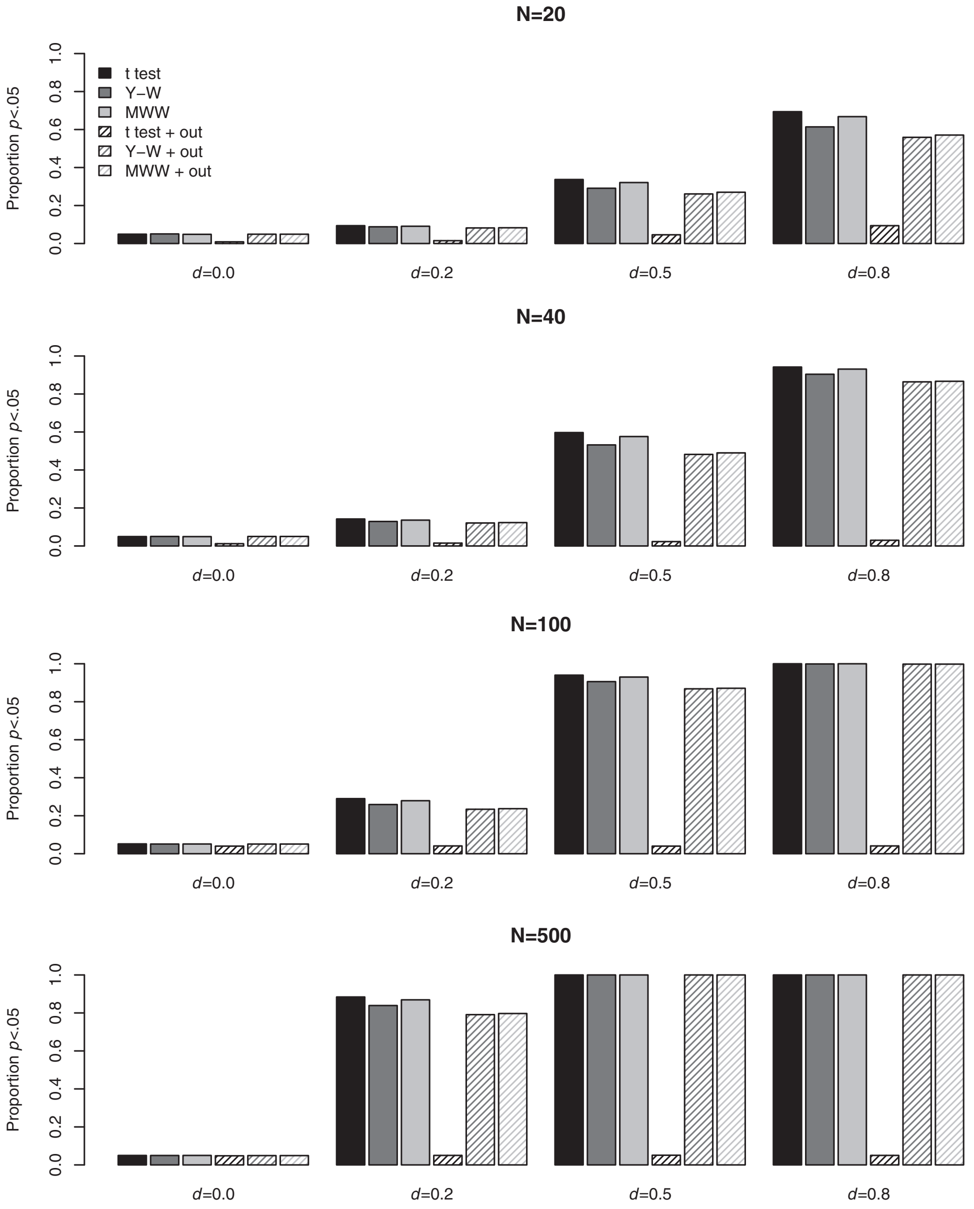

*Figure 8.* Proportion of statistically significant *p*-values of the *t* test, Yuen-Welch test (Y-W), and Mann-Whitney-Wilcoxon test (MWW) of normally distributed sum scores, with and without outliers, for different underlying effect sizes (*d*). The dark solid bar (*t* test; no outliers) is the nominal power. The three bars on the left represent the three methods without outliers, whereas the three bars on the right represent the same methods when outliers are added to the data.

$Z$ score, with $Z = 3$ as the most common threshold value. However, threshold values of $Z = 2$ were not uncommon.

We investigated the effect of outlier removal on the Type I error rate with a simulation study in which the data were nonnormal because of the use of sum scores. We simulated sum scores based on both dichotomous items and polytomous items, as analyses of sum scores on the basis of tests and questionnaires are common in psychological research. We also simulated more difficult tests, which result in skewed distributions, and used actual data to empirically confirm our results. Results suggest that with a threshold value of $Z = 3$ or larger and a not overly skewed distribution, the Type I error rate remains around the nominal value after removal of outliers from the data. However, when the distribution is skewed, even a threshold value of $Z = 3$ will inflate the Type I error rate substantially (to values of .175 in our simulations). Such a threshold for outlier removal is therefore not recommended for skewed distributions. Furthermore, as the distribution of psychological variables is often not normal (Micceri, 1989) and determining the actual form of the underlying distribution is difficult (especially with small data samples), we do not recommend using

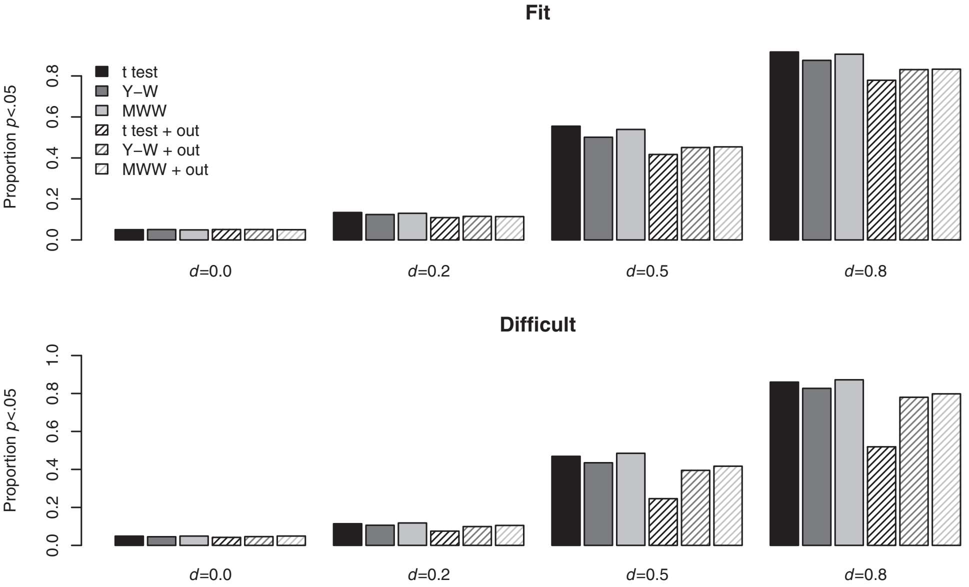

*Figure 9.* Proportion of statistically significant *p*-values of the *t* test, Yuen-Welch test (Y-W), and Mann-Whitney-Wilcoxon test (MWW) of graded response model–based sum scores based on a test with 40 items and 40 subjects, with and without outliers, for a fitting test (upper row) and for a difficult test (bottom row). out = outlier.

this threshold value. A threshold value of $Z = 2$, which is quite commonly used in practice, will inflate the Type I error rate up to .222 in our simulations.

Furthermore, our examination of several psychology journals suggests that thresholds for outlier deletion often appear chosen without clear external (or a priori determined) guidelines. Together with survey results from John et al. (2012), who found that 38% of psychological researchers admitted to having decided to exclude data after looking at the impact of doing so on the results, this suggests that the subjective use of the threshold for excluding outliers is not uncommon. Our simulation shows that subjective use of threshold values can inflate the Type I error rate to values as large as .452, especially when the population distribution is skewed. Furthermore, the influence of removing outliers on the Type I error rate could be even greater because we did not take into account differences in sample size and heteroscedasticity, which can also inflate the Type I error rate (Sawilowsky & Blair, 1992). Therefore, we do not recommend removing outliers before applying the actual analyses as this will lead to an increased Type I error rate. Moreover, when the (subjective) removal of outliers is combined with other commonly used QRPs, the Type I error rate will be inflated even more (Bakker et al., 2012; Simmons et al., 2011). Inflated Type I error rates result in the publication of false-positive findings. Particularly in combination with publication bias, such false-positive findings may be difficult to correct (Asendorpf et al., 2013; Bakker et al., 2012; Ferguson & Heene, 2012; Ioannidis, 2012; Pashler & Harris, 2012; Rosenthal, 1979).

On the other side, keeping extreme contaminants or extreme values from a nonnormal distribution in the data can lower the power of standard analyses, which makes it harder to detect a genuine effect. We saw an especially profound effect on the power of the *t* test when the scores were directly generated from a mixed-normal distribution (the power did not pass .1 for a large effect, while the power of the *t* test without outliers in the data lay between .61 and 1.00, depending on the sample size). Yet, as questionnaire data are bounded, extreme values of sum scores are also bounded. Especially in very short tests that fit the ability of the test-taker, the *t* test showed acceptable power even with the presence of outliers in the data. However, in longer tests and/or in difficult tests, the loss in power of the *t* test is substantial. Alternatives like the robust Y-W test and the nonparametric MWW test are less dependent on the actual distribution of the data and therefore less influenced by the presence of outliers in the data than the *t* test. Both the Y-W test and the MWW test had higher power than the *t* test (except for very short tests) when outliers were present and performed well under most conditions of the second simulation study. Furthermore, when no outliers are part of the data, the Type I error rate and the power of the Y-W and MWW tests appear to be comparable to the *t* test's. Therefore, both tests are a good replacement for the *t* test, also when no outliers are part of the data.

The MWW test has a somewhat higher power that the Y-W test, and the Y-W test can be somewhat conservative. Therefore, in the investigated situations, the MWW test is preferred above the Y-W test (but see Footnote 5 for two other alternatives). However, because the MWW test is sensitive to heteroscedasticity (especially when sample sizes differ; Erceg-Hurn & Mirosevich, 2008; Wilcox, 1998; Zimmerman, 1994, 1998), the Y-W test might be a better replacement of the *t* test when heteroscedasticity is expected. Heteroscedasticity is often encountered in psychological data. For example, Ruscio and Roche (2012) found a variance ratio between groups larger than three in 23% of the 453 published studies that they investigated. Heteroscedasticy is expected when groups differ in latent trait variances, but in other cases as well. Specifically, even if latent trait variances are equal, mean latent trait differences can lead to group differences in sum score variances if item difficulties are better suited for one group as opposed to the other group. For instance, if items in a test for depression are particularly well suited to measure depression, one could expect lower sum score variance in healthy samples for which items are relatively difficult (see, e.g., Wicherts & Johnson, 2009, who discussed this problem in the realm of variance decompositions). Further research is needed to investigate the performance of the Y-W test and the MWW test in the presence of outliers, combined with different sample sizes and heteroscedasticity.

Because we simulated sum scores based on item response models, another option would be to model the data with generalized item response models with groups as an additional predictor. However, the typically small sample sizes in psychological research will often not be large enough to estimate the parameters accurately. In addition, for questionnaire data, new ways to detect outliers are being developed that take into account the scores on all the different items (Zijlstra, Van der Ark, & Sijtsma, 2007). This will make it possible to better determine potential outliers in questionnaire data when sample sizes are sufficiently large.

In this study, we have focused on the comparison of the means of two groups, which is a basic and often-used research design in psychology. In correlational research (or in other research designs), the effect of outliers on the results of the statistical analyses can be quite profound too. In further research, these other research designs should be investigated as well. For many of these research designs, nonparametric (Gibbons & Chakraborti, 2003) or robust (Wilcox, 2012) statistical methods are available. However, not all methods currently have nonparametric or robust counterparts. Moreover, we did not consider in detail all existing alternatives to the *t* test when the data are skewed or when the data include outliers. Notably, Keselman et al. (2004) showed that the Y-W test can also be combined with bootstrapping and a transformation for skewness to control the Type I error rate. Other alternatives are a permutation test and two variants of the MWW test that take ties into account (Cliff, 1996; Brunner & Munzel, 2000; cf. Wilcox, 2012). In our additional simulations, these alternatives performed quite similarly to the standard MWW and Y-W tests (see Footnote 5). The MWW test can be executed in standard statistical packages like SPSS. The WRS package for R contains functions for the Y-W test and the other discussed tests.[6]

One might notice that the term *outlier* is only used in 7% of the articles that we examined. However, other terms and sentences like *extreme values* or *we removed all values with a Z value larger than* are probably also used to describe outlier identification and removal. This study was not designed to provide exact estimates of the different outlier handling methods but merely to give an indication of the common practice of removing outliers before the actual statistical analyses. Eighteen percent of the authors reported that they did analyses with and without the outliers. This is a better procedure and often recommended in statistical textbooks (e.g., Agresti & Franklin, 2007; Howitt & Cramer, 2011; Stevens, 2001), but if outcomes differ substantially, it is still not clear what to do.

In many of the investigated scenarios, the rise in Type I error rate is quite small and might therefore not be that influential. However, when these practices are combined with other common QRPs as described by John et al. (2012), the chance of finding a false-positive result can be equal to .50 (Bakker et al., 2012). Since conducting and publishing of replication studies are still not the standard, the correction of false-positive findings remains suboptimal and slow (Pashler & Harris, 2012). Researchers need to become aware of the influence of common decisions in the analysis of the results and use methods that minimize both the Type I and Type II error rates. Therefore, we present the following recommendations.

## Recommendations

- Correct or delete erroneous values.
- Based on prior research, it is not recommended to use *Z* scores to identify outliers. We recommend methods that suffer less from masking like the IQR or the MAD-median rule instead.
- Decide on outlier handling *before* seeing the results of the main analyses, and if possible, preregister the study at, for example, the Open Science Framework (http://opensciencefframework.org/).
- If preregistration is not possible, report the outcomes both with and without outliers or on the basis of alternative methods.
- Report transparently about how outliers were handled.
- Do not carelessly remove outliers as this increases the probability of finding a false positive, especially when using a threshold value of *Z* lower than 3 or when the data are skewed.
- Use methods that are less influenced by outliers like nonparametric or robust methods such as the Mann-Whitney-Wilcoxon test and the Yuen-Welch test, or researchers may choose to conduct bootstrapping (all without removing outliers).

Whenever there are likely outliers in the data or when data are nonnormal for the typical psychometric reasons we have described, these recommendations could help researchers properly control Type I and Type II error rates.

[6] The functions `yuen()`, `cidv2()`, `bmp()`, `permg()`, and `yuenbt()` of the WRS package for R were used to execute the Y-W test, Cliff's method, the Brunner-Munzel procedure, a permutation test, and a bootstrap version of the Y-W test, respectively.

## References

Agresti, A., & Franklin, C. (2007). *Statistics: The art and science of learning from data*. Upper Saddle River, NJ: Pearson Prentice Hall.

Aron, A., Aron, E. N., & Coups, E. J. (2009). *Statistics for psychology* (5th ed.). Upper Saddle River, NJ: Pearson.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119. doi:10.1002/per.1919

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543–554. doi:10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43,* 666–678. doi:10.3758/s13428-011-0089-5

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. Chichester, England: Wiley.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B.* Methodological, *57,* 289–300.

Bergman, L. R. (1988). You can't classify all of the people all of the time. *Multivariate Behavioral Research, 23,* 425–441. doi:10.1207/s15327906mbr2304_1

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and small-sample approximation. *Biometrical Journal, 42,* 17–25. doi:10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U

Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Erlbaum.

Collett, D., & Lewis, T. (1976). The subjective nature of outlier rejection procedures. *Applied Statistics, 25,* 228–237.

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology* (4th ed.). Harlow, England: Pearson.

Derogatis, L. R. (1994). *Symptom Checklist 90–R: Administration, scoring, and procedures manual* (3rd ed.). Minneapolis, MN: National Computer Systems.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12,* 105–120. doi:10.1037/1082-989X.12.1.105

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63,* 591–601. doi:10.1037/0003-066X.63.7.591

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7,* 555–561. doi:10.1177/1745691612459059

Field, A. (2013). *Discovering statistics using IBM SPSS Statistic* (4th ed.). London, England: Sage.

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). New York, NY: Norton.

Gibbons, J. D., & Chakraborti, S. (2003). *Nonparametric statistical inference*. Boca Raton, FL: CRC Press.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68,* 155–165. doi:10.1037/0022-006X.68.1.155

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. New York, NY: Wiley.

Hoekstra, R., Kiers, H. A. L., & Johnson, A. (2012). Assumptions for well-known statistical techniques: Disturbing explanations for why they are seldom checked. *Frontiers in Psychology, 3,* Article 137. doi:10.3389/fpsyg.2012.00137

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Howitt, D., & Cramer, D. (2011). *Introduction to statistics in psychology* (5th ed.). London, England: Pearson.

Huber, P. J. (1981). *Robust statistics*. doi:10.1002/0471725250

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* Article e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7,* 645–654. doi:10.1177/1745691612464056

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*. Advance online publication. Retrieved from http://ssrn.com/abstract=1996631

Keselman, H. J., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample *t* test. *Psychological Science, 15,* 47–51. doi:10.1111/j.0963-7214.2004.01501008.x

Langdridge, D., & Hagger-Johnson, G. (2009). *Introduction to research methods and data analysis in psychology* (2nd ed.). Harlow, England: Pearson.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., & Tucker Smith, C. (2013). PsychDisclosure.org: Grassroots support for reforming reporting standards in psychology. *Perspectives on Psychological Science, 8,* 424–432. doi:10.1177/1745691613491437

Lorr, M., McNair, D. M., & Droppleman, L. F. (1992). *Manual for the Profile of Mood States*. San Diego, CA: Educational and Industrial Testing Service.

MacDonald, P. (1999). Power, Type I, and Type III error rates of parametric and nonparametric statistical tests. *Journal of Experimental Education, 67,* 367–379. doi:10.1080/00220979909598489

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18,* 50–60. doi:10.1214/aoms/1177730491

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166. doi:10.1037/0033-2909.105.1.156

Miller, J. (1991). Short report: Reaction time analysis with outlier exclusion: Bias varies with sample size. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 43*(A), 907–912. doi:10.1080/14640749108400962

Moore, D. S., McCabe, G. P., & Craig, B. A. (2009). *Introduction to the practice of statistics* (6th ed.). New York, NY: Freeman.

Neuhäuser, M., Lösch, C., & Jöckel, K. H. (2007). The Chen-Luo test in case of heteroscedasticity. *Computational Statistics & Data Analysis, 51,* 5055–5060. doi:10.1016/j.csda.2006.04.025

Nolan, S. A., & Heinzen, T. E. (2007). *Statistics for the behavioral sciences*. New York, NY: Worth Publishers.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation, 9,* Article 6. Retrieved from http://PAREonline.net/getvn.asp?v=9&n=6

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536. doi:10.1177/1745691612463401

Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Pearson Assessment.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638–641. doi:10.1037/0033-2909.86.3.638

Ruscio, J., & Roche, B. (2012). Variance heterogeneity in published psychological research: A review and a new index. *Methodology, 8,* 1–11. doi:10.1027/1614-2241/a000034

Samejima, F. (1997). The graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). doi:10.1007/978-1-4757-2691-6_5

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110–114. doi:10.2307/3002019

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111,* 352–360. doi:10.1037/0033-2909.111.2.352

Shiffler, R. E. (1988). Maximum *Z* scores and outliers. *American Statistician, 42,* 79–80. doi:10.2307/2685269

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Staudte, R. G., & Sheater, S. J. (1990). *Robust estimation and testing*. doi:10.1002/9781118165485

Stevens, J. (2001). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Erlbaum.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Needham Heights, MA: Allyn & Bacon.

Taleb, N. N. (2007). *The black swan*. New York, NY: Random House.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 47*(A), 631–650. doi:10.1080/14640749408401131

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426–432. doi:10.1037/a0022790

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350–362.

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34,* 28–35.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6,* 291–298. doi:10.1177/1745691611406923

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE, 6,* Article e26828. doi:10.1371/journal.pone.0026828

Wicherts, J. M., & Johnson, W. (2009). Group differences in the heritability of items and test scores. *Proceedings of the Royal Society: Series B. Biological Sciences, 276,* 2675–2683. doi:10.1098/rspb.2009.0238

Wicherts, J. M., & Vorst, H. C. M. (2004). Modelpassing van de verkorte profile of mood states en meetinvariantie over mannen en vrouwen [Model fit of the shortened Profile of Mood States and measurement invariance across men and women]. *Nederlands Tijdschrift voor de Psychologie, 59,* 12–21. doi:10.1007/BF03062320

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods. *American Psychologist, 53,* 300–314. doi:10.1037/0003-066X.53.3.300

Wilcox, R. (2012). *Modern statistics for the social and behavioral sciences: A practical introduction*. Boca Raton, FL: CRC Press.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1,* 80–83. doi:10.2307/3001968

Wilson, S., & MacLean, R. (2011). *Research methods and data analysis for psychology*. Berkshire, England: McGraw-Hill.

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika, 61,* 165–170. doi:10.1093/biomet/61.1.165

Zijlstra, W. P., Van der Ark, A., & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research, 42,* 531–555. doi:10.1080/00273170701384340

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. *Journal of General Psychology, 121,* 391–401. doi:10.1080/00221309.1994.9921213

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67,* 55–68. doi:10.1080/00220979809598344

Zimmerman, D. W., & Zumbo, B. D. (1992). Parametric alternatives to the Student *t* test under violation of normality and homogeneity of variance. *Perceptual and Motor Skills, 74,* 835–844. doi:10.2466/pms.1992.74.3.835

(*Appendix follows*)

# Appendix

## Description of the Mann-Whitney-Wilcoxon Test and the Yuen-Welch Test

### Mann-Whitney-Wilcoxon Test

A nonparametric method of testing to compare two populations is the Mann-Whitney-Wilcoxon test (Mann & Whitney, 1947; Wilcoxon, 1945). Both independent samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ are put in ascending order. All values are replaced by ranks ranging from 1 to $m + n$. When there are tied groups, take the rank to be equal to the midpoint of the group. The ranks of each group are added, and then the lowest of these ranks is the test statistic $W$. $W$ can be transformed in a $Z$ score by

$$Z = \frac{W - \overline{W}}{SE_{\overline{W}}}, \text{ where}$$

$$SE_{\overline{W}} = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}.$$

The null hypothesis is rejected if

$$|Z| \geq z_{1-\alpha/2},$$

where $z_{1-\alpha/2}$ is the $1 - a/2$ quantile of standard normal distribution.

### Yuen-Welch Test

A robust method for comparing trimmed means is the Yuen-Welch Test (Yuen, 1974). $n_j$ is the sample size associated with the $j$th group, and $h_j$ is the number of observations left in the $j$th group after trimming. Put the remaining observations in ascending order yielding $X_{(1j)} \leq \ldots \leq X_{(nj)}$. The trimmed mean of the $j$th group can be estimated with

$$\overline{X}_{tj} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j - g_j} X_{(ij)},$$

and the Winsorized mean and variance with

$$\overline{X}_{wj} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{wij}, \text{ where}$$

$$\begin{aligned} X_{wij} &= X_{(g_j+1)j} \text{ if } X_{ij} \leq X_{(g_j+1)j} \\ &= X_{ij} \text{ if } X_{(g_j+1)j} < X_{ij} < X_{(n_j - g_j)j} \\ &= X_{(n_j - g_j)j} \text{ if } X_{ij} \geq X_{(n_j - g_j)j} \end{aligned}$$

$$s_{wj}^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_{wj})^2.$$

Yuen's test statistic is calculated with

$$T_y = \frac{\overline{X}_{t1} - \overline{X}_{t2}}{\sqrt{d_1 + d_2}}, \text{ where}$$

$$d_j = \frac{(n_j - 1) s_{wj}^2}{h_j (h_j - 1)}.$$

The degrees of freedom are

$$\widehat{V}_y = \frac{(d_1 + d_2)^2}{\frac{d_1^2}{h_1 - 1} + \frac{d_2^2}{h_2 - 1}}.$$

The null hypothesis is rejected if

$$|T_y| \geq t,$$

where $t$ is the $1 - a/2$ quantile of Student's $T$ distribution with $v_y$ degrees of freedom.

Received January 18, 2013
Revision received November 9, 2013
Accepted November 12, 2013 ■