



Normative comparisons for large neuropsychological test batteries: User-friendly and sensitive solutions to minimize familywise false positives

Hilde M. Huizenga, Joost A. Agelink van Rentergem, Raoul P. P. P. Grasman, Dino Muslimovic & Ben Schmand

To cite this article: Hilde M. Huizenga, Joost A. Agelink van Rentergem, Raoul P. P. P. Grasman, Dino Muslimovic & Ben Schmand (2016) Normative comparisons for large neuropsychological test batteries: User-friendly and sensitive solutions to minimize familywise false positives, *Journal of Clinical and Experimental Neuropsychology*, 38:6, 611-629, DOI: [10.1080/13803395.2015.1132299](https://doi.org/10.1080/13803395.2015.1132299)

To link to this article: <http://dx.doi.org/10.1080/13803395.2015.1132299>



Published online: 10 Apr 2016.



Submit your article to this journal [↗](#)



Article views: 108



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

Normative comparisons for large neuropsychological test batteries: User-friendly and sensitive solutions to minimize familywise false positives

Hilde M. Huizenga^{a,b,c}, Joost A. Agelink van Rentergem^a, Raoul P. P. P. Grasman^{a,b}, Dino Muslimovic^d and Ben Schmand^{a,b,e}

^aDepartment of Psychology, University of Amsterdam, Amsterdam, The Netherlands; ^bAmsterdam Brain and Cognition Center, University of Amsterdam, Amsterdam, The Netherlands; ^cResearch Priority Area Yield, University of Amsterdam, Amsterdam, The Netherlands; ^dDepartment of Medical Psychology, Groene Hart Hospital, Gouda, The Netherlands; ^eDepartment of Medical Psychology, Academic Medical Center, Amsterdam, The Netherlands

ABSTRACT

Introduction. In neuropsychological research and clinical practice, a large battery of tests is often administered to determine whether an individual deviates from the norm. We formulate three criteria for such large battery normative comparisons. First, familywise false-positive error rate (i.e., the complement of specificity) should be controlled at, or below, a prespecified level. Second, sensitivity to detect genuine deviations from the norm should be high. Third, the comparisons should be easy enough for routine application, not only in research, but also in clinical practice. Here we show that these criteria are satisfied for current procedures used to assess an overall deviation from the norm—that is, a deviation given all test results. However, we also show that these criteria are not satisfied for current procedures used to assess test-specific deviations, which are required, for example, to investigate dissociations in a test profile. We therefore propose several new procedures to assess such test-specific deviations. These new procedures are expected to satisfy all three criteria. **Method.** In Monte Carlo simulations and in an applied example pertaining to Parkinson disease, we compare current procedures to assess test-specific deviations (uncorrected and Bonferroni normative comparisons) to new procedures (Holm, one-step resampling, and step-down resampling normative comparisons). **Results.** The new procedures are shown to: (a) control familywise false-positive error rate, whereas uncorrected comparisons do not; (b) have higher sensitivity than Bonferroni corrected comparisons, where especially step-down resampling is favorable in this respect; (c) be user-friendly as they are implemented in a user-friendly normative comparisons website, and as the required normative data are provided by a database. **Conclusion.** These new normative comparisons procedures, especially step-down resampling, are valuable additional tools to assess test-specific deviations from the norm in large test batteries.

ARTICLE HISTORY

Received 27 March 2015

Accepted 11 December 2015



KEYWORDS

Criteria for abnormality in neuropsychology; familywise false-positive error rate; Bonferroni and Holm one-step and step-down resampling; normative comparisons; simulations

In neuropsychological assessment, an individual is often administered a large battery of tests (e.g., Arenas-Pinto et al., 2014; Binder, Iverson, & Brooks, 2009; Brooks, 2010; Crawford, Garthwaite, & Gault, 2007; Larrabee, 2014; Schretlen, Testa, Winicki, Pearlson, & Gordon, 2008; S. J. Wilson et al., 2015). The score on each test is then compared to its normative data, to assess deviations from the norm. This paper addresses the question of how to perform such

large battery normative comparisons in a valid and easy way, thereby facilitating routine application in neuropsychological research and in neuropsychological practice.

Large battery comparisons are ubiquitous. In clinical practice, they are used to inform diagnosis and/or guide tailored treatment (Lezak, Howieson, Bigler, & Tranel, 2012). In research, they serve two purposes. First, they may be used to classify participants into impaired versus nonimpaired groups.

CONTACT Hilde M. Huizenga  h.m.huizenga@uva.nl  Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129B, 1018 WS Amsterdam, The Netherlands

The authors do not report any financial or other relationships that could be interpreted as a conflict of interest affecting this manuscript.

© 2016 Informa UK Limited, trading as Taylor & Francis Group

These groups are then studied to investigate prevalence, demographic factors, biomarkers, or treatment effects (e.g., Meyer, Boscardin, Kwasa, & Price, 2013). Second, the classification into impaired versus unimpaired serves in some treatment effect studies as a dependent variable. That is, treatment effects are assessed not only in a continuous fashion—that is, whether a mean memory score improves under a new treatment as compared to treatment as usual—but also in a discrete manner—that is, whether the percentage of participants with a memory impairment reduced under a new treatment as compared to treatment as usual (cf. Kazdin, 2008; Kraemer & Kupfer, 2006).

Adequate procedures for normative comparisons of a single test have already been proposed (Crawford & Howell, 1998). These procedures have been extended in various ways—for example, to yield effect sizes and confidence intervals (Crawford & Garthwaite, 2002) and to account for background variables like an individual's age or level of education (Crawford & Garthwaite, 2006). In such single test normative comparisons, a test score falling below a percentile criterion of the normative data is considered to be abnormal. For example, the percentile criterion may be set at 5%. This 5% criterion implies that the false-positive error rate—that is, the chances of deciding that an individual deviates from the norm whereas she or he actually does not—is 5%.¹

In case of large test batteries, the 5th percentile criterion implies that the false-positive error rate is 5% for *each* test separately, corresponding to a specificity of 95%. These false-positive errors accumulate when multiple tests are administered, yielding an *overall* false-positive error rate, the familywise false-positive error rate, which will exceed 5%. More specifically, if M tests are administered, the familywise false-positive error rate, from now on the familywise error, is $[1 - (1 - 0.05)^M] \times 100\%$, provided that tests are uncorrelated in the normative sample. For example, the familywise error for $M = 13$ uncorrelated tests is then 49%. That is, a *healthy* individual has a 50–50 chance to be classified as deviating on at least one test (cf. Huizenga, Smeding, Grasman, & Schmand, 2007). Although this familywise error will be lower if tests are correlated in the normative sample, it will often substantially exceed 5% (Crawford et al., 2007; Huizenga et al., 2007).

There is an increasing awareness in the neuropsychological community that it is necessary to control familywise error at prespecified levels. This awareness is present in the group means testing context, where it is, for example, tested whether group means differ (two-sample t tests) or whether group means differ from a hypothesized value (one-sample t tests) on multiple neuropsychological tests. In such a group means testing context, it has been argued that a lack of control over familywise error may give rise to overinterpretation of chance findings (e.g., Bell, Olivier, & King, 2013; Blakesley et al., 2009; Eichstaedt, Kovatch, & Maroof, 2013; Levav et al., 2002; Lewis, Maruff, Silbert, Evered, & Scott, 2006; Schatz, Jay, McComb, & McLaughlin, 2005; C. E. Wilson et al., 2014; cf. Ioannidis, 2005; Miguel et al., 2014; Simmons, Nelson, & Simonsohn, 2011). In an excellent review specifically aimed at the neuropsychological community, Blakesley et al. (2009) reviewed several procedures to control familywise error in the group means testing context. They studied, for example, the well-known Bonferroni procedure, the Holm procedure (Holm, 1979), and various resampling procedures (Westfall & Young, 1993). Simulation studies indicated that these alternative procedures all controlled familywise error.

The familywise error issue also is prominent in the normative comparisons context (e.g., Berthelson, Mulchan, Odland, Miller, & Mittenberg, 2013; Bilder, Sugar, & Helleman, 2014; Brooks, 2010; Crawford et al., 2007; Davis & Millis, 2014; Larrabee, 2008, 2014; Loewenstein et al., 2006; Meyers et al., 2014; Naglieri & Paolitto, 2010; Palmer, Boone, Lesser, & Wohl, 1998; Proto et al., 2014; Schretlen et al., 2008). It has been argued that in clinical practice, lack of control over familywise error in normative comparisons may result in overdiagnosis and unnecessary treatment, increasing patient burden and unnecessary costs to the health care system (Binder et al., 2009; Brooks, Iverson, Holdnack, & Feldman, 2008; Gisslén, Price, & Nilsson, 2011; Torti, Focà, Cesana, & Lescure, 2011). In neuropsychological research, it has been argued that lack of control has two disadvantages. First, if normative comparisons are used to assign participants to impaired and nonimpaired groups, lack of control will lead to

¹Although we adhere in this paper to a required false-positive rate of 5%, other prespecified rates may also be imposed, without loss of generality.

the inclusion of false positives into the impaired sample, resulting in heterogeneity, and thus in less powerful studies of, for example, prevalence, risk factors, biomarkers, and treatment effects (Blackford & La Rue, 1989; Brooks, Iverson, Feldman, & Holdnack, 2009; Höfler, 2005; Meyer et al., 2013). Second, if normative comparisons are used to assess deviations from the norm after treatment, lack of control may lead to the conclusion that many participants still deviate from the norm, whereas the treatment was actually quite effective. So, we require that procedures for normative comparisons control familywise error at prespecified levels.

We also require that procedures have adequate sensitivity to detect genuine deviations from the norm. Detection of genuine deficits is important in neuropsychological research. First, it offers the opportunity to precisely investigate prevalence and progression of these deficits. Second, it allows identification of all deficits associated with a disorder, thereby offering the opportunity to gain more insight into the mechanisms underlying the disorder (Lezak et al., 2012). Detection of genuine deficits is also important in neuropsychological clinical practice, as it offers the opportunity to target interventions to these deficits (e.g., Constantinidou, Wertheimer, Tsanadis, Evans, & Paul, 2012; Sander, Nakase-Richardson, Constantinidou, Wertheimer, & Paul, 2007).

In addition to these familywise error and sensitivity criteria, we also require that procedures are easy to apply, as they should offer the possibility of routine application in neuropsychological assessment, not only in research but also in clinical practice. Procedures that are not user-friendly because they require a statistical background and programming skills and/or large normative datasets will not be used very often. Therefore, we require that a procedure should be user-friendly.

Before reviewing potential procedures that may satisfy the three criteria, it is informative to make a distinction between two main aims of large battery comparisons (cf. Huberty & Morris, 1989). First, large battery comparisons are used to classify individuals as *overall* impaired or unimpaired given all tests. Second, large battery comparisons are also used to provide *test-specific* classifications as impaired or unimpaired—for example, to investigate dissociations in the test profile. For example, in test-specific classifications an individual may be classified as impaired on a memory test but as

unimpaired on the other neuropsychological tests. In the following we review whether current procedures for overall and test-specific classification satisfy the familywise error, sensitivity, and user-friendliness criteria.

Overall classification as impaired or unimpaired

One procedure for overall classification is to require deviations on multiple tests (e.g., Arenas-Pinto et al., 2014; Axelrod & Wall, 2007; Grunseit, Perdices, Dunbar, & Cooper, 1994; Ingraham & Aiken, 1996; Proto et al., 2014). Several approaches have been adopted to determine the number of required deviations. Among them, approaches taking the dependency between test scores into account are to be preferred (Berthelson et al., 2013; Crawford et al., 2007; Muslimovic, Post, Speelman, & Schmand, 2005; Schagen, Muller, Booger, Mellenbergh, & van Dam, 2006; Schretlen et al., 2008), as they satisfy the three criteria (e.g., Crawford et al., 2007). That is, they control familywise error at prespecified levels, have adequate sensitivity, and are relatively easy to apply as software exists to determine the number of required deviations (Crawford, 2016).

A second procedure for overall classification is to perform a multivariate normative comparison (e.g., Cohen et al., 2015; González-Redondo et al., 2012; Smeding, Speelman, Huizenga, Schuurman, & Schmand, 2011; Su et al., 2015). In a multivariate comparison, it is determined whether an entire test profile—that is, an individual's combination of test scores—differs from that in the normative sample (Crawford & Allan, 1994; Grasman, Huizenga, & Geurts, 2010; Huba, 1985; Huizenga et al., 2007). This method satisfies the three criteria (Huizenga et al., 2007). That is, familywise error is controlled, sensitivity is adequate, and it is easy to apply as the procedure is implemented in a webpage (Multivariate normative comparisons, 2016).

In sum, the overall classification procedures satisfy the three criteria. However, this is not the case for current test-specific classification procedures, as we outline next.

Test-specific classification as impaired or unimpaired: Current procedures

The first common procedure for test-specific classifications is to perform uncorrected comparisons

—that is, to treat each test as if it was the only test that was administered. As indicated earlier, these uncorrected comparisons do not control familywise error. As a result, sensitivity is very high. The procedure is very user-friendly, as no additional computations are required. So uncorrected comparisons do not satisfy the familywise error criterion, yet they do satisfy the sensitivity and user-friendliness criteria.

The second procedure is a Bonferroni normative comparison (e.g., Huizenga et al., 2007). If tests are uncorrelated in the normative sample, this correction yields a familywise error never exceeding 5%. However, if test scores are correlated, which is much more common, Bonferroni correction results in a familywise error that is too low and, consequently, with a decreased sensitivity to detect genuine deviations from the norm (e.g., Huizenga et al., 2007). Therefore, Bonferroni normative comparisons satisfy the familywise error criterion, but the sensitivity criterion is not satisfied. The user-friendliness criterion is satisfied, as the procedure is relatively simple to apply.

Test-specific classification as impaired or unimpaired: New procedures

As uncorrected and Bonferroni normative comparisons do not satisfy all criteria, we propose three alternatives: Holm, one-step, and step-down resampling normative comparisons. Below we only indicate whether these procedures are likely to satisfy the three criteria; the procedures are described in more detail in the Method section.

The first new procedure is based on the Holm method (Holm, 1979). In the usual group means testing context, it has been shown that Holm controls familywise error. It has also been shown that the Holm method is characterized by higher sensitivity than Bonferroni, although sensitivity is still too low if test scores are correlated (Blakesley et al., 2009; Eichstaedt et al., 2013; Holm, 1979). Up to now the Holm method has only been applied in the group means testing context, but we will show that it can easily be extended to normative comparisons. In order to promote user-friendliness, we implemented Holm normative comparisons in a user-friendly Normative Comparisons website (Agelink van Rentergem & Huizenga, 2016).

The second new procedure is based on one-step resampling (Blakesley et al., 2009; Nichols & Holmes, 2002 for a general introduction; Westfall & Young, 1993 for a more specific treatment). In the group means testing context, it has been shown that one-step resampling controls familywise error and outperforms Bonferroni in terms of sensitivity if test scores are correlated. Up to now, one-step resampling has only been applied in the group means testing context, but we will show that it can easily be extended to normative comparisons.

The third new procedure is based on step-down resampling (Westfall & Young, 1993). In the mean testing context, it has been shown that step-down resampling controls familywise error and outperforms one-step resampling in terms of sensitivity. We will again show that generalization to the normative comparisons context is easy.

With respect to user-friendliness of the resampling approaches, two important issues deserve attention. First, the resampling normative comparisons procedures require experience with programming, for example in R (R Core Team, 2015) and therefore are not user-friendly. To address this, we implemented them in the user-friendly Normative Comparisons website (Agelink van Rentergem & Huizenga, 2016). A second issue relates to the fact that resampling normative comparisons require access to raw normative data; means and standard deviations of normative data are not sufficient. Raw normative data are generally available in research settings, as scientific studies often compare patient samples to healthy control samples. However, in neuropsychological practice, raw normative data are usually unavailable. To address this issue, we aggregated healthy control data from neuropsychological scientific studies into a single database. This database will be made available, without any costs, for qualified² neuropsychologists in the very near future (ANDI; Advanced Neuropsychological Diagnostics Infrastructure, 2016). Currently, investigators of 90 studies donated healthy control data of over 25,000 participants together completing over 50 neuropsychological tests. This offers the possibility to provide the normative data required for resampling normative comparisons.

²In the first year after release of the database, Dutch qualified neuropsychologists will be given access. Qualifications can be checked easily, as every licensed neuropsychologist is registered by the Dutch ministry of health (BIG register, 2016). After this first year, international extensions will be considered.

We first outline the new normative comparison procedures in more detail. We then report the results of a Monte Carlo simulation study in which we compared the usual uncorrected and Bonferroni normative comparisons to the new Holm, one-step resampling, and step-down resampling normative comparisons. In these simulations we assess familywise false-positive error and the sensitivity to detect genuine deviations from the norm. We also illustrate the normative comparisons website with an application to the neuropsychological evaluation of patients with Parkinson disease (Muslimovic et al., 2005). Finally, we summarize results and discuss potential limitations and solutions.

Method

We first describe a single normative comparison and then proceed with Bonferroni, Holm, one-step resampling, and step-down resampling. More detail and computer code are given in the Appendix.

Normative comparisons: Single neuropsychological test

First, consider a single neuropsychological test used to compare an individual to a normative sample of N persons. Let x denote the score of the individual, and let y_n , with $n = 1, \dots, N$, denote scores in the normative sample. It is convenient (cf. Appendix) to center normative scores and the individual's score at the normative sample mean \bar{y} . That is, $y_n^* = y_n - \bar{y}$, and $x^* = x - \bar{y}$, where $*$ denotes that a variable is centered. The statistic required for a single normative comparison equals (Crawford, Howell, & Garthwaite, 1998; Sokal & Rohlf, 1995):

$$t_{\text{norm}} = \frac{x^* - \bar{y}^*}{\text{sd}(y^*)/\sqrt{N}} \times \text{scaling factor} \quad (1)$$

Note that \bar{y}^* equals zero due to centering. In equation (1), $\text{sd}(y^*)$ denotes the usual estimate of the standard deviation of y^* :

$$\text{sd}(y^*) = \sqrt{\frac{\sum_{n=1}^N (y_n^* - \bar{y}^*)^2}{(N-1)}} \quad (2)$$

The scaling factor equals $1/\sqrt{N+1}$. To understand why this is the case, suppose first it instead equals 1. In that case, equation (1) is the common one-sample t_{test} statistic, used to test whether x^* differs from the mean \bar{y}^* . More specifically, in t_{test} , $x^* - \bar{y}^*$ is divided by the standard deviation of the mean \bar{y}^* , that is, by its standard error $\text{sd}(y^*)/\sqrt{N}$:

$$t_{\text{test}} = \frac{x^* - \bar{y}^*}{\text{sd}(y^*)/\sqrt{N}} \quad (3)$$

However, in the current normative comparisons context, we do not aim to test whether x^* deviates from the *mean* \bar{y}^* , but to test whether it deviates from the *distribution* of y^* . Therefore $x^* - \bar{y}^*$ should not be divided by the standard deviation of the mean \bar{y}^* , but by the standard deviation of the *distribution* of y^* , that is, by $\text{sd}(y^*)$. This is effectuated by setting the scaling factor in equation (1) roughly equal to $1/\sqrt{N}$ instead of 1. More precisely it should equal $1/\sqrt{N+1}$ (for an extensive treatment: Sokal & Rohlf, 1995, p. 227–228).

Whereas t_{test} is used to determine whether a value deviates from the *mean* of a distribution of observations (group means testing context), t_{norm} is used to determine whether a value deviates from a *distribution* of observations (normative comparisons context). In both contexts, the statistics t_{test} and t_{norm} have to be compared to the distribution of t_{test} under the null hypothesis $x^* - \bar{y}^* = 0$ (Crawford et al., 1998). This is the Student t distribution with $N-1$ degrees of freedom. So, if we aim to determine whether a score deviates from the norm, we compare the t_{norm} statistic to the distribution of t_{test} under the null hypothesis, and the resulting p -value is indicative of the abnormality of t_{norm} . If t_{norm} is located in the outer tails of this distribution, we decide that the score deviates from the norm. The choice of a critical value for the outer tails determines the false-positive rate.³ For example, in the case of a one-sided normative comparison, testing the hypothesis that an individual scores less than the norm, a critical value of .05 for the lower tail yields a false-positive rate of 5%.

This close resemblance between group means testing and normative comparisons—statistics differ by a scaling factor but the required distribution is the same—allows us to extend procedures from a group means testing context to a normative comparisons context, as is outlined next.

³Provided that the usual assumptions of a t test are met.

Bonferroni normative comparisons

If a familywise error of 5% is desired and if M neuropsychological tests are administered, the p -values (cf. previous section) of all t_{norm} statistics are multiplied by M . This yields the Bonferroni corrected p -values.

Holm normative comparisons

The Holm procedure (cf. Holm, 1979 for the group means testing context) is a so-called step-down version of the Bonferroni procedure. Correction proceeds in two steps: from p -values to step-down p -values, and from step-down p -values to corrected p -values. First, the p -value of the largest absolute t_{norm} statistic is multiplied by M , the second largest by $(M - 1)$, and so on. This yields step-down p -values. Thereafter, a correction is applied, ensuring that smaller absolute t -statistics do not have smaller p -values than larger absolute t -statistics. To accomplish this, the corrected p -value of a t_{norm} statistic is the maximum of its step-down p -value and the corrected p -values of larger absolute t_{norm} statistics.

One-step resampling normative comparisons

In uncorrected comparisons, the t_{norm} statistic is compared to the distribution of t_{test} under the null hypothesis. In one-step resampling normative comparisons, the absolute t_{norm} statistic is compared to the distribution of the *maximum* over M absolute t_{test} statistics under the null hypothesis (cf. Nichols & Holmes, 2002; Westfall & Young, 1993, for the mean-testing context). Whereas the distribution of t_{test} under the null hypothesis is known (the Student t distribution), the distribution of the maximum over M absolute t_{test} statistics, the so-called max distribution, is unknown and therefore has to be obtained by resampling (cf. Nichols & Holmes, 2002). That is, by resampling the original dataset it is possible (cf. Appendix) to create a new dataset that satisfies the null hypothesis of no differences between x^* and \bar{y}^* on any of the M neuropsychological tests. From this new dataset, we determine and store the maximum over its M absolute t_{test} statistics. This resampling procedure is repeated many—for example, 2000—times, thereby generating 2000 maximum absolute t_{test} statistics under the null hypothesis and thus the required max distribution (cf. Figure 1). After this

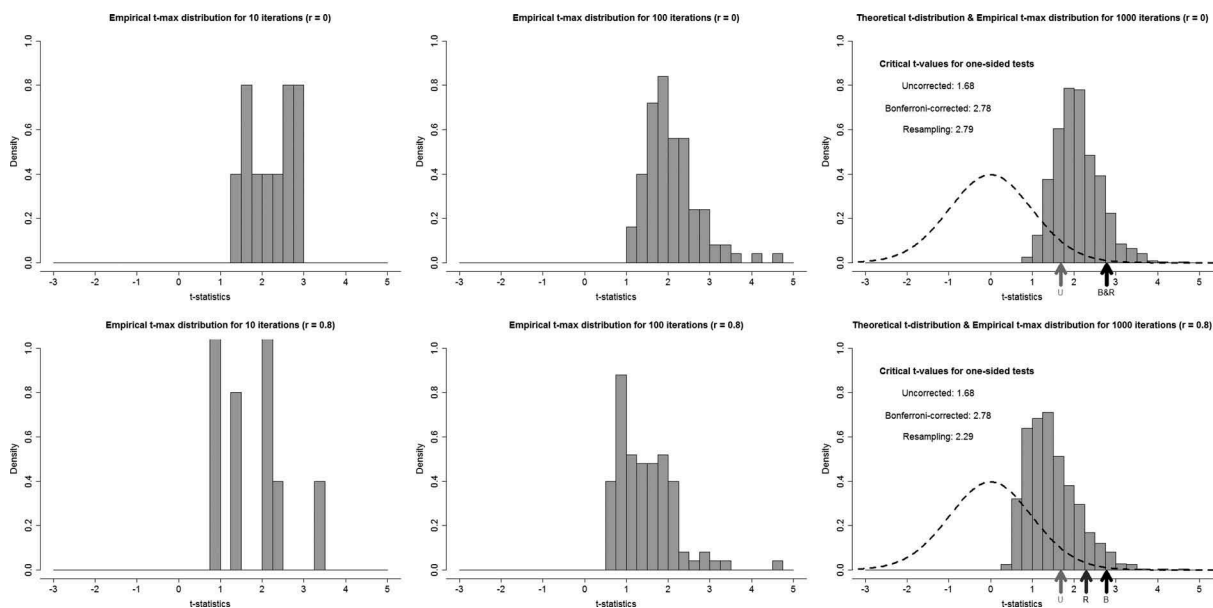


Figure 1. An illustration of the one-step resampling approach. This figure contains max distributions obtained in a condition where the normative sample consists of 50 participants and where 13 uncorrelated tests (top row) or 13 correlated tests (bottom row) have been administered. In each row, the three figures refer to max distributions derived from 10, 100, and 1000 resamples: It can be seen that smoothness of the distribution increases with an increasing number of resamples. The theoretical Student- t distribution is depicted in the max distribution derived from 1000 resamples. If tests are uncorrelated, Bonferroni and resampling critical values (arrows) are equal. If tests are correlated, the resampling critical value is less stringent.

max distribution has been obtained, each of the M absolute t_{norm} statistics is compared to the max distribution. A more technical description is given in the [Appendix](#).

Step-down resampling normative comparisons

In step-down resampling normative comparisons (Westfall & Young, 1993, for the mean-testing context), the largest absolute t_{norm} statistic is compared to the max distribution over all M neuropsychological tests, as in the one-step resampling procedure. However, the next largest absolute t_{norm} statistic is referred to the max distribution computed from all neuropsychological tests except the one giving rise to the largest absolute t_{norm} statistic. The second next largest statistic is referred to the max distribution computed from all neuropsychological tests except the first two, and so on. Afterwards, a correction is applied, ensuring that smaller absolute t -statistics do not have smaller p -values than larger absolute t -statistics, akin to the correction used in the Holm procedure. Please refer to the [Appendix](#) for the technical description.

Monte Carlo simulations

The goal of the simulations was to assess familywise error (i.e., the complement of specificity) and sensitivity for uncorrected, Bonferroni, Holm, one-step, and step-down resampling comparisons. In the resampling comparisons we derived the max distributions by computing 2000 resamples.

Simulation method

We simulated multivariate normally distributed data for 50 persons as the normative sample, and data for one individual that was compared to this normative sample (cf. for a similar approach,

Crawford & Garthwaite, 2006; Huizenga et al., 2007). This procedure was repeated 5000 times in each simulation condition. We combined three factors. First, we included conditions with either 10 or 30 neuropsychological tests. Second, we included conditions in which correlations between tests in the normative sample were set to .0, .5, or .8. Third, we simulated a difference from the norm by giving the individual a score of 0, 2, 2.5, 3, 3.5, or 4 standard deviations from the normative data mean. In the case of a difference from the norm, this difference was present on the first five neuropsychological tests. For example, in the case of 30 tests, a difference—for example, of 3 standard deviations—was present on the first five tests, but not on the remaining 25.

The normative comparisons procedures were implemented as outlined in the R-code in the [Appendix](#). In one-step and step-down resampling, we computed 2000 resamples.

An estimate of familywise error was obtained from conditions in which there was no simulated difference between the individual and the normative sample. Familywise error was defined as the percentage of simulations in which one or more of the test results indicated a deviation from the norm. An estimate of sensitivity was obtained from conditions in which there was a simulated difference. Sensitivity was defined as the percentage of simulations in which the individual deviated on the first test.

Simulation results

[Table 1](#) indicates that familywise error differs markedly between uncorrected comparisons and the other types of comparisons. Uncorrected comparisons are characterized by too high familywise error. In the worst case, in which 30 uncorrelated tests are administered, it is nearly 80% instead of the intended 5%. Although familywise error decreases with the number of tests and with the

Table 1. Familywise error rate as a function of the number of neuropsychological tests and correlations between these tests in the normative sample.

No. of tests	Correlation	Uncorrected (%)	Bonferroni (%)	Holm (%)	Resampling	
					One-step (%)	Step-down (%)
10	0	40.0	5.4	5.4	5.5	5.6
10	.5	25.8	3.9	3.9	4.9	4.9
10	.8	14.9	2.1	2.1	4.8	4.8
30	0	78.5	5.1	5.1	5.3	5.3
30	.5	41.2	3.4	3.4	5.2	5.2
30	.8	21.2	1.3	1.3	4.6	4.6

Note. Familywise error rate should be 5%.

correlation between them, the most favorable condition—that is, 10 tests that are .8 correlated—still yields a familywise error of about 15%. Bonferroni and Holm comparisons are characterized by a familywise error at or below 5%. One-step and step-down resampling comparisons always have a familywise error of about 5%. So Bonferroni, Holm, and one-step and step-down resampling, but not the usual uncorrected comparisons, keep familywise error at or below 5%.

Sensitivity is depicted in Figure 2. Although uncorrected comparisons are characterized by an unacceptably large familywise error, their results are plotted to provide some sort of upper bound to attainable sensitivity. First consider the situation in which test scores are uncorrelated (left-hand panels). In these cases all procedures have equal sensitivity.

Second, if variables are correlated (middle and right-hand panels), resampling comparisons are characterized by highest sensitivity, with step-down resampling slightly outperforming one-step resampling.

In sum, among the procedures with an acceptable familywise error, step-down resampling has to be preferred as it has the highest sensitivity. As compared to Bonferroni, a sensitivity advantage up to 20% can be attained.

Illustrative application

Muslimovic et al. (2005) compared the cognitive profile of 115 patients with newly diagnosed Parkinson disease to that of 70 healthy controls. As an illustration we compare each of these patients

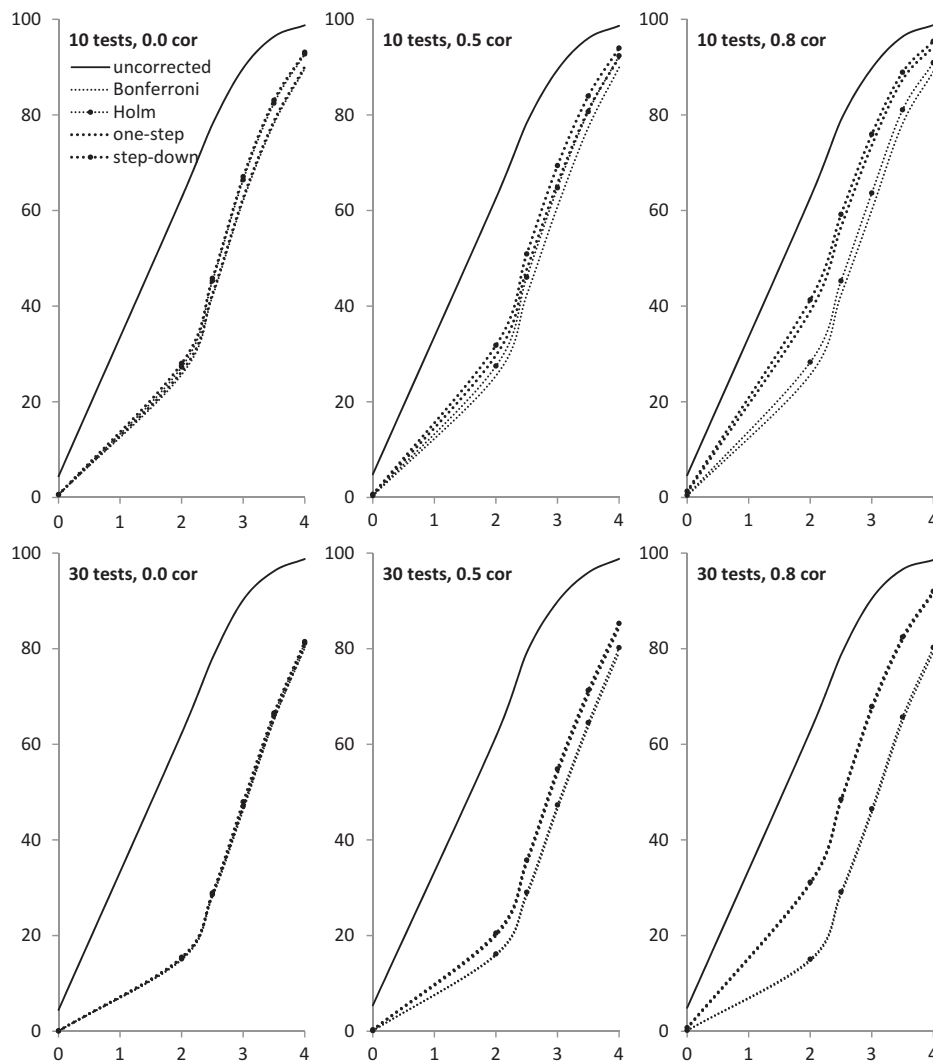


Figure 2. Sensitivity as a function of correlations in the normative sample and as a function of the magnitude of the simulated differences to the norm.

to the control sample using the common uncorrected and Bonferroni normative comparisons, and the new Holm, one-step resampling, and step-down resampling normative comparisons.

Twenty-three neuropsychological test variables were included in the analysis. Only participants with complete data were included, leaving 84 patients and 65 controls for further analysis. The patient and control samples differed significantly in age; therefore we used scores that were standardized with respect to published norms, or which were standardized by means of a regression approach (for further details on standardization: Muslimovic et al., 2005). All normative comparisons were one-sided, because we hypothesized that patients perform worse than the control sample. We required that individual scores were located below the usual 5th percentile—that is, we used the $\alpha = .05$ criterion.

The average correlation between variables was not very high (.15), but some variables correlated in the .6–.9 range (cf. Figure 3). Therefore we expected the resampling approaches, as compared to Bonferroni and Holm, to show a higher percentage of deviations.

Uncorrected comparisons reveal that 89% of the newly diagnosed Parkinson patients show a deviation on at least 1 neuropsychological test variable. This percentage is 17% for Bonferroni and Holm and 19% for one step and step-down resampling.

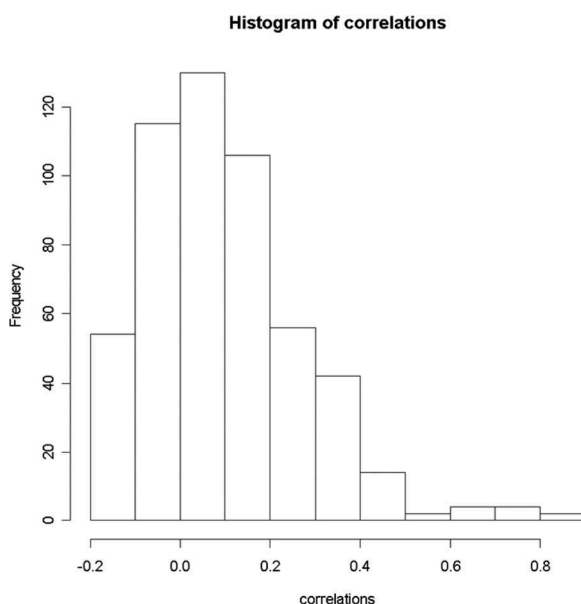


Figure 3. Histogram of correlations between normative test scores in the empirical illustration.

Two patients are not classified as deviating with Bonferroni and Holm, but are so with the resampling approaches.

As an illustration, consider how one of these patients, patient 3075, is analyzed with the Normative Comparisons website (Figure 4). Input options are displayed on the left, whereas output, both in graphical form (Figure 4, upper panel) and in tabular form (Figure 4 lower panel), is displayed on the right. With respect to input, we uploaded two datasets, one for controls and one for patients, containing ID numbers and test scores. We also selected the type of normative comparisons: step-down resampling, one-sided comparisons, deciding whether scores are lower than the norm, with the usual $\alpha = .05$ criterion. The graphical output (Figure 4, upper panel) and matching tabular output (Figure 4, lower panel) indicates that this patient deviates on the Tower of London test, but not on the other tests.

Discussion

Large battery normative comparisons are ubiquitous in neuropsychological practice and research. Therefore, it is important that these comparisons are carried out in a valid, sensitive, and user-friendly way. First, adequate large battery normative comparisons should control familywise false-positive error rate at a prespecified level in order to guarantee high specificity. Second, they should have sufficient sensitivity to detect genuine deviations from the norm. Third they should be user-friendly to allow routine application in neuropsychological practice and research. We noted that several procedures for *overall* normative comparisons satisfy these three criteria, but that current standard procedures for *test-specific* comparisons do not. Therefore, the aim of the current paper was to develop test-specific normative comparisons procedures meeting all three criteria. We compared these new procedures to standard procedures by means of simulations and by means of an empirical example.

Results of our simulation study indicate that traditional uncorrected comparisons do not control familywise false-positive error. In the worst case, a familywise error approaching 80% instead of the intended 5% was observed. Only the Bonferroni, Holm, one-step resampling, and step-down resampling procedures control familywise error at or below 5%. Resampling outperforms

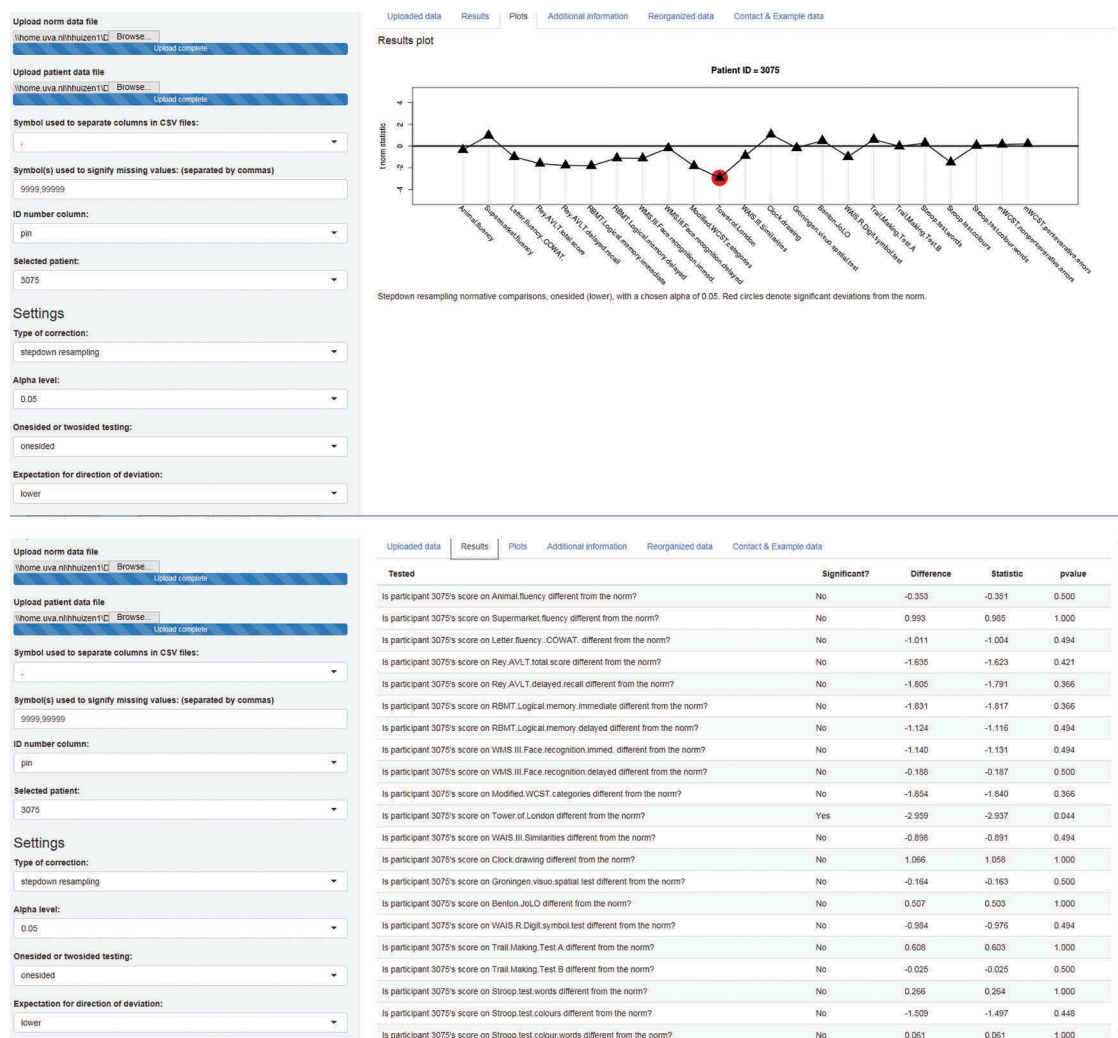


Figure 4. Illustration of the Normative Comparisons website. To view a color version of this figure, please see the online issue of the Journal.

Bonferroni in terms of sensitivity, with a slight advantage of step-down resampling over one-step resampling. Our simulations indicate that a sensitivity advantage of up to 20% over Bonferroni can be obtained. Let us suppose that the sensitivity advantage is 10%. This implies that an additional 10 out of 100 individuals will be correctly characterized as deviating from the norm. In neuropsychological practice, these individuals may then, for example, profit from interventions, which otherwise would not be available to them. In neuropsychological research, this heightened sensitivity will offer the opportunity to gain more insight into the mechanisms underlying a disorder (Lezak et al., 2012).

The increase in sensitivity as compared to Bonferroni depends on the magnitude of

correlations between neuropsychological tests. It is difficult to give a general indication of the sensitivity advantage that is to be expected in neuropsychology, since the magnitude of correlations is unknown in most situations. In our Parkinson example, correlations varied between $-.20$ and $.80$, and the average correlation was $.15$. Although the average correlation was small, resampling methods did classify two individuals as deviating that Bonferroni did not.

Several issues deserve attention. First, we only investigated performance of procedures for normally distributed normative data. Crawford, Garthwaite, Azzalini, Howell, and Laws (2006) indicated that the t -statistic approach, which lies at the heart of uncorrected, Bonferroni, and Holm normative comparisons, is affected by non-

normality (cf. Grasman et al., 2010). Resampling approaches to mean testing are generally less affected by non-normality than *t* tests. Therefore, resampling approaches to normative comparisons might also be beneficial in this respect, yet this requires further investigation.

Second, base rates of impairment may vary between patient samples. The current resampling procedures may allow for such base rate information in two ways. First, base rates may be included as priors in a Bayesian approach. Ibrahim, Chen, and Gray (2002) proposed a Bayesian extension of the one-step resampling approach in a group means testing context. An extension to the current normative comparisons context might therefore be feasible. Note, however, that a Bayesian approach is hardly ever used in neuropsychological practice (Elwood, 2007; Gavett, 2015). Instead, neuropsychologists include base rates informally by using lenient cutoff criteria—for example, by choosing a nominal alpha of 20% instead of 5% (Elwood, 2007; Meehl & Rosen, 1955). Accordingly, the second way to incorporate base rate information into resampling procedures is to use such lenient cutoff criteria.

Third, and related to the previous point, as compared to the usual uncorrected comparisons, Bonferroni, Holm, one-step resampling, and step-down resampling comparisons are characterized by decreased sensitivity. If high sensitivity is required, we argue that it is better not to use uncorrected comparisons, as this will not provide insight into the actual familywise error. In such circumstances, we suggest using an elevated criterion—for example, to change the required familywise error from 5 to 10 or 20%. For example, if an effective and safe treatment for cognitive impairment would be available, up to 20% false positives might be preferred to minimize the risk that patients are denied access to this effective treatment.

To conclude, the present study indicates that large battery test-specific normative comparisons are best carried out by resampling normative comparisons, especially by step-down resampling comparisons. They control familywise error, they have the highest sensitivity to detect genuine deviations, and they are user-friendly, since the Normative Comparisons website promotes their routine use in neuropsychological research and clinical practice.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

H.M.H. is supported by a VICI grant awarded by the Netherlands Organization of Scientific Research (NWO) [grant number 453-12-005]. J.A.R. is supported by a grant awarded by the NWO [grant number MaGW 480-12-015]. D.M. was supported by the Prinses Beatrix Fonds [grant number PGO01-0138]. R.P.P.G. was supported by a VENI grant awarded by the NWO [grant number C.2523.0079].

References

- Advanced Neuropsychological Diagnostics Infrastructure. (2016, January 21). Retrieved from <http://www.andi.nl/home/>
- Agelink van Rentergem, J. A., & Huizenga, H. M. (2016, January 21). Retrieved from <https://eclip.shinyapps.io/NormativeComparisons>
- Anderson, M. J., & Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation*, 62(3), 271–303. doi:10.1080/00949659908811936
- Arenas-Pinto, A., Winston, A., Stöhr, W., Day, J., Wiggins, R., Quah, S. P., ... Paton, N. I. (2014). Neurocognitive function in HIV-infected patients: Comparison of two methods to define impairment. *PLoS One*, 9(7), e103498. doi:10.1371/journal.pone.0103498
- Axelrod, B. N., & Wall, J. R. (2007). Expectancy of impaired neuropsychological test scores in a non-clinical sample. *The International Journal of Neuroscience*, 117(11), 1591–1602. doi:10.1080/00207450600941189
- Bell, M. L., Olivier, J., & King, M. T. (2013). Scientific rigour in psycho-oncology trials: Why and how to avoid common statistical errors. *Psycho-Oncology*, 505(22), 499–505.
- Berthelson, L., Mulchan, S. S., Odland, A. P., Miller, L. J., & Mittenberg, W. (2013). False positive diagnosis of malingering due to the use of multiple effort tests. *Brain Injury*, 27(7–8), 909–916. doi:10.3109/02699052.2013.793400
- BIG register. (2016, March 16). Retrieved from <https://www.bigregister.nl/en/>
- Bilder, R. M., Sugar, C. A., & Helleman, G. S. (2014). Cumulative false positive rates given multiple performance validity tests: Commentary on Davis and Millis (2014) and Larrabee (2014). *The Clinical Neuropsychologist*, 28(8), 1212–1223. doi:10.1080/13854046.2014.969774

- Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "Abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 24(1), 31–46. doi:10.1093/arclin/acn001
- Blackford, R. C., & La Rue, A. (1989). Criteria for diagnosing age-associated memory impairment: Proposed improvements from the field. *Developmental Neuropsychology*, 5(4), 295–306. doi:10.1080/87565648909540440
- Blakesley, R. E., Mazumdar, S., Dew, M. A., Houck, P. R., Tang, G., Reynolds, C. F., & Butters, M. A. (2009). Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology*, 23(2), 255–264. doi:10.1037/a0012850
- Brooks, B. (2010). Seeing the forest for the trees: Prevalence of low scores on the Wechsler Intelligence Scale for Children, fourth edition (WISC-IV). *Psychological Assessment*, 22(3), 650–656. doi:10.1037/a0019781
- Brooks, B. L., Iverson, G. L., Feldman, H. H., & Holdnack, J. A. (2009). Minimizing misdiagnosis: Psychometric criteria for possible or probable memory impairment. *Dementia and Geriatric Cognitive Disorders*, 27(5), 439–450. doi:10.1159/000215390
- Brooks, B. L., Iverson, G. L., Holdnack, J. A., & Feldman, H. H. (2008). Potential for misclassification of mild cognitive impairment: A study of memory scores on the Wechsler Memory Scale-III in healthy older adults. *Journal of the International Neuropsychological Society: JINS*, 14(3), 463–478. doi:10.1017/S1355617708080521
- Cohen, S., Ter Stege, J. A., Geurtsen, G. J., Scherpbier, H. J., Kuijpers, T. W., Reiss, P., ... Pajkrt, D. (2015). Poorer cognitive performance in perinatally HIV-infected children versus healthy socioeconomically matched controls. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 1(60), 1111–1119. doi:10.1093/cid/ciu1144
- Constantinidou, F., Wertheimer, J. C., Tsanadis, J., Evans, C., & Paul, D. R. (2012). Assessment of executive functioning in brain injury: Collaboration between speech-language pathology and neuropsychology for an integrative neuropsychological perspective. *Brain Injury*, 26(13–14), 1549–1563. doi:10.3109/02699052.2012.698786
- Crawford, J., & Allan, K. (1994). The Mahalanobis Distance index of WAIS-R subtest scatter: Psychometric properties in a healthy UK sample. *British Journal of Clinical Psychology*, 33, 65–69. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8260.1994.tb01094.x/full>
- Crawford, J. R. (2016, January 21). *Estimating Percentage of Population Exhibiting Abnormally Low Scores and Score Differences*. Retrieved from <http://homepages.abdn.ac.uk/j.crawford/pages/dept/PercentAbnormKtests.htm>
- Crawford, J. R., & Garthwaite, P. H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. *Neuropsychologia*, 40(8), 1196–1208. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11931923>
- Crawford, J. R., & Garthwaite, P. H. (2006). Comparing patients' predicted test scores from a regression equation with their obtained scores: A significance test and point estimate of abnormality with accompanying confidence limits. *Neuropsychology*, 20(3), 259–271. doi:10.1037/0894-4105.20.3.259
- Crawford, J. R., Garthwaite, P. H., Azzalini, A., Howell, D. C., & Laws, K. R. (2006). Testing for a deficit in single-case studies: Effects of departures from normality. *Neuropsychologia*, 44(4), 666–677. doi:10.1016/j.neuropsychologia.2005.06.001
- Crawford, J. R., Garthwaite, P. H., & Gault, C. B. (2007). Estimating the percentage of the population with abnormally low scores (or abnormally large score differences) on standardized neuropsychological test batteries: A generic method with applications. *Neuropsychology*, 21(4), 419–430. doi:10.1037/0894-4105.21.4.419
- Crawford, J. R., & Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology*, 20(5), 755–762. doi:10.1076/jcen.20.5.755.1132
- Crawford, J. R., Howell, D. C., & Garthwaite, P. H. (1998). Payne and Jones revisited: Estimating the abnormality of test score differences using a modified paired samples *t* test. *Journal of Clinical and Experimental Neuropsychology*, 20(6), 898–905. doi:10.1076/jcen.20.6.898.1112
- Davis, J. J., & Millis, S. R. (2014). Examination of performance validity test failure in relation to number of tests administered. *The Clinical Neuropsychologist*, 28(2), 199–214. doi:10.1080/13854046.2014.884633
- Eichstaedt, K. E., Kovatch, K., & Maroof, D. A. (2013). A less conservative method to adjust for familywise error rate in neuropsychological research: The Holm's sequential Bonferroni procedure. *NeuroRehabilitation*, 32(3), 693–696. doi:10.3233/NRE-130893
- Elwood, R. W. (2007). Clinical discriminations and neuropsychological tests: An appeal to Bayes' theorem. *Clinical Neuropsychologist*, 7(2), 224–233. doi:10.1080/13854049308401527
- Gavett, B. E. (2015). The value of Bayes' theorem for interpreting abnormal test scores in cognitively healthy and clinical samples. *Journal of the International Neuropsychological Society*, 21(3), 249–257. doi:10.1017/S1355617715000168
- Gisslén, M., Price, R. W., & Nilsson, S. (2011). The definition of HIV-associated neurocognitive disorders: Are we overestimating the real prevalence? *BMC Infectious Diseases*, 11(1), 356. doi:10.1186/1471-2334-11-356

- González-Redondo, R., Toledo, J., Clavero, P., Lamet, I., García-García, D., García-Eulate, R.,... Rodríguez-Oroz, M. C. (2012). The impact of silent vascular brain burden in cognitive impairment in Parkinson's disease. *European Journal of Neurology: The Official Journal of the European Federation of Neurological Societies*, 19(8), 1100–1107. doi:10.1111/j.1468-1331.2012.03682.x
- Grasman, R. P. P. P., Huizenga, H. M., & Geurts, H. M. (2010). Departure from normality in multivariate normative comparison: The Cramér alternative for Hotelling's T². *Neuropsychologia*, 48(5), 1510–1516. doi:10.1016/j.neuropsychologia.2009.11.016
- Grunseit, A. C., Perdices, M., Dunbar, N., & Cooper, D. A. (1994). Neuropsychological function in asymptomatic HIV-1 infection: Methodological issues. *Journal of Clinical and Experimental Neuropsychology*, 16(6), 898–910. doi:10.1080/01688639408402701
- Höfler, M. (2005). The effect of misclassification on the estimation of association: A review. *International Journal of Methods in Psychiatric Research*, 14(2), 92–101. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/mpr.20/abstract>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. doi:10.2307/4615733
- Huba, G. J. (1985). How unusual is a profile of test scores? *Journal of Psychoeducational Assessment*, 3(4), 321–325.
- Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin*, 105(2), 302–308. doi:10.1037/0033-2909.105.2.302
- Huizenga, H. M., Smeding, H., Grasman, R. P. P. P., & Schmand, B. (2007). Multivariate normative comparisons. *Neuropsychologia*, 45(11), 2534–2542. doi:10.1016/j.neuropsychologia.2007.03.011
- Ibrahim, J. G., Chen, M.-H., & Gray, R. J. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, 97(457), 88–99. doi:10.1198/016214502753479257
- Ingraham, L. J., & Aiken, C. B. (1996). An empirical approach to determining criteria for abnormality in test batteries with multiple measures. *Neuropsychology*, 10(1), 120–124. doi:10.1037//0894-4105.10.1.120
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. doi:10.1371/journal.pmed.0020124
- Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *The American Psychologist*, 63(3), 146–159. doi:10.1037/0003-066X.63.3.146
- Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, 59(11), 990–996. doi:10.1016/j.biopsych.2005.09.014
- Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologist*, 22(4), 666–679. doi:10.1080/13854040701494987
- Larrabee, G. J. (2014). False-positive rates associated with the use of multiple performance and symptom validity tests. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 29(4), 364–373. doi:10.1093/arclin/acu019
- Levav, M., Mirsky, A. F., Herault, J., Xiong, L., Amir, N., & Andermann, E. (2002). Familial association of neuropsychological traits in patients with generalized and partial seizure disorders. *Journal of Clinical and Experimental Neuropsychology*, 24(3), 311–326. doi:10.1076/jcen.24.3.311.985
- Lewis, M. S., Maruff, P., Silbert, B. S., Evered, L. A., & Scott, D. A. (2006). Detection of postoperative cognitive decline after coronary artery bypass graft surgery is affected by the number of neuropsychological tests in the assessment battery. *The Annals of Thoracic Surgery*, 81(6), 2097–2104. doi:10.1016/j.athoracsur.2006.01.044
- Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). New York, NY: Oxford University Press.
- Loewenstein, D., Acevedo, A., Ownby, R., Agron, J., Barker, W., Isaacson, R.,... Duara, R. (2006). Using different memory cutoffs to assess mild cognitive impairment. *The American Journal of Geriatric Psychiatry*, 14(11), 911–919. Retrieved from <http://www.sciencedirect.com/science/article/pii/S1064748112608707>
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52(3), 194–216. doi:10.1037/h0048070
- Meyer, A.-C. L., Boscardin, W. J., Kwasa, J. K., & Price, R. W. (2013). Is it time to rethink how neuropsychological tests are used to diagnose mild forms of HIV-associated neurocognitive disorders? Impact of false-positive rates on prevalence and power. *Neuroepidemiology*, 41(3–4), 208–216. doi:10.1159/000354629
- Meyers, J. E., Miller, R. M., Thompson, L. M., Scalese, A. M., Allred, B. C., Rupp, Z. W.,... Junghyun Lee, A. (2014). Using likelihood ratios to detect invalid performance with performance validity measures. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 29(3), 224–235. doi:10.1093/arclin/acu001
- Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K., Gerber, A.,... van der Laan, M. (2014). Promoting transparency in social science research. *Science*, 343, 30–31. Retrieved from http://e-gap.org/wp/wp-content/uploads/2014/04/Transparency-UCB_2014-04-09.pdf
- Multivariate normative comparisons. (2016, January 21). Retrieved from <http://purl.org/net/rgrasman/mnc>
- Muslimovic, D., Post, B., Speelman, J. D., & Schmand, B. (2005). Cognitive profile of patients with newly

- diagnosed Parkinson disease. *Neurology*, 65(8), 1239–1245. doi:10.1212/01.wnl.0000180516.69442.95
- Naglieri, J. A., & Paolitto, A. W. (2005). Ipsative comparisons of WISC-IV index scores. *Applied Neuropsychology*, 12(4), 208–211. doi:10.1207/s15324826an1204
- Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1), 1–25. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11747097>
- Palmer, B. W., Boone, K. B., Lesser, I. M., & Wohl, M. A. (1998). Base rates of “impaired” neuropsychological test performance among healthy older adults. *Archives of Clinical Neuropsychology*, 13(6), 503–511. doi:10.1093/arclin/13.6.503
- Proto, D. A., Pastorek, N. J., Miller, B. I., Romesser, J. M., Sim, A. H., & Linck, J. F. (2014). The dangers of failing one or more performance validity tests in individuals claiming mild traumatic brain injury-related postconcussive symptoms. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 29(7), 614–624. doi:10.1093/arclin/acu044
- R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Sander, A., Nakase-Richardson, R., Constantinidou, F., Wertheimer, J., & Paul, D. R. (2007). Memory assessment on an interdisciplinary rehabilitation team: A theoretically based framework. *American Journal of Speech-Language Pathology*, 16, 316–331. Retrieved from <http://ajslp.pubs.asha.org/article.aspx?articleid=1757586>
- Schagen, S. B., Muller, M. J., Boogerd, W., Mellenbergh, G. J., & van Dam, F. S. A. M. (2006). Change in cognitive function after chemotherapy: A prospective longitudinal study in breast cancer patients. *Journal of the National Cancer Institute*, 98(23), 1742–1745. doi:10.1093/jnci/djj470
- Schatz, P., Jay, K. A., McComb, J., & McLaughlin, J. R. (2005). Misuse of statistical tests in Archives of Clinical Neuropsychology publications. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, 20(8), 1053–1059. doi:10.1016/j.acn.2005.06.006
- Schretlen, D. J., Testa, S. M., Winicki, J. M., Pearlson, G. D., & Gordon, B. (2008). Frequency and bases of abnormal performance by healthy adults on neuropsychological testing. *Journal of the International Neuropsychological Society: JINS*, 14(3), 436–445. doi:10.1017/S155617708080387
- 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(11), 1359–1366. doi:10.1177/0956797611417632
- Smeding, H. M. M., Speelman, J. D., Huizenga, H. M., Schuurman, P. R., & Schmand, B. (2011). Predictors of cognitive and psychosocial outcome after STN DBS in Parkinson’s disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, 82(7), 754–760. doi:10.1136/jnnp.2007.140012
- Sokal, R. R., & Rohlf, J. F. (1995). *Biometry*. San Francisco, CA: W. H. Freeman.
- Su, T., Schouten, J., Geurtsen, G. J., Wit, F. W., Stolte, I. G., Prins, M., ... Schmand, B. A. (2015). Multivariate normative comparison, a novel method for more reliably detecting cognitive impairment in HIV infection. *AIDS*, 29(5), 547–557. doi:10.1097/QAD.0000000000000573
- Torti, C., Focà, E., Cesana, B. M., & Lescure, F. X. (2011). Asymptomatic neurocognitive disorders in patients infected by HIV: Fact or fiction? *BMC Medicine*, 9(1), 138. doi:10.1186/1741-7015-9-138
- Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Vol. 279). Hoboken, NJ: John Wiley & Sons.
- Wilson, C. E., Happé, F., Wheelwright, S. J., Ecker, C., Lombardo, M. V., Johnston, P., ... Murphy, D. G. M. (2014). The neuropsychology of male adults with high-functioning autism or Asperger syndrome. *Autism Research: Official Journal of the International Society for Autism Research*, 7(5), 568–581. doi:10.1002/aur.1394
- Wilson, S. J., Baxendale, S., Barr, W., Hamed, S., Langfitt, J., Samson, S., ... Smith, M.-L. (2015). Indications and expectations for neuropsychological assessment in routine epilepsy care: Report of the ILAE Neuropsychology Task Force, Diagnostic Methods Commission, 2013–2017. *Epilepsia*, 56(5), 674–681. doi:10.1111/epi.12962

Appendix

Algorithms

In this appendix we first give a more detailed description of the one-step resampling algorithm and then indicate how it is extended to the step-down resampling algorithm. R code is also provided.

One-step resampling algorithm

The one-step algorithm yields the distribution of the maximum absolute t_{test} statistic under the null hypothesis, the max distribution, by means of the so-called permutation approach to resampling (cf. Anderson & Legendre, 1999; Nichols & Holmes, 2002). Thereafter, each absolute t_{norm} statistic of an individual is compared to this max distribution.

Let $n = 1, \dots, N$ denote N participants in the normative sample, and let $m = 1, \dots, M$ denote M neuropsychological tests. A vector \mathbf{y}_m^* contains N centered test scores on the m th neuropsychological test. For example, if $N = 6$, \mathbf{y}_m^* may equal $[2, 4, -2, -4, 2, -2]$ —that is, on test m , the first participant in the normative sample has a centered score of 2, the second participant has a centered score of 4, and so on. The scalar x_m^* denotes an individual’s centered score on the m th test. Then perform the following computations (C1–C5).

- C1: As normative data have been centered, a resample can be obtained by multiplying each element in \mathbf{y}_m^* by a randomly chosen +1 or a -1 (cf. Nichols & Holmes, 2002). For example, the centered normative data on the m th test $\mathbf{y}_m^* = [2, 4, -2, -4, 2, -2]$ are multiplied by a randomly generated vector $[+1, -1, +1, +1, -1, +1]$ yielding the resampling data $\mathbf{y}_m' = [2, -4, -2, -4, -2, -2]$. In order to leave the correlation structure between variables intact, it is crucial that each test is multiplied by the same randomly generated vector, so $\mathbf{y}_1^*, \mathbf{y}_1', \dots, \mathbf{y}_M^*$ are all multiplied by $[+1, -1, +1, +1, -1, +1]$.
- C2: Compute the t_{test} statistic in this resampling dataset for each of the $m = 1, \dots, M$ tests. In computing this statistic, x_m^* is set to zero, as we are interested in the distribution under the null hypothesis.
- C3: Determine the maximum over the M absolute t_{test} statistics obtained in Step 2. This yields the max statistic. Repeat C1 to C3 several, say L , times. In the present simulation study, L is set to 2000.
- C4: Store the L max statistics; this yields the required distribution of the maximum absolute t_{test} statistic under the null hypothesis, the max distribution.
- C5: Determine, for each variable m , where the absolute t_{norm} statistic is located in the max distribution. In the case of a two-sided hypothesis, if an absolute t_{norm} statistic is located beyond the 95th percentile of the max distribution, this indicates that the individual deviates from the norm on that particular neuropsychological test. In case of a one-sided hypothesis—that is, that an individual performs worse than

the norm—two conditions should be satisfied: The sign of t_{norm} should be in the expected direction, and the absolute value of t_{norm} should be located beyond the 90th percentile of the max distribution. Note that the two- and one-sided critical values are 90% and 95% and not 95% and 97.5 since the max distribution concerns maxima of absolute t -values.

Step-down resampling algorithm

Absolute t_{norm} statistics are first ordered from high to low. The highest absolute t_{norm} statistic is referred to the tmax distribution, as outlined above. The next highest t_{norm} statistic is referred to the tmax distribution derived over all neuropsychological tests, except the test for which the highest t_{norm} statistic was observed. The second next highest t_{norm} statistic is referred to the tmax distribution derived over all neuropsychological tests, except the two tests for which the two highest t_{norm} statistics were observed, and so on. The p -values thus obtained are subjected to a correction, imposing that p -values of low absolute t_{norm} statistics can never be lower than p -values of high absolute t_{norm} statistics. That is, a p -value is the maximum of the current p -value and the corrected p -values associated with higher absolute t_{norm} statistics.

To view a color version of the R-code algorithm, please see the online issue of the Journal.



```

1 #Large battery normative comparisons
2 #Tests one sided negative effect: uncorrected, Bonferroni, Holm, one-step resampling, step-down resampling
3
4 rm(list=ls(all=TRUE))
5 library(matrixStats)
6
7 library(MASS)
8
9 ##START INPUT##
10 nsamp=2000
11 alpha=0.05
12 #number of resamples
13 #nominal alpha
14 #should contain id, group
15 #identifier (1 for patient group, patient group goes first) and nvar tests
16 ##END INPUT
17
18 npat=sum(dats[,2])
19 nnorm=dim(dats)[1]-npat
20 nvar=dim(dats)[2]-2
21 #number of patients
22 #number of controls
23 #number of tests
24 dats=as.matrix(dats)
25
26 #statistics on normative sample
27 normdat=dats[((npat+1):(npat+nnorm)),(3:(nvar+2))]
28 normmean=colMeans(normdat)
29 #norm mean
30 normdat=normdat-c(rep(1,nnorm))%*%t(normmean)
31 #center
32 normse =colSds(normdat)
33 #se
34 normse =normse/sqrt(nnorm)
35
36 resmat=c(); resmat_overall=c()
37 for (k in 1:npat) {
38   pval=c(); pval_holm=c(); pval_bonf_fin=c(); pval_holm_fin=c(); pval_res1_fin=c();
39   pval_res2_fin=c()
40   hlp=c(rep(0,nvar*5))
41
42   #statistics on individual patient

```

```

1 outvec=dat[k,1]
2 indtat=dat[k,(3:(nvar+2))]
3 indtat=indtat-normmean
4 normstat=(indtat/normse)/sqrt(nnorm+1)
5 ord=order(-abs(normstat))
6
7 ##UNCORRECTED, BONFERRONI, HOLM
8 for (i in 1:nvar) {
9   j=ord[i]
10  aa=1-pt(abs(normstat[j]),(nnorm-1))
11  bb=aa*(nvar+1-i)
12  pval=c(pval,aa)
13  pval_holm=c(pval_holm,bb)
14 }
15
16 #enforce monotonicity on Holm p-values
17 pval_holmh=pval_holm
18 for (i in 1:nvar) {
19   if (i==1) pval_holm[i]=pval_holmh[i]
20   if (i> 1) pval_holm[i]=max(pval_holmh[i],pval_holm[i-1])
21 }
22
23 for (i in 1:nvar) {
24   j=which(ord==i)
25   pval_fin= c(pval_fin,pval[j])
26   pval_bonf_fin=c(pval_bonf_fin,pval_fin[i]*nvar)
27   pval_holm_fin=c(pval_holm_fin,pval_holm[j])
28   if(((indtat[i])<0) & (pval_fin[i]) <alpha)
29   if(((indtat[i])<0) & (pval_bonf_fin[i])<alpha)
30   if(((indtat[i])<0) & (pval_holm_fin[i])<alpha)
31 }
32
33 fwe_u=sum(hlp[(1:nvar)])
34 fwe_b=sum(hlp[((nvar+1):(2*nvar))])

```

#patient identifier
#data
#center
#normmean is zero due to centering
#first: index normstat highest absolute t value
#uncorrected
#Bonferroni
#Holm



```

1 fwe_h=sum(hlp[((2*nvar+1):(3*nvar))])
2 if(fwe_u != 0) outvec=c(outvec,1) else outvec=c(outvec,0)
3 if(fwe_b != 0) outvec=c(outvec,1) else outvec=c(outvec,0)
4 if(fwe_h != 0) outvec=c(outvec,1) else outvec=c(outvec,0)
5
6
7 ##ONE-STEP RESAMPLING AND STEP-DOWN RESAMPLING
8 tboot=c()
9 tmaxboot=c()
10 for (j in 1:nsamp) {
11   rand_ones=rbinom(nnorm,1,0.5)
12   for (i in 1:(nnorm)) {if(rand_ones[i]==0.0) rand_ones[i]==-1}
13   bootsampmat=matrix(rep(rand_ones,nvar),nnorm,nvar)
14   bootsampmat=bootsampmat*normdat
15   tstats=as.double(colMeans(bootsampmat)/(colSds(bootsampmat)/sqrt(nnorm)))
16   tmaxboot=c(tmaxboot,max(abs(tstats)))
17   tboot=rbind(tboot,abs(tstats))
18 }
19
20 #one step
21 pval_res1_fin=c()
22 for (i in 1:nvar) {
23   pval_res1_fin =c(pval_res1_fin,length(which(as.double(tmaxboot) >= abs(normstat[i])))/nsamp)
24   if((inddat[i]<0 & pval_res1_fin[i]<(alpha*2)) hlp[(3*nvar+i)]=1
25 }
26
27 #step-down
28 tststepboot=c(rep(0, nsamp))
29 psmaxh=c()
30 for (i in nvar:1) {
31   j=ord[i]
32   if (i==nvar) tststepboot=tboot[,j]
33   if (i<nvar) tststepboot=apply(cbind(tboot[,j],tststepboot),1,max)

```

#familywise uncorrected
#familywise Bonf corrected
#familywise Holm corrected

```

1  psmaxh =c(psmaxh,(length(which(as.double(tstepboot) >= abs(normstat[j])))/nsamp))
2  }
3
4  psmax=psmaxh
5  for (i in nvar:1) {
6    if (i==nvar)psmax[i]=psmaxh[i]
7    if (i<nvar) psmax[i]=max(psmaxh[i],psmax[i+1])      #enforce monotonicity
8  }
9
10 #reorder
11 for (i in 1:nvar) {
12   j=nvar+1-which(ord==i)
13   pval_res2_fin=c(pval_res2_fin,psmax[j])
14   if((inddat[i]<0 & pval_res2_fin[i]<(alpha*2)) hlp[(4*nvar+i)]=1
15 }
16
17 fwe_r1=sum(hlp[((3*nvar+1):(4*nvar))])
18 if(fwe_r1 !=0) outvec=c(outvec,1) else outvec=c(outvec,0)
19 fwe_r=sum(hlp[((4*nvar+1):(5*nvar))])
20 if(fwe_r1 !=0) outvec=c(outvec,1) else outvec=c(outvec,0)
21
22 blup=9
23 resmat=rbind(resmat,
24   c(outvec[1],"", "blup", round(normstat, 3)),
25   c(outvec[1],"uncor", outvec[2], round(pval_fin, 3)),
26   c(outvec[1],"Bonf ", outvec[3], round(pval_bonf_fin,3)),
27   c(outvec[1],"Holm", outvec[4], round(pval_holm_fin,3)),
28   c(outvec[1],"Res1", outvec[5], round(pval_res1_fin,3)),
29   c(outvec[1],"Res2", outvec[6], round(pval_res2_fin,3)))
30
31 resmat_overall=rbind(resmat_overall,outvec)
32 }
33 resmat

1 colMeans(resmat_overall)
2 save.image("d.Rdata")

```