The Spanish Journal of Psychology
2011, Vol. 14, No. 2, On Line First
http://dx.doi.org/10.5209/rev_SJOP.2011.v14.n2.48

ISSN 1138-7416

Testing Equivalence with Repeated Measures: Tests of the Difference Model of Two-Alternative Forced-Choice Performance

Miguel A. García-Pérez and Rocío Alcalá-Quintana

Universidad Complutense (Spain)

Solving theoretical or empirical issues sometimes involves establishing the equality of two variables with repeated measures. This defies the logic of null hypothesis significance testing, which aims at assessing evidence against the null hypothesis of equality, not for it. In some contexts, equivalence is assessed through regression analysis by testing for zero intercept and unit slope (or simply for unit slope in case that regression is forced through the origin). This paper shows that this approach renders highly inflated Type I error rates under the most common sampling models implied in studies of equivalence. We propose an alternative approach based on omnibus tests of equality of means and variances and in subject-by-subject analyses (where applicable), and we show that these tests have adequate Type I error rates and power. The approach is illustrated with a re-analysis of published data from a signal detection theory experiment with which several hypotheses of equivalence had been tested using only regression analysis. Some further errors and inadequacies of the original analyses are described, and further scrutiny of the data contradict the conclusions raised through inadequate application of regression analyses.
Keywords: statistical equivalence, repeated measures, Signal Detection Theory, Yes–No, 2AFC, interval bias, Standard Difference Model.

Resolver problemas teóricos o empíricos requiere en ocasiones contrastar la equivalencia de dos variables usando medidas repetidas. El mero planteamiento de este objetivo supone un desafío para la lógica subyacente a los métodos de contraste de hipótesis estadísticas, que están diseñados para evaluar la magnitud de la evidencia contraria a la hipótesis nula y de ningún modo permiten evaluar la evidencia a favor de ella. En algunos contextos aplicados se ha abordado el problema utilizando métodos de regresión y contrastando la hipótesis de que la pendiente es 1 y la hipótesis de que la ordenada en el origen es 0 (o simplemente la primera de ellas cuando se fuerza la regresión "por el origen"). Este trabajo muestra que esa estrategia conlleva tasas empíricas de error tipo I muy superiores a las tasas nominales bajo cualquiera de los modelos de muestreo más comúnmente implicados en estudios de equivalencia. Como alternativa, se propone una estrategia basada tanto en pruebas tipo ómnibus que incluyen contrastes de medias y varianzas como en análisis sujeto a sujeto (cuando la situación lo permita). Un estudio de simulación con estas pruebas muestra que la tasa empírica de error tipo I se ajusta a la tasa nominal y que la potencia de los contrastes es adecuada. A modo de ilustración, se aplican estos contrastes para re-analizar los datos de un experimento psicofísico sobre detección de contraste que originalmente sólo fueron analizados mediante regresión por parte de los autores del estudio, pese a que todas las hipótesis consideradas implicaban equivalencia con medidas repetidas. Nuestro re-análisis permite una inspección más minuciosa de los datos que revela contradicciones entre las características empíricas de los datos y las conclusiones extraídas mediante la aplicación inadecuada de métodos de regresión. Los resultados de este re-análisis también invalidan las conclusiones extraídas en la publicación original.
Palabras clave: equivalencia estadística, medidas repetidas, Teoría de Detección de Señales, Sí–No, elección forzada entre dos alternativas, efectos de orden.

Supported by grant PSI2009-08800 from Ministerio de Ciencia e Innovación (Spain). We thank Marisa Carrasco for sharing the data from their study.

Correspondence concerning this article should be addressed to Miguel A. García-Pérez. Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas. 28223 Madrid (Spain). Phone: +34-913943061. Fax: +34-913943189. E-mail: miguel@psi.ucm.es

Over the years of reviewing manuscripts, I had been developing the nervous conviction that there was too much emphasis on pouring raw data into SPSS or BMD programs and simply accepting whatever numbers emerged—chi-squares, F ratios, whatever—as the conclusion, without further ado. This didn't seem to me like a very imaginative or fruitful way to go about analyzing data. It seemed to me that off-the-shelf statistical analysis programs were producing off-the-shelf results.

Loftus (1985, p. 149)

The reason that, say, Cronbach's alpha and principal components analysis are so popular in psychology (...) is that they are default options in certain mouse-click sequences of certain popular statistics programs. Since psychologists are monogamous in their use of such software (most in my department are wedded to SPSS) there is little chance of convincing them to use a model—any model—that is not "clickable" in the menus of major statistical programs.

Borsboom (2006, p. 433)

The need to establish statistical equivalence arises in a number of contexts, such as when seeking evidence that two or more groups are matched with respect to some control variable (a methodological question), when seeking evidence that two or more treatments are equally effective (a practical question), or when seeking evidence that two or more experimental manipulations produce the indistinguishable effects that some model predicts (a theoretical question). In all of these occasions the experimental hypothesis (i.e., that groups are matched, that treatments are equally effective, or that manipulations produce the same effect) translates into not rejecting the null hypothesis, and the researcher actually seeks to find evidence supporting the experimental hypothesis. As is well known, null hypothesis significance testing (NHST) is not useful when the aim of the experimenter is, loosely speaking, to accept the null hypothesis of no difference. As Blackwelder (1982, p. 346) put it, "*p* is a measure of the evidence against the null hypothesis, not for it, and insufficient evidence to reject the null hypothesis does not imply sufficient evidence to accept it." This problem arises mostly because the logic of NHST rarely fits researchers' goals (Dixon & O'Reilly, 1999). Tests of equivalence are needed in these cases, and some have been developed over the past few decades for use with independent groups (Anderson & Hauck, 1983; Blackwelder, 1982; Dunnett & Gent, 1977; Edgell, 1995; Frick, 1995a, 1995b; Kirkwood, 1981; Metzler, 1974; Rogers, Howard, & Vessey, 1993; Selwyn, Demptster, & Hall, 1981; Selwyn & Hall, 1984; Stegner, Bostrom, & Greenfield, 1996; Tryon, 2001; Tryon & Lewis, 2008; Westlake, 1976, 1979, 1981).

The tests described in the papers just mentioned represent slight variants of conventional methods in NHST, like using strategies that increase statistical power (as part of the broader *good-effort criterion* of reasonably seeking to detect an effect if it existed; Frick, 1995a, 1995b), defining specific types of confidence intervals, or defining indifference regions around the null hypothesis (for an application of this latter idea in the context of mastery decisions in educational measurement, see van den Brink & Koele, 1980; García-Pérez, 1989). All of these methods require users to specify the maximum amount of difference that is regarded as negligible in practice, and the methods themselves are generally aimed at testing equality of means (for testing for lack of association, see Goertzen & Cribbie, 2010). These two characteristics represent drawbacks for the type of application that we will be discussing in this paper, both because the maximum difference that is negligible is difficult to agree upon in many cases and also because we will be considering instances in which equivalence implies not only identity of means but also of variances at the very least. More importantly, all the methods described in the papers listed above were developed in the context of establishing the equivalence of independent groups of observations, but this case does not exhaust all of the situations in which a researcher might be interested in testing equivalence. Actually, experiments in which participants serve under several treatments lend themselves to equivalence tests with repeated measures. For example, one may set out to determine whether sensory thresholds obtained with alternative psychophysical methods are equivalent (Alcalá-Quintana & García-Pérez, 2007), whether paper-and-pencil and computer administrations of a test are equivalent (Hays & McCallum, 2005), or whether cross-modal interactions affecting saccade latencies are invariant under certain types of experimental manipulations (Diederich & Colonius, 2011). Also, and to anticipate the experimental context in which equivalence tests with repeated measures will be illustrated in this paper, one might set out to test the signal detection theory tenet that performance in two-alternative-forced choice (2AFC) trials is the same whether the signal is presented in the first or in the second interval. In this latter case, repeated measures are not an option but an inescapable consequence of experimental control (i.e., all participants must serve in an experiment in which the signal is randomly presented in the first or the second interval across a series of 2AFC trials).

Equivalence tests with repeated measures do not seem to have been developed as such. Yet, a large body of literature under the general labels of "determining agreement between instruments" or "method comparison studies" describes tests that are applicable in these conditions (see, e.g., Altman & Bland, 1983; Astrua, Ichim, Pennecchi, & Pisani, 2007; Bland & Altman, 1986, 1999; Cox, 2006; Dunn & Roberts, 1999; Hawkins, 2002; Lin, 1989, 1992, 2000; Lin, Hedayat, Sinha, & Yang, 2002; van Stralen, Jager, Zoccali, & Dekker, 2008; Wang & Iyer, 2008; Westgard & Hunt, 1973). The context in which these tests were originally developed involved a comparison of two alternative

instruments (or methods) measuring the same variable and the goal was to assess the agreement (or equivalence) between the measures provided by either instrument or method. This literature has almost exclusively considered correlation and linear regression as the tools to carry out the equivalence tests, but there are obvious ways in which correlation is inadequate and linear regression is insufficient for this purpose (for a thorough discussion of the theoretical inadequacy of regression analyses in the context of method comparison studies, see Bland & Altman, 1986, 2003; Lin, 1989, 1992, 2000). The equivalence tests that are generally carried out for this purpose consist of testing the null hypotheses that the regression slope is unity and that the regression intercept is zero; in some occasions, linear regression through the origin (Turner, 1960) has also been used, which forces the intercept to zero and thus only involves a test that the regression slope is unity.[1]

The goal of this paper is three-fold. First, to provide evidence that linear regression (whether unconstrained or through the origin) is not advisable because the Type I error rates of the tests for regression slope and intercept overwhelmingly exceed their nominal rates under the most common circumstances in equivalence testing. Thus, this demonstration does not rely on rhetorical arguments regarding whether regression seems reasonable for the purpose but, rather, in actual evidence regarding its failure as a statistical test in this context. Second, to propose an alternative package of statistical tests for testing equivalence with repeated measures and to show that the Type I error rate and power of these tests is adequate. Third, to re-analyze a published data set in which equivalence was tested using the inadequate method of regression through the origin. Our re-analyses do not support the conclusions that were originally raised. Our overall purpose is to describe and illustrate the use of an adequate statistical package to substantiate a decision about rejection or not rejection of equivalence. Nevertheless, this decision will never turn into strictly accepting equivalence although the package indeed provides the occasion for a lack of equivalence to manifest in one way or another, thus abiding by the subjective good-effort criterion that is usual in tests of equivalence with independent measures (see Frick, 1995a, 1995b; Tryon, 2001) and that is widely adopted in empirical practice (Baguley, Landsdale, Lines, & Parkin, 2006; van Berkum, 1997; Corina, 1999; Cusack & Carlyon, 2003; Dierdorff & Morgeson, 2007; Ferrand, 1999; Hietanen & Leppänen, 2003; Hollands & Spence, 1998; Huntsman, 1998; Jordan & Troth, 2004; Kane, Poole, Tuholski, & Engle, 2006; Los, 2004; Perea & Rosa, 2002; Rorden, Karnath, & Driver, 2001; Russo, Fox, & Bowles, 1999; Saint-Aubin & Poirier, 1999; Segrin, 2004; Smith & Kounios, 1996; Spence & Driver, 1997, 1998; Tipples & Sharma, 2000; Vatakis & Spence, 2008; Vatakis, Ghazanfar, & Spence, 2008; Zampini et al., 2005).

The inadequacy of regression for testing equivalence with repeated measures

Consider the case of two variables, X and Y, measured in the same sample of individuals. Observed measurements in X and Y are both affected by error, and errors may not have the same statistical characteristics for X and Y. For instance, the two variables could be measured with procedures that differ as to bias and precision. Thus, observed values in X and Y will differ from the underlying true values by a random additive error whose mean reflects measurement bias and whose variance reflects measurement precision, and the mean and variance of these errors may vary across variables. We will adopt the usual notation where n is the size of the sample of paired observations, $\bar{X}$ and $\bar{Y}$ are the sample means, s_x^2 and s_y^2 are the sample variances, and r_{xy} is the sample product-moment correlation.

Under the conventional *measurement model*, if X and Y are observed measures of the same latent variable T, measurement error makes $X = T + \varepsilon_X$ and $Y = T + \varepsilon_Y$, where ε_X and ε_Y are random variables with means μ_{ε_X} and μ_{ε_Y} (reflecting bias when these means are different from zero) and non-null variances $\sigma^2_{\varepsilon_X}$ and $\sigma^2_{\varepsilon_Y}$ (reflecting precision). Differences in the bias and precision with which the two variables are measured is the only threat to the statistical equivalence of X and Y in this case. This is of outmost interest in method comparison studies and in assessing agreement between instruments because in these cases a common latent variable is generally guaranteed. However, when the latent variables are not the same and the research question is purely theoretical (e.g., assessing whether the threshold for discriminating the length of temporal intervals is the same whether the interval is delimited by auditory or by visual stimuli), differences in the bias and precision with which the two variables are measured (which are theoretically irrelevant) may seriously hamper the quest for equivalence. In these cases X and Y are not guaranteed to share a latent variable so that measurement error makes $X = T_X + \varepsilon_X$ and $Y = T_Y + \varepsilon_Y$, where T_X and T_Y are the true values and ε_X and ε_Y are random errors as before. The relevant theoretical question is whether the latent variables T_X and T_Y are equivalent, but testing this equivalence is hampered by the fact that the manifest variables X and Y

[1] Whether the regression intercept should be forced to zero or estimated from the data is certainly a controversial issue in the literature, but we will not discuss it here. The interested reader can ponder the various aspects of this issue in Hahn (1977), Casella (1983), Mukherjee, White, and Wuyts (1998), Eisenhauer (2003), or Freund, Wilson, and Sa (2006).

which are used in the test are affected by measurement errors that may disguise the potential equivalence of the latent variables. Note, however, that whether or not $T_X = T_Y$, potential differences in the distributions of ε_X and ε_Y may result in a lack of equivalence of X and Y. Classical test theory (see Gulliksen, 1950) describes the relations between observed and latent variables under this measurement model.

But it is also possible that the two variables X and Y can only be described by their joint probability distribution. Then, the experiment that is used to collect the data with which to test equivalence would be regarded as involving *bivariate sampling* from a distribution in which variables X and Y have means μ_X and μ_Y, variances σ_x^2 and σ_y^2, and a correlation ρ_{XY}.

These two types of sampling differ from that assumed in regression models. Under *regression sampling*, the observations in X are taken at specified fixed levels (i.e., they are not random) and they are measured without error whereas the observations in Y are assumed to be normally distributed with mean $\beta_0 + \beta_1 X$ (in ordinary least-squares linear regression) or $\beta_1 X$ (in regression through the origin) and with a variance σ_e^2 that does not vary with X (the homoscedasticity assumption). It is important to realize that the conventional test statistics for β_0 and β_1 (or only for the latter in the case of regression through the origin) were derived under the assumption of regression sampling and, hence, that their performance under bivariate sampling or under the measurement model described above is not guaranteed to be accurate. To investigate their behavior in these three cases and also under a slight variant of regression sampling whereby the observations in X are random, a simulation study was carried out as described next.

Simulation method

Twenty-thousand samples were drawn according to each of the four sampling models described above. Under the measurement model, T was normally distributed with mean and variance that varied across simulations whereas ε_X and ε_Y were independent and identically normally distributed with mean 0 and a variance that also varied across simulations with the constraint that $\sigma_x^2 = \sigma_y^2 = \sigma_T^2 + \sigma_{\varepsilon_X}^2 = \sigma_T^2 + \sigma_{\varepsilon_Y}^2$. Under bivariate sampling, X and Y had a bivariate normal distribution with means $\mu_X = \mu_Y$ and variances $\sigma_x^2 = \sigma_y^2$ that varied across simulations whereas ρ_{XY} also varied between .69 and .99 across simulations. Under regression sampling, the number of levels for X varied between 3 and 23 across simulations and these levels were symmetrically placed around μ_X with constant spacing whereas observations in Y were normally distributed with $\mu_Y = X$ (i.e., $\beta_0 = 0$ and $\beta_1 = 1$) and variance σ_y^2, which also varied across simulations as a result of variations in σ_e^2. Finally, a variant of regression sampling was also considered that merely differed in that values for X were randomly drawn from a normal distribution with mean and variance that varied across simulations. For each sample thus drawn, the coefficients of ordinary least-squares regression and regression through the origin were computed, and the two-tailed significance

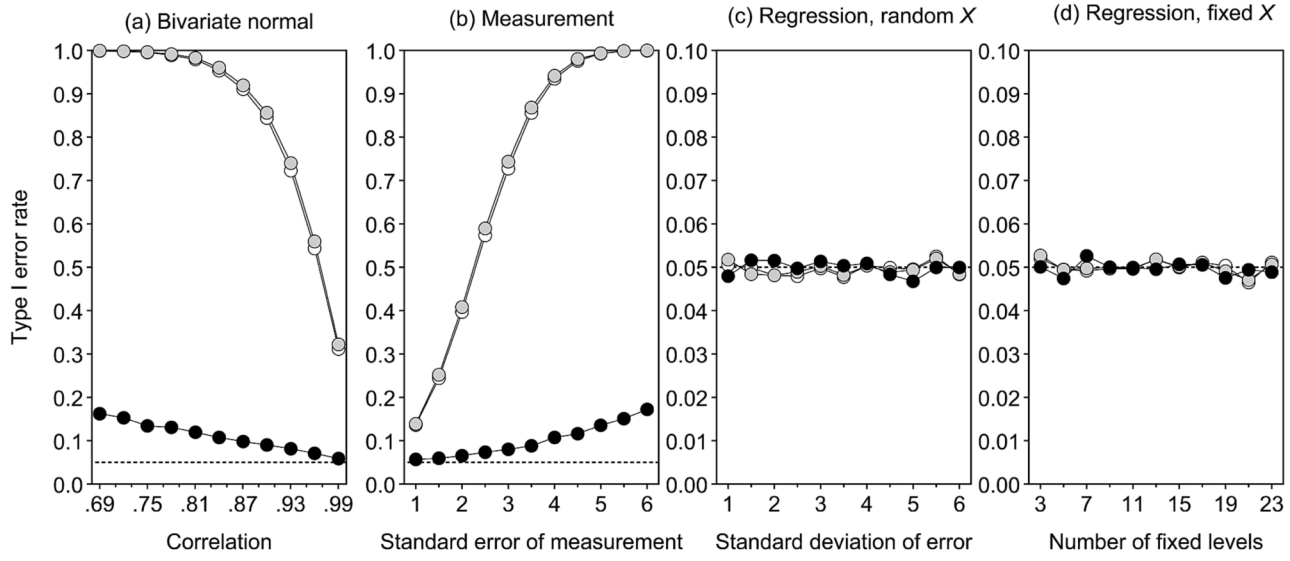

Figure 1. Type I error rate of the test that the regression intercept is zero (open symbols) and that the regression slope is unity (gray symbols) in ordinary least-squares regression and of the test that slope is unity in regression through the origin (solid symbols) under bivariate sampling as a function of the correlation between X and Y (a), under the measurement model as a function of the standard error of measurement (b), under lax regression sampling with random X as a function of error variance (c), and under strict regression sampling with error variance $\sigma_e^2 = 121$ as a function of the number of fixed levels for the observations in X (d). Nominal test size is $\alpha = .05$ and sample size is $n = 150$.

of the statistical tests of the null hypotheses $\beta_0 = 0$ and $\beta_1 = 1$ (in ordinary regression) or $\beta_1 = 1$ (in regression through the origin) was computed. When the entire set of 20,000 samples had been drawn, the empirical Type I error rate of each test was estimated as the proportion of samples for which the two-tailed probability was below .05 (for a nominal size-.05 test).

Results

Figure 1 shows representative results from one set of simulations in which $n = 150$, $\mu_X = 50$ and $\sigma_x^2 = 100$. Clearly, bivariate sampling when $\mu_Y = 50$ and $\sigma_y^2 = 100$ also (Figure 1a) rendered data for which the null hypotheses $\beta_0 = 0$ (open symbols) and $\beta_1 = 1$ (gray symbols) in ordinary regression were rejected overwhelmingly even when ρ_{XY} was as high as .99; the same was true for the hypothesis $\beta_1 = 1$ in regression through the origin (solid symbols). Ordinary regression and regression through the origin behaved similarly inappropriately under the measurement model when $\mu_X = \mu_Y = 50$ and $\sigma_x^2 = \sigma_y^2 = 100$ (Figure 1b), only approaching the nominal Type I error rate when the standard error of measurement was unrealistically low (i.e., when $\sigma_{\varepsilon_x}^2 = \sigma_{\varepsilon_y}^2 < 1$ which, given that $\sigma_x^2 = \sigma_y^2 = 100$, implies a reliability in excess of .99). In contrast, all tests turned out to be accurate under lax regression sampling (i.e., when X is random rather than fixed; Figure 1c) and, naturally, under strict regression sampling (Figure 1d). The results of other simulations in which sample size (in the range from $n = 20$ to $n = 300$) or the ranges and parameters of the distributions of the variables involved varied (i.e., X and Y in bivariate sampling, or T, ε_X, and ε_Y under the measurement model) were similar in that lax and strict regression sampling always rendered accurate tests, whereas bivariate sampling and the measurement model always rendered tests whose inaccuracy was evident. The particular test that most departed from its nominal size varied greatly across these conditions so that, for instance, the performance of all tests deteriorated meaningfully although at different rates as sample size increased; also, the test for $\beta_0 = 0$ seemed accurate under bivariate sampling only when $\mu_X = \mu_Y = 0$ but, in these conditions, the Type I error rate of the test for $\beta_1 = 1$ in regression through the origin was overly inaccurate. These results, then, show the inadequacy of NHST of regression parameters for testing equivalence with repeated measures under the most common empirical circumstances (i.e., bivariate sampling or measurement models).

A package for testing equivalence with repeated measures

Generally, the equivalence of two variables has several observable manifestations. Thus, thorough analyses that explore all the implications of equivalence are in order. We will start considering two approaches to testing equivalence, namely, omnibus tests and subject-by-subject tests. The former are useful for testing equivalence at the population level (i.e., when some model states that two variables should be identical in the population, as is the case for test scores on presumed parallel tests) whereas the latter are useful for testing equivalence on an individual by individual basis (e.g., when a model states that the equality of two variables may or may not hold according to individual characteristics so that whether or not equality holds at the population level is immaterial; a typical situation is the analysis of reaction times under various manipulations, where participants using different strategies may end up producing similar or different distributions of reaction times across conditions).

Omnibus tests

Perhaps the most stringent test of equivalence would be based on a package of statistical tests including a *t*-test for the equality of two related means through the well-known statistic

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

which is distributed as t with $n - 1$ degrees of freedom, a *t*-test for the equality of two related variances through the well-known statistic

$$T_v = \frac{\sqrt{n-2}(s_x^2 - s_y^2)}{2s_x s_y\Big/\sqrt{1-r_{xy}^2}}, \qquad (2)$$

which is distributed as t with $n - 2$ degrees of freedom, and a chi-square test of homogeneity of distributions. These tests are known to be robust to violation of their assumptions even with small samples (Benjamini, 1983; Cressie, 1980; García-Pérez & Núñez-Antón, 2009; Good & Hardin, 2006, Ch. 5) and they thus represent a useful starting point for equivalence tests. Arguably, the equivalence would be rejected if at least one of the three tests in this package rejects its null hypothesis, although in different practical applications a rejection by some of these tests may be regarded as less critical than a rejection by others (e.g., rejecting only the null hypothesis of equality of variances may not be regarded as critical when there is evidence that the two variables are measured with different precision). In any case, because several independent tests are applied to the same data, it is advisable to carry them out under the typical Bonferroni correction for the case of k independent tests, namely, the overall size-α test of equivalence would be rejected when at least one of the k tests was rejected at $\alpha^* = \alpha/k$. This is actually the typical approach to assessing parallelism in classical test theory

(García-Pérez, 2010), which also amounts to establishing the equivalence of repeated measures.

Besides the package just described, a one-shot approach to testing simultaneously for equality of means and variances was proposed by Bradley and Blackwood (1989), which uses the statistic

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

where *SSE* is the residual sum of squares from the regression of $D = X - Y$ on $S = X + Y$. If $\mu_X = \mu_Y$ and $\sigma_x^2 = \sigma_y^2$, this test statistic is distributed F with 2 and $n - 2$ degrees of freedom.

Although significance tests for correlations and regression coefficients are inadequate, the various omnibus tests just described must necessarily be complemented with evidence of a positive relation between X and Y. The reason is that all of the tests just discussed may concur in not rejecting the null hypothesis when X and Y differ blatantly. Consider the case of two variables that are highly negatively correlated with the same means, variances, and distributions: in such case, equivalence will not occur because $Y = 2\mu_X - X \neq X$ despite the fact that $\mu_X = \mu_Y$ and $\sigma_x^2 = \sigma_y^2$. For this purpose, we propose the use of the concordance correlation coefficient $\hat{\rho}_c$ (Lin, 1989, 1992, 2000), which is a scaled average measure of the squared perpendicular deviation of the data from the identity line in a scatter plot. The concordance coefficient, whose values range between −1 and 1, is defined as

$$\hat{\rho}_c = \frac{2 r_{xy} s_x s_y}{s_x^2 + s_y^2 + (\overline{X} - \overline{Y})^2} \quad (4)$$

and two of its interesting properties are that $\hat{\rho}_c = 0$ if and only if $r_{xy} = 0$ and that $\hat{\rho}_c = r_{xy}$ if and only if $\overline{X} = \overline{Y}$ and $s_x^2 = s_y^2$, whereas $|\hat{\rho}_c| \leq |r_{xy}|$ otherwise. Lin (1989) also shows that the Z-transformed concordance coefficient $\hat{Z} = 1/2 \ln[(1 + \hat{\rho}_c)/(1 - \hat{\rho}_c)]$ is asymptotically normally distributed with known mean and variance, which allows for NHST and the construction of confidence intervals. Although the concordance correlation coefficient outperforms the product-moment correlation for our present purposes, it is less clear that it can actually be used in an equivalence test. The reason is that strict equivalence occurs when $\rho_c = 1$, but it is not always clear what precise reference value $\rho_c < 1$ should be adopted in an empirical evaluation of equivalence using NHST or confidence intervals for ρ_c. We mention this coefficient here because it may eventually be useful, but we will not include it in our package except to provide the required additional evidence that X and Y are positively related.

To gather evidence as to the adequacy of the test package and the one-shot approach, simulations were carried out along the lines described in the preceding section to investigate accuracy and power. These simulations only considered bivariate sampling and the measurement model (the two cases that more often hold in empirical tests of equivalence) and excluded regression sampling for obvious reasons: The regression model assumes $\sigma_y^2 = \sigma_x^2 + \sigma_e^2$ so that the null hypothesis $\sigma_y^2 = \sigma_x^2$ included in the package can never be true under regression sampling. Also, the homogeneity test was not included in our simulation study because the typically small sample sizes in empirical studies of equivalence rarely allow for it.

The top part of Figure 2 shows the accuracy of the *t*-test for equality of two related means (open circles), the *t*-test for equality of two related variances (gray circles), the package consisting of both tests with the Bonferroni correction (solid circles), and the one-shot Bradley–Blackwood test (small open squares) under bivariate sampling with $\mu_X = \mu_Y = 50$ and $\sigma_x^2 = \sigma_y^2 = 100$ as a function of ρ_{XY} (left panel) and under the measurement model with $\mu_T = 50$, $\mu_{\varepsilon_x} = \mu_{\varepsilon_y}$ (so that $\mu_X = \mu_T = 50$), $\sigma_x^2 = \sigma_y^2 = \sigma_T^2 + \sigma_{\varepsilon_x}^2 = \sigma_T^2 + \sigma_{\varepsilon_y}^2 = 100$ as a function of σ_{ε_x} (right panel). Nominal test size was $\alpha = .05$ and sample size was $n = 150$. Across the board, both *t*-tests are accurate when considered separately, and their joint application with the Bonferroni correction is only minimally more accurate; the Bradley–Blackwood test, on the other hand, is similarly adequate. The bottom part of Figure 2 shows the power (also evaluated at $\alpha = .05$ with $n = 150$) of either *t*-test, of their joint application with the Bonferroni correction, and of the Bradley–Blackwood test as a function of variations in the mean of Y (upper row), in the variance of Y (center row), or in both (lower row), also under bivariate sampling with $\rho_{XY} = .7$ (left column) and under the measurement model with $\mu_X = \mu_T = 50$, $\sigma_x^2 = 100$, $\sigma_T^2 = 84$, and $\sigma_{\varepsilon_x}^2 = 16$ (right column). Under the measurement model, variations in the mean and variance of Y were accomplished by varying μ_{ε_y} and $\sigma_{\varepsilon_y}^2$ as needed. The package (solid circles) and the Bradley–Blackwood test (small open squares) are slightly less powerful than the applicable *t*-test (open and gray circles) when X and Y differ only as to mean or as to variance, but they are both certainly more powerful when X and Y differ as to both (the typical case when X and Y actually differ). In sum, then, the package of *t*-tests for means and variances with a Bonferroni correction is appropriate for testing equivalence, and so is the Bradley–Blackwood test, which in addition is slightly more powerful when means and variances both differ.

Subject-by-subject tests

The omnibus approach described thus far aims at assessing equivalence in the population from which the pairs of observations are sampled. However, some situations demand tests of equality at the individual level, particularly in cases in which a theoretical model indicates that

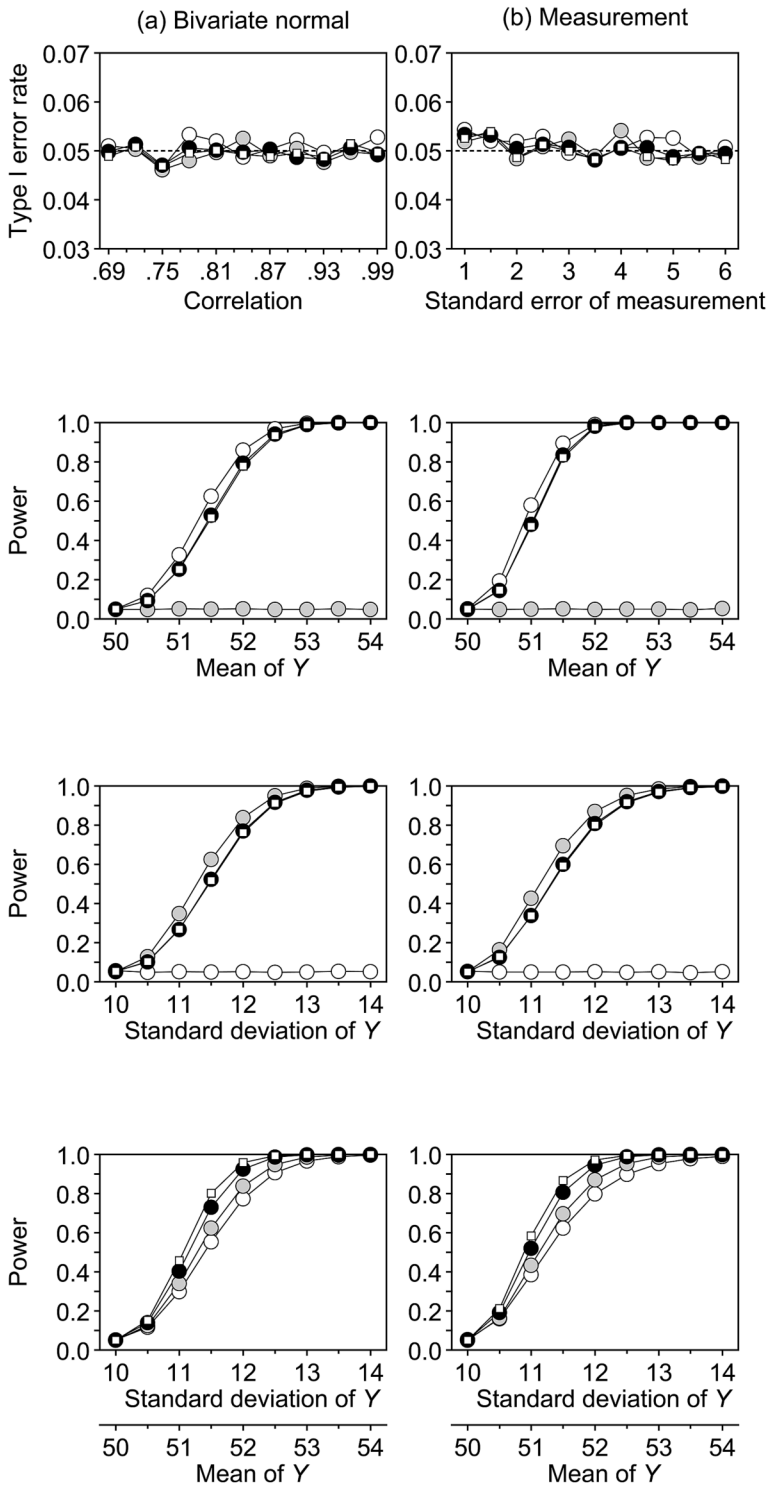

Figure 2. Top part: Type I error rates under bivariate sampling as a function of the correlation between X and Y (a) and under the conventional measurement model as a function of the standard error of measurement (b). Open circles reflect results for the t-test of equality of means, gray circles reflect results for the t-test of equality of variances, solid circles reflect results for the joint package with a Bonferroni correction, and small open squares reflect results for the Bradley–Blackwood test. Nominal test size is $\alpha = .05$ and sample size is $n = 150$. Bottom part: Power of the tests (graphical conventions as before) across variations in the mean of Y (upper row), in the variance of Y (center row), or in both (lower row). In all cases the mean of X was 50 and its variance was 100.

equivalence may or may not hold separately for each individual while making no claim as to how equivalence (as a binary categorical variable reflecting an individual characteristic) is distributed in the population. For instance, a given pair of observations in X and Y may represent performance measures derived from proportion correct across sets of trials in each of two different conditions, or average reaction times across a number of repeat trials in each of two different conditions. In these cases, the two conditions may interact with individual characteristics in ways that equivalence may hold for some participants and not for others. Error models for proportions (or transformations thereof) or error models for means, variances, and counts can then be used to test equivalence on a subject-by-subject basis. In such cases, the equality of X and Y would be separately tested for each individual in the sample using the particular statistical test or package that is appropriate given the nature of the two variables. If X and Y are means, t-tests may be appropriate; if they are proportions, a test of homogeneity of distributions would be appropriate and performed on a 2-way contingency table with rows (or columns) representing the variable (X or Y) and columns (or rows) representing the type of response (correct or incorrect). On the other hand, if the entire distributions of measurements under each condition are usable, then these subject-by-subject tests may provide the conditions for equivalence tests with independent measures (e.g., the set of N_X reaction times recorded for some individual under condition X and the set of N_Y reaction times recorded for the same individual under condition Y). We will not consider these cases here because they can be treated with the methods described in the introduction, which were designed for use with independent measures.

The distinction between omnibus and subject-by-subject tests is more important than it may seem at first glance. Consider the case in which an omnibus test reveals significant differences between X and Y. Subject-by-subject tests on the data, when feasible, might indicate that X and Y also differ significantly and in the same direction for the vast majority of individuals in the sample, which warrants the overall conclusion of the omnibus test. But these subject-by-subject tests might reveal instead that X and Y only differ significantly for a rather small subset of the individuals in the sample. The overall conclusion from the omnibus test is then unwarranted, because this conclusion entails that X differs from Y in the population when the truth is that X and Y differ significantly for only a few individuals but in a way that triggers rejection of the null hypothesis of equality by the omnibus test. It is incumbent on the researcher to define whether equivalence is expected to hold at the population level or, rather, it must be regarded as an individual characteristic subject to differences across participants.

Empirical illustration

The remainder of this paper illustrates the use of the set of tests just described in a particular application and also comments on some issues that may arise in such a quest for equivalence. The data that will be used for this purpose come from a recent study (Yeshurun, Carrasco, & Maloney, 2008; henceforth referred to as YCM) which tested four separate hypotheses associated with the standard difference model of 2AFC performance. The reason for using these data is mainly that it allows for a thorough discussion of several applications of equivalence tests with repeated measures, but also that further analyses and scrutiny of the data are possible to complement the quest for equivalence. It will be shown along the way that YCM committed other errors besides the use of regression analysis to answer their research questions; all of them will also be fixed in our application of the tests discussed in this paper, whose results do not support the conclusions originally raised by YCM. Before we describe all of these analyses, the next subsection provides the necessary background by describing the standard difference model of 2AFC performance and some of its variants, the four specific hypotheses that YCM set out to test, and the experiments that were designed to gather the data with which the hypotheses were tested.

The model, the experiments, and the data

The basic assumptions of signal detection theory (SDT) and the ensuing models of performance are described in a number of sources (e.g., Macmillan & Creelman, 2005; McNicol, 2005; Wickens, 2002). They are perhaps sufficiently well known also, but a brief description will be useful here if only to introduce our notation. SDT posits that the sensory effect elicited by presentation of a stimulus is a continuous random variable with some distribution. It is generally assumed without loss of generality that the distribution is normal with a mean μ that increases with stimulus level and with a variance that is independent of stimulus level and arbitrarily assumed to be unity.[2] Signal detection experiments involve a large number of trials under Yes–No tasks or under 2AFC tasks, among other tasks.

[2] Determining whether the variance of sensory effects is constant (the so-called fixed-noise assumption) or changes with stimulus level (the so-called variable-noise assumption) has proved elusive (see García-Pérez & Alcalá-Quintana, 2009), but this issue is inconsequential for the present analysis. It should nevertheless be stressed that YCM also adopted the fixed-noise assumption that we are endorsing here so that our analyses will not differ from theirs in this respect.

A Yes–No trial consists of a single temporal interval in which either the stimulus or a blank (chosen at random with equiprobability) is presented. If the stimulus was presented, the sensory effect S_s has mean $\mu_s > 0$; otherwise, the sensory effect S_b has mean $\mu_b = 0$. Observers are asked to indicate whether or not a signal had been presented, and they are assumed to set an unknown cutpoint c (often referred to as "criterion") such that they respond 'Yes' if the sensory effect is larger than c and they respond 'No' otherwise. If the stimulus has actually been presented (defining signal trials), a 'Yes' response is scored as a hit and a 'No' response is scored as a miss; if the stimulus has not been presented (defining no-signal trials), a 'Yes' response is scored as a false alarm and a 'No' response is scored as a correct rejection. From the empirical proportion $\hat{p}_h$ of hits across a series of signal trials and the empirical proportion $\hat{p}_{fa}$ of false alarms across a series of no-signal trials (with signal and no-signal trials randomly interwoven within a session), estimates of parameters μ and c are obtained as[3]

$$\hat{\mu} = \Phi^{-1}(\hat{p}_h) - \Phi^{-1}(\hat{p}_{fa}), \tag{5a}$$

$$\hat{c} = \Phi^{-1}(1 - \hat{p}_{fa}) \tag{5b}$$

(see, e.g., Wickens, 2002, ch. 2), where Φ is the unit-normal distribution function. In Yes–No tasks, $\hat{\mu}$ as estimated from Equation (5a) is also taken to be the estimate of sensitivity referred to as *d'* in SDT.

A temporal 2AFC trial in a signal detection experiment consists of two consecutive intervals one of which (chosen at random with equiprobability) presents the stimulus (and, thus, elicits a sensory effect S_s with mean $\mu_s > 0$) whereas the other presents a blank (and, thus, elicits a sensory effect S_b with mean $\mu_b = 0$). Observers are asked to report which interval presented the stimulus, and they are assumed to use a decision rule such that the reported interval is that which elicited the larger sensory effect. This decision rule subsumes a decision variable D representing the difference, say, between the sensory effect S_2 elicited in the second interval and the sensory effect S_1 elicited in the first interval so that the observer responds 'interval 1' when $D < 0$ and 'interval 2' when $D > 0$ (see Figure 3a). Therefore, observers also use a cutpoint but, unlike with the Yes–No paradigm, it is assumed to be fixed at $c = 0$. The observer's response is scored as correct if it matches the interval that actually presented the stimulus and otherwise it is scored as incorrect. In these conditions, the absolute value of the mean of the decision variable is estimated as

$$\hat{\mu} = \sqrt{2}\,\Phi^{-1}(\hat{p}), \tag{6}$$

where $\hat{p}$ is the overall empirical proportion of correct responses across interval-1 and interval-2 presentations of the stimulus.[4]

The foregoing description reflects what is known as the *standard difference model of 2AFC performance*, which assumes no differences in sensitivity across intervals (i.e., the mean of D is the same—except for a change of sign—whether the signal is presented in the first or the second interval) and also assumes unbiased observers (i.e., the cutpoint is at $c = 0$). According to this model, the proportion $\hat{p}_1$ of correct responses to stimuli presented in the first interval should be the same (within sampling error) as the proportion $\hat{p}_2$ of correct responses to stimuli presented in the second interval, but this is not always observed empirically. Two variants of the standard difference model compete to account for empirical cases in which $\hat{p}_1$ and $\hat{p}_2$ differ significantly. One of them states that there are actual differences in sensitivity across intervals. This yields the model illustrated in Figure 3b, where the observer is still unbiased (i.e., the cutpoint remains at $c = 0$) but the mean of D differs by more than a sign reversal when the stimulus is presented in the first or second intervals. Under the model of Figure 3b, these means are respectively estimated as

$$\hat{\mu}_1 = \sqrt{2}\,\Phi^{-1}(\hat{p}_1), \tag{7a}$$

$$\hat{\mu}_2 = \sqrt{2}\,\Phi^{-1}(\hat{p}_2), \tag{7b}$$

and note that values $\hat{\mu}_1$ and $\hat{\mu}_2$ can always be found under this model so that the observed $\hat{p}_1$ and $\hat{p}_2$ are exactly reproduced.

---

[3] Since the average sensory effect of the blank is assumed to be null and is not estimated, the only remaining parameters in the model are the cutpoint c and the average sensory effect of the stimulus, which we will subsequently denote μ rather than μ_s.

[4] For simplicity, we are referring only to the mean of D in order to avoid at this point the decision on a metric for *d'*, which also involves considerations as to how to treat the standard deviation of D. To illustrate, if *d'* in a 2AFC task is defined as the mean of the distribution of the sensory effects of the signal (e.g., Wickens, 2002, p. 97), $d' = \hat{\mu} = \sqrt{2}\Phi^{-1}(\hat{p})$ from Equation (6) and thus *d'* has the same value in Yes–No and 2AFC tasks; if, on the contrary, *d'* is defined as the distance between the two distributions in Figure 3a divided by their (common) standard deviation (e.g., McNicol, 2005, p. 67; Wickens, 2002, p. 100), the use of Equation (6) yields $d' = 2\hat{\mu}/\sqrt{2} = \sqrt{2}\hat{\mu} = 2\Phi^{-1}(\hat{p})$ and then *d'* is $\sqrt{2}$ times larger in a 2AFC task than it is in a Yes–No task. Then, the advantage of $\hat{\mu}$ at this point is that it is unequivocal.

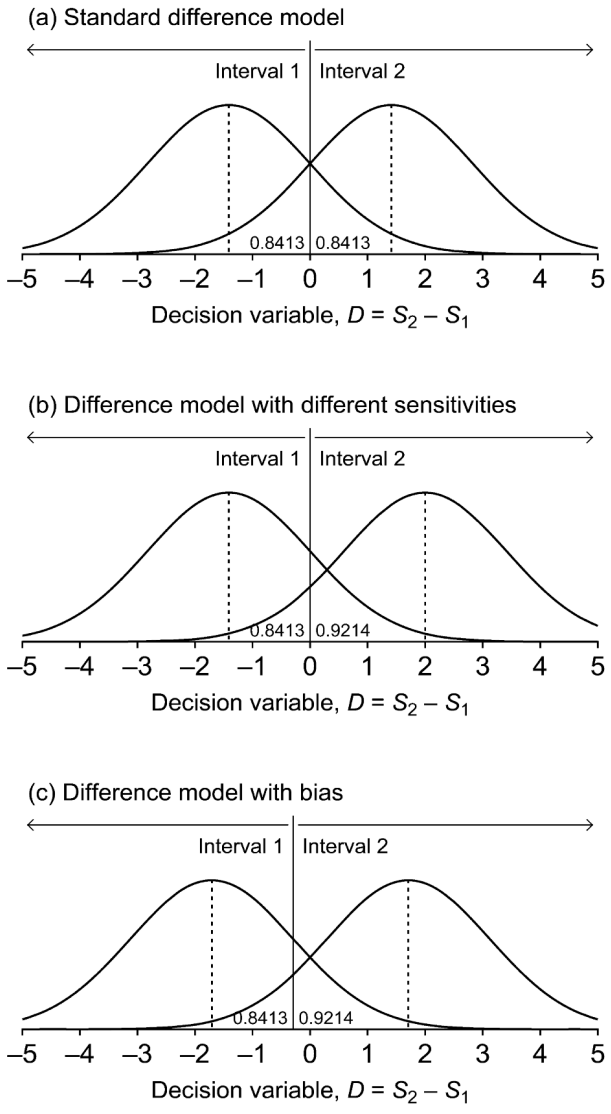

Figure 3. Variants of the difference model for 2AFC procedures. In each panel, the distribution on the left applies when the target is presented in the first interval, the distribution on the right applies when the target is presented in the second interval. A continuous vertical line partitions the horizontal axis at some point c into regions that lead the observer to give the responses indicated at the top. The probability that the response indicated at the top of each region is correct is given by the numerals near the bottom, which represent the area under the distribution in the applicable region. (a) Standard difference model, in which the mean of D is $|\mu| = \sqrt{2}$ and $c = 0$. (b) Difference model with differences in sensitivity across intervals such that $|\mu_1| = \sqrt{2}$ and $|\mu_2| = 1$ (μ_i being the mean of D when the target is presented in interval i) and with $c = 0$ also. (c) Difference model with bias, in which $|\mu_1| = |\mu_2| = \mu = 1.2069$ and $c = 0.2071$. Note that the two latter models predict interval bias in exactly the same amount.

The competing explanation for unequal empirical proportions of correct responses across intervals is depicted in Figure 3c. This alternative model assumes instead that sensitivity is the same in both intervals (i.e., $\mu_1 = \mu_2 = \mu$ again) but the observer is biased and sets the cutpoint at some $c \neq 0$. The model parameters are estimated as

$$\hat{\mu} = [\Phi^{-1}(\hat{p}_1) + \Phi^{-1}(\hat{p}_2)]/\sqrt{2}\,, \tag{8a}$$

$$\hat{c} = [\Phi^{-1}(\hat{p}_1) - \Phi^{-1}(\hat{p}_2)]/\sqrt{2}\,. \tag{8b}$$

Note that values $\hat{\mu}$ and $\hat{c}$ can also always be found under this model so that the observed $\hat{p}_1$ and $\hat{p}_2$ are exactly reproduced. Thus, significant differences in proportion correct across intervals can arise either as a result of actual differences in sensitivity across intervals without criterion bias (Figure 3b) or as a result of a non-zero criterion without differences in sensitivity across intervals (Figure 3c). A mixture of both characteristics is also possible, but we will not consider it here.

Yeshurun et al. (2008) analyzed data from 17 different 2AFC experiments in the literature and they reported significant differences in proportion correct across intervals, something that is usually dubbed 'interval bias.' They then set out to investigate four claims associated with the standard difference model of 2AFC performance, namely,

(1) that the 2AFC procedure is unbiased in the sense that the empirical proportion of correct responses is the same when the stimulus is presented in the first or the second interval,

(2) that the structure of the 2AFC procedure does not alter sensitivity in any way so that *d'* is the same in the first and the second interval,

(3) that *d'* from a 2AFC procedure is √2 times larger than *d'* from a Yes–No procedure (which should hold when the observer's sensitivity is the same in the two 2AFC intervals), and

(4) that *d'* from a 2AFC procedure is larger than *d'* from a Yes–No procedure even in the presence of differences in sensitivity across 2AFC intervals.

To investigate these issues, YCM carried out an experiment comprising two parts. One of them (which they referred to as a "2-way task") involved a conventional 2AFC procedure with 395–408 trials in which a signal was displayed in only one of two temporal intervals (with equiprobability) and the observer had to indicate in which interval the signal had been presented. The other (which they referred to as a "4-way task") was designed such that a signal could be presented in the first, the second, neither, or both intervals of a trial that was identical in all respects to trials in the 2-way task, and the observer had to indicate which of these four patterns had been presented. The overall number of trials (of the four types) varied between 402 and 408 across observers. The 2-way task provided

proportions of correct responses when the signal was presented in the first and the second intervals and also yielded a sensitivity estimate $\hat{d}'_{FC}$. The 4-way task was regarded as two consecutive Yes–No tasks and yielded empirical proportions of hits and false alarms in each Yes–No task as well as two estimates of sensitivity, $\hat{d}'_1$ and $\hat{d}'_2$, one from the Yes–No task in each interval. YCM found in their results "little evidence supporting the claims that [2AFC] is unbiased and that it does not alter sensitivity" and they also "reject[ed] the two claims associated with the difference model as a model of performance" (YCM, p. 1837). In sum, then, they rejected the four claims listed above.

On testing SDT models and the four claims just described, YCM did not carry out any tests for equality of means or variances and three of the claims were tested via regression through the origin. We re-analyzed their data thoroughly and through the package defined earlier, also using subject-by-subject tests where appropriate. We also carried out further analyses that may shed additional light on the issues that YCM investigated. By carrying out these additional analyses, we want to stress that equivalence with repeated measures often has additional implications and that researchers should use all available chances to scrutinize the data. The results of these re-analyses and new analyses are presented next. We used YCM's actual data, which were kindly provided by Dr. Carrasco. Table 1 lists, for each of the 20 observers in their experiments, the proportions correct in each interval of the 2-way task as well as the values of *d'* and *c* that they estimated for the 2-way task and for each interval of the 4-way task,[5] all taken from the files supplied by Dr. Carrasco.

The first claim, $p_1 = p_2$

This claim only involves manifest variables. YCM's 2-way task as well as their literature review unequivocally reveal interval bias in that $p_1 = p_2$ does not generally hold at the individual level (i.e., according to statistical tests under the subject-by-subject approach discussed earlier). Actually, interval bias was reportedly *not* shown by almost two thirds of the observers in YCM's 2-way task, and a non-negligible proportion of observers in the 17 experiments that they reviewed failed to show interval bias too. Jäkel and Wichmann (2006) also reported results revealing that only a subset of their observers showed interval bias.

One reason that we refer to this claim here is to correct an error in the computation of the subject-by-subject test carried out by YCM. They reported that the test was rejected for eight of their observers, and in their Figure 7 they indicated the significance level for each observer with a square whose side increased proportionally to minus the logarithm of the *p*-value of the test. They plotted two data points with the size that corresponds to significant *p*-values between .05 and .01, but our own computation of the test described in their Appendix B.1 indicates that the *p*-values for these two observers (#4 and #10 in Table 1) are, respectively, .081 and .092 so that these data points should have been plotted as single dots and they should not have been counted as instances of rejection of the null hypothesis by their criterion of a 95% significance level. Then, only six of their observers showed a significant interval bias, and these are marked in Table 1 with a star between the columns for $\hat{p}_1$ and $\hat{p}_2$. The final figure is that 70% (14/20) of their observers did not show any significant interval bias. The importance of this fact lies in that observers failing to show interval bias in YCM's 2-way task thus *seem* to be behaving according to the standard difference model and, then, their performance and sensitivity estimates in the 4-way task should perhaps stand out as different from those of observers showing interval bias in the 2-way task. In other words, if the performance of these observers in the 2-way task is consistent with the standard difference model, one should reasonably expect that the performance of these observers in other tasks is also consistent with predictions from the same model. On the other hand, the 2-way performance of the remaining observers is instead consistent with the models in Figures 3b and 3c and, again, their performance in other tasks is expected to be consistent with predictions of one or the other of those models. Checking out this consistency thus becomes a further test of the difference model, and the results of these additional tests will be presented in the next section.

Although the hypothesis that $p_1 = p_2$ lends itself to subject-by-subject analyses, we should mention that size-.05 omnibus tests (i.e., with $\alpha^* = .025$, given that $k = 2$ and $\alpha = .05$) only rejected the null hypothesis of equality of variances ($t_{18} = -2.4816$, $p = .023$) and that the value of the concordance coefficient was .472. On the other hand, the Bradley–Blackwood test also rejected the null hypothesis of equality of means and variances ($F_{2, 18} = 6.5665$; $p = .007$). In sum, on a subject-by-subject basis, the hypothesis

[5] We should stress that estimates of the criterion *c* in the 2-way task were incorrectly computed by YCM. According to Equation (6) in YCM's Appendix B.2, the estimate should be computed as $\hat{c} = \Phi^{-1}(1 - \hat{p}_{1|2})$, where $\hat{p}_{1|2}$ is the proportion of incorrect responses when the signal was presented in interval 2, so that $1 - \hat{p}_{1|2}$ is actually the proportion correct when the signal was presented in interval 2, listed as $\hat{p}_2$ in Table 1. Thus, for observer #1 in Table 1, $\hat{c} = \Phi^{-1}(.9902) = 2.3339$ instead of 1.9746. We have been unable to work out the computation that may have rendered the estimates obtained by YCM, but we should also note that this error is inconsequential because YCM did not use criterion estimates in their analyses.

is rejected for only 6 out of 20 observers and, then, the data indicate that only a few observers (30% of the sample) show significant interval bias. Admittedly, insufficient evidence to reject the null (which occurred for 70% of the observers) is not to be taken as evidence for it (Blackwelder, 1982; Frick, 1995a). But, at the same time, rejecting the null (which occurred for 30% of the observers) does not imply accepting the alternative (Goodman & Royall, 1988; Hacking, 1965). In any case, and beyond this nihilism, evidence against the standard difference model claim that $p_1 = p_2$ is far from overwhelming in YCM's 2-way task.

The second claim, $d'_1 = d'_2$

The presence of interval bias for some observers in the 2-way task led YCM to seek evidence of differences in sensitivity when the target is presented in the first or the second 2AFC interval. They thus devised a 4-way task that, in their words, "directly measure[s] the sensitivity of the observer in the two intervals of a [2AFC] task" (YCM, p. 1843). Sensitivities d'_1 and d'_2 were thus estimated from the 4-way task under the untested and suspect assumption that the observer performs an independent Yes–No task on each interval of the 4-way trial. (A formal proof that this assumption is implied in YCM's estimation method is provided in Appendix A, where the implications of this assumption are also discussed.) Further, regression through the origin rejected the null hypothesis of unit slope and led YCM to conclude that "the observers are, overall, slightly more sensitive in the first interval than the second" (YCM, p. 1844). In other words, d'_1 and d'_2 differed significantly by YCM's analysis, thus supporting the model in Figure 3b.

But regression results are suspect when the data do not come from regression sampling, as illustrated in Figure 1 above. What do dependable tests of equivalence say instead? As for *t*-tests, the average $\hat{d}'_1$ was 2.725 with a standard deviation of 0.925 and the average $\hat{d}'_2$ was 2.488 with a standard deviation of 0.893, so that a paired-samples *t*-test for equality of means yields $t_{19} = 2.5636$ and a two-tailed *p*-value of .019 whereas a paired-samples *t*-test for equality of variances yields $t_{18} = 0.3357$ and a two-tailed *p*-value of .741. Then, because $k = 2$ and $\alpha^* = .025$ when $\alpha = .05$, equality of means is rejected and, then, the package of omnibus tests rejects the equivalence of d'_1 and d'_2. The value of the concordance coefficient was .871. In contrast, the alternative Bradley–Blackwood test does not reject the null hypothesis of equality of means and variances ($F_{2, 18} = 3.1890$; $p = .065$). These conflicting statistical conclusions and the borderline nature of the rejection (in one case) or not rejection (in the other) of the null indicates that the evidence against the hypothesis of equal sensitivities in both intervals is far from overwhelming.

But it is somewhat surprising that YCM tested the hypothesis that $p_1 = p_2$ on a subject-by-subject basis and then switched to linear regression through the origin for an omnibus test of the hypothesis that $d'_1 = d'_2$. Equality of d'_1 and d'_2 may hold for some observers and not for others just as equality of p_1 and p_2 holds for some observers and not for others. In fact, there is no reason to think that all observers will be more sensitive in one interval than in the other, just as there is no reason to think that all observers will show interval bias (and empirical evidence indicates that only a few observers actually show it). The approach described by Macmillan and Creelman (2005, p. 328), which we also describe and exemplify in Appendix B, could actually have been used for testing equality of d'_1 and d'_2 on a subject-by-subject basis, since $\hat{d}'_1$ and $\hat{d}'_2$ from YCM's 4-way task were estimated as independent Yes–No measures of sensitivity (as Appendix A proves). Application of this test with $\alpha = .05$ rejects the null hypothesis of equality for only five observers (indicated with a star on the left of the column for $\hat{d}'_2$ in Table 1). Then, YCM's claim that "observers are, overall, slightly more sensitive in the first interval than the second" misstates the facts because 15 of 20 observers actually did not show any significant difference in sensitivity across intervals, one was significantly more sensitive in the second interval, and only four were significantly more sensitive in the first interval. Then, the data indicate that, with a few exceptions, observers' sensitivities do not differ significantly across intervals.

We should note at this point that a quick look at the location of the stars in Table 1 reveals that out of the five observers for whom $d'_1 = d'_2$ is rejected, $p_1 = p_2$ is rejected for only two (observers #12 and #14 in Table 1). And, remarkably, for observer #12, $\hat{p}_1 < \hat{p}_2$ while $\hat{d}'_1 > \hat{d}'_2$. It is obvious that a perfect match in the outcomes of the two tests should not be expected, and also that any non-significant difference between $\hat{p}_1$ and $\hat{p}_2$ (alternatively, $\hat{d}'_1$ and $\hat{d}'_2$) should not necessarily have the same sign as a significant difference between $\hat{d}'_1$ and $\hat{d}'_2$ (alternatively, $\hat{p}_1$ and $\hat{p}_2$). But the divergent results of these two tests deserve further scrutiny, if only because each observers' performance is likely to have been produced by a particular version of the difference model (whichever it was for each observer) and, hence, some consistencies in the data should be expected.

Furthermore, if YCM's contention is correct that the 4-way task provides estimates of sensitivity in the two intervals of the 2-way task, $\hat{d}'_1$ and $\hat{d}'_2$ estimated from the 4-way task should have predictive validity and account (within sampling error) for observers' performance in the 2-way task. This is best understood by noting the implications of true differences in sensitivity across intervals of the 2-way task, which were illustrated in Figure 3b. If $\hat{d}'_1$ and $\hat{d}'_2$ from YCM's 4-way task are estimates of μ_1 and μ_2 during each interval of the 2-way task (as YCM explicitly claimed in the first paragraph of their Appendix B.3), these values should account through the model in Figure 3b for the actual proportion correct of their observers in each interval of

the 2-way task. We have compared actual performance in the 2-way task with the performance predicted through the model in Figure 3b under the assumption that $\hat{d}'_1$ and $\hat{d}'_2$ actually describe sensitivity in each of the intervals of the 2-way task. Details of how this prediction was obtained are given in Appendix C and the results are shown in Figure 4.

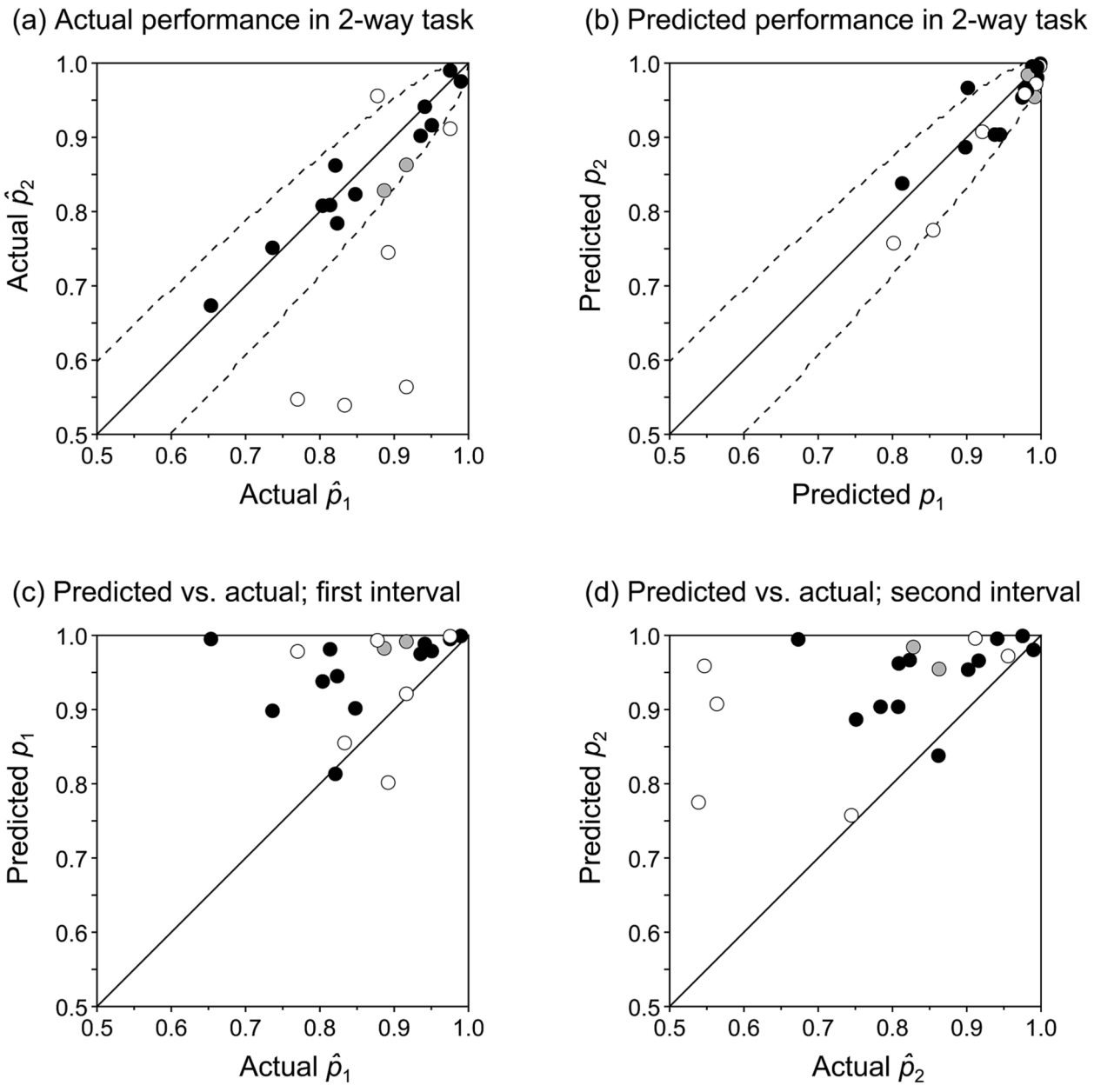

Figure 4. (a) Proportion correct in the second interval plotted against proportion correct in the first interval of the 2-way task for each of the observers participating in YCM's experiments. Dashed curves indicate the 95% confidence region around the null hypothesis H_0: $p_1 = p_2$ (represented by the diagonal line) according to the test used by YCM (see their Appendix B.1). Data points within this confidence region would not reject the null; the two gray symbols denote observers whose data reject the null according to YCM's report but do not reject it according to our own computation of their test; the six open symbols denote observers whose data reject the null in YCM's and our own computations. (b) Proportion correct in each interval of the 2-way task predicted by the sensitivity measures determined through the 4-way task. Symbols denote observers showing or not showing interval bias as described in the preceding panel. (c, d) Predicted versus actual performance in the first (c) and second (d) intervals of the 2-way task. The prediction is again obtained from sensitivity measures determined through the 4-way task. Symbols denote observers showing or not showing interval bias as described in the preceding panels.

Table 1
Data and parameter estimates in the 2-way and 4-way tasks of Yeshurun et al. (2008)

	2-way task				4-way task				
Observer	$\hat{p}_1$		$\hat{p}_2$	$\hat{d}'$	$\hat{c}$	$\hat{d}'_1$	$\hat{c}_1$		$\hat{d}'_2$	$\hat{c}_2$
1	0.9755		0.9902	4.3029	1.9746	3.6247	2.0599	*	2.9144	1.8895
2	0.8235		0.7843	1.7164	0.9015	2.2580	1.2121		1.8421	1.0074
3	0.9901		0.9755	4.3008	2.3282	4.3957	2.3338		4.4722	1.8895
4	0.9167		0.8627	2.4779	1.3527	3.3829	2.3337	*	2.3904	1.3830
5	0.9167	*	0.5637	1.5930	1.1321	1.9975	1.1594		1.8762	1.3490
6	0.9412		0.9412	3.1295	1.5647	3.2099	1.7599		3.6640	2.1779
7	0.8480		0.8235	1.9573	1.0120	1.8245	1.0707	*	2.5911	1.8209
8	0.8209		0.8621	2.0095	0.9517	1.2590	0.8100		1.3954	0.9188
9	0.9755	*	0.9118	3.3257	1.9402	4.3022	2.3337		3.7426	2.1779
10	0.8867		0.8284	2.1591	1.1760	2.9823	1.8895		3.0374	1.6544
11	0.9360		0.9020	2.8155	1.5063	2.7674	1.3517		2.3778	1.4861
12	0.8775	*	0.9559	2.8733	1.2074	3.5116	2.5827	*	2.7034	1.3517
13	0.8137		0.8088	1.7653	0.8883	2.9428	1.7048		2.5127	1.5647
14	0.8333	*	0.5392	1.0879	0.7189	1.4972	0.5131	*	1.0686	0.8382
15	0.7363		0.7512	1.3105	0.6440	1.7995	1.2564		1.7089	0.8779
16	0.8922	*	0.7451	1.9081	1.1434	1.1986	0.6140		0.9876	1.0491
17	0.8039		0.8079	1.7258	0.8557	2.1723	1.5622		1.8426	1.2622
18	0.9507		0.9163	3.0337	1.6350	2.8705	1.6055		2.5758	1.4861
19	0.7700	*	0.5473	0.8673	0.5392	2.8628	1.6006		2.4555	1.3463
20	0.6533		0.6735	0.8437	0.4035	3.6509	2.3319		3.5940	2.3319

Data reported for the 2-way task are proportion correct in the first and second intervals ($\hat{p}_1$ and $\hat{p}_2$), sensitivity $\hat{d}'$, and criterion $\hat{c}$; a star between the columns labeled $\hat{p}_1$ and $\hat{p}_2$ indicates that proportion correct differs significantly across intervals for the corresponding observer. Data reported for the 4-way task are sensitivity $\hat{d}'$ and criterion $\hat{c}$ for the Yes–No tasks carried out in the first and second intervals (indicated by subscripts); a star on the left of the column labeled $\hat{d}'_2$ indicates that sensitivity differs significantly across intervals for the corresponding observer.

Data points in Figure 4a show, for each observer, actual proportion correct in the second interval of the 2-way task against actual proportion correct in the first interval. Data points from the 12 observers reported by YCM to *not show* interval bias (i.e., observers for whom $p_1 = p_2$ within sampling error) are indicated with solid symbols; of the remaining eight observers (whom YCM tagged as showing interval bias), two did not actually show interval bias as discussed in the preceding section, and their data points (indicated with gray symbols in Figure 4a) actually fall within the 95% confidence region (enclosed by dashed curves) around the null hypothesis $p_1 = p_2$ (diagonal line). Figure 4b shows an analogous plot of predicted performance obtained from the model of Figure 3b on the assumption that μ_1 and μ_2 in the 2-way task are given by $\hat{d}'_1$ and $\hat{d}'_2$ in the 4-way task. Predicted data for the observers showing or not showing interval bias in the actual 2-way task are still indicated with symbols of different shade. Two characteristics of these plots are worth pointing out. First, the large interval bias that some observers showed in the 2-way task does not come out in these predictions: Sensitivities estimated through the 4-way task predict that only one observer should show a minimal interval bias in the 2-way task, but this particular observer did not actually show it. Second, predicted proportion correct in either interval of the 2-way task is noticeably better than actual performance (this point is best appreciated in Figures 4c and 4d), with no observer expected to perform below the 75%-correct level on either interval of the 2-way task (compared to six observers actually performing below this level on at least one of the intervals), with 16 observers expected to give more than 90% correct responses in either interval (compared to only five observers actually showing this performance level), and with 12 observers expected to give more than 95% correct responses in either interval of the 2-way task (when only two actually showed this performance level).

Since the difference model with different sensitivities in each interval (given by $\hat{d}'_1$ and $\hat{d}'_2$) cannot predict observed performance in each interval of the 2-way task, these results may be viewed as rejecting the difference model as a model of performance in 2AFC tasks. But another possibility is that $\hat{d}'_1$ and $\hat{d}'_2$ are not valid estimates of sensitivity during the intervals of a 2AFC task. Actually, our demonstration and discussion in Appendix A suggests that this is likely to be the case. Interestingly, the difference model can be put aside completely in this inquiry by looking at the relation

between differences in $\hat{d}'_1$ and $\hat{d}'_2$ (which are estimated without recourse to the difference model) and differences in $\hat{p}_1$ and $\hat{p}_2$ (which are observed quantities that do not come from any imposed model). If interval bias in the 2-way task (the difference $\hat{p}_1 - \hat{p}_2$) is caused by differences in sensitivity across intervals and these latter are in turn revealed by the difference $\hat{d}'_1 - \hat{d}'_2$ in the 4-way task, one should expect a strong and positive relation between these differences. The scatter plot in Figure 5 shows instead that the difference $\hat{p}_1 - \hat{p}_2$ in the 2-way task is unrelated to the difference $\hat{d}'_1 - \hat{d}'_2$ in the 4-way task. Surprisingly enough, observers with the largest absolute values of $\hat{d}'_1 - \hat{d}'_2$ in the 4-way task (data points in the far left and far right of Figure 5) show non-significant and virtually null interval bias ($\hat{p}_1 - \hat{p}_2$) in the 2-way task, whereas observers with large and significant interval bias (open symbols in the upper part of Figure 5) show differences $\hat{d}'_1 - \hat{d}'_2$ that are non-significant and smaller than those of many observers for whom interval bias $\hat{p}_1 - \hat{p}_2$ is virtually null. In other words, interval bias in the 2-way task does not occur for observers showing large differences between $\hat{d}'_1$ and $\hat{d}'_2$ in the 4-way task, and vice versa. It thus seems untenable that $\hat{d}'_1$ and $\hat{d}'_2$ actually indicate sensitivities during the first and second intervals of the 2-way task and, in these circumstances, the difference model cannot be blamed for the failed predictions in Figure 4.

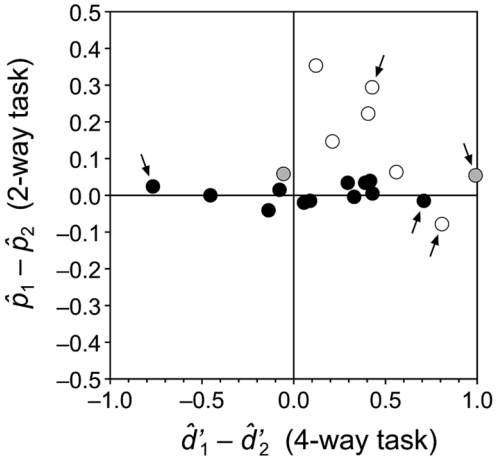

Figure 5. Relation between interval bias (the difference $\hat{p}_1 - \hat{p}_2$) in the 2-way task and differences in sensitivity across intervals (the difference $\hat{d}'_1 - \hat{d}'_2$) estimated with the 4-way task. The obvious lack of relation indicates that interval bias is not caused by (presumed) differences in sensitivity across intervals. Symbols of different shade denote observers showing or not showing interval bias as described in Figure 4a. Arrows indicate observers for whom $\hat{d}'_1$ and $\hat{d}'_2$ differed significantly.

The conclusion at this point is then that measures of sensitivity obtained with the 4-way task are likely to be contaminated by response bias and that they cannot be interpreted as measures of sensitivity during each of the 2AFC intervals. No further analyses are then justifiable to pursue the issues investigated by YCM and, thus, we will not evaluate claims (3) and (4). But we should point out that YCM again addressed these claims inadequately by testing for unit slope in regression through the origin. Nevertheless, we should turn to a discussion of claim (3) to point out other errors in YCM's analyses that should not spread in future research on these issues.

The third claim, $d'_{FC} = \tau d'_{YN}$

Citing Wickens (2002, p. 100ff), YCM argued that, when sensitivity differs across intervals, overall 2AFC sensitivity is given by $d'_{FC} = \sqrt{d'^2_1 + d'^2_2} = d'_1\sqrt{1 + \rho^2}$, where $\rho = d'_2/d'_1$ and d'_1 and d'_2 are the Yes–No sensitivities during each 2AFC interval. They then estimated ρ as the slope of the regression of d'_2 on d'_1 through the origin (which yielded $\hat{\rho} = 0.908$; see Appendix D for a critique of this approach to estimating ρ), defined $\hat{\tau} = \sqrt{1 + \hat{\rho}^2} = 1.35$, and used regression through the origin again to test the claim $d'_{FC} = \tau d'_{YN}$. This regression analysis led them to conclude that $\tau d'_{YN}$ differs significantly from d'_{FC}, and they rejected claim (3) and the difference model.

An important issue concerning claim (3) is the theoretical justification of the relation that YCM started with, namely, $d'_{FC} = \sqrt{d'^2_1 + d'^2_2}$. Wickens (2002, p. 122) proved the Pythagorean relation $d'_{AB} = \sqrt{d'^2_A + d'^2_B}$ in a completely different context, namely, when (1) A and B are stimuli that differ in quantity along some dimension but also in quality on two orthogonal dimensions, (2) d'_A represents Yes–No sensitivity to stimulus A, (3) d'_B represents Yes–No sensitivity to stimulus B, and (4) d'_{AB} represents *discrimination* sensitivity on single-presentation trials in which either stimulus A or stimulus B is presented and the observer must indicate whether A or B had been presented. (This case, and a similar proof, is also discussed in Section 7.2 of Macmillan & Creelman, 2005.)[6] The right-hand side of the expression certainly involves Yes–No sensitivities, but the left-hand side is not comparable to sensitivity in 2AFC trials in which one of the intervals presents a blank and the other presents a signal. Then, the Pythagorean relation that YCM assumed still has to be proved.

Consider for this purpose a one-dimensional representation of the 2AFC task for the case in which sensitivity differs across intervals (Figure 6a; adapted from YCM's Figure 5b). Thus, when the signal is presented in interval 1, the mean

[6] Wickens (2002) also proved a Pythagorean relation in another case, namely, Yes–No detection of compound stimuli compared to Yes–No detection of each component (see Sections 10.1 and 10.3 in Wickens, 2002; see also Equation 6.9 in Macmillan & Creelman, 2005). But this case is also not equivalent to 2AFC sensitivity compared to Yes–No sensitivities in each of the intervals of the 2AFC task.

of the decision variable is $-\mu_1$; when the signal is presented in interval 2, the mean is μ_2. These means are also the Yes–No sensitivities that would hold during each 2AFC interval (YCM, p. 1841), that is, $\mu_1 = d'_1$ and $\mu_2 = d'_2$. Then, under the conventional definition that sensitivity is given by the distance between those means divided by the common standard deviation, we arrive at $d'_{FC} = (d'_1 + d'_2)/\sqrt{2}$.

We can only think of one reasoning under which YCM's Pythagorean relation will (incorrectly) arise, and it is by using a two-dimensional representation such as that in Figure 6b (adapted from YCM's Figure 5a). Using the same definition of sensitivity as above, the distance between the means is actually $\sqrt{d'^2_1 + d'^2_2}$, which could be mistaken to be d'_{FC} because of the unit standard deviation in this representation. This demonstration is fallacious because in a two-dimensional representation sensitivity is not given by the distance between the means but by the sum of the distances from each mean to the decision boundary given by the dotted diagonal line (for a discussion and an illustration of this principle in the case of the reminder task, see pp. 180–181 and Figure 7.4 in Macmillan & Creelman, 2005).

In sum, when sensitivity differs across 2AFC intervals, $d'_{FC} = (d'_1 + d'_2)/\sqrt{2}$. This relation can also be written in the form $d'_{FC} = d'_1\,(1 + \rho)/\sqrt{2}$, where $\rho = d'_2/d'_1$, so that the relation $d'_{FC} = \tau d'_1$ involves $\tau = (1 + \rho)/\sqrt{2}$ instead of YCM's $\tau = \sqrt{1 + \rho^2}$. From the 4-way data in Table 1, YCM's erroneous estimate of τ through regression analysis was 1.351, whereas the correct estimate from the relation derived here and using the method described in Appendix D is 1.358. The difference is minimal for these data, but the correct relation should have been used and, more importantly, the incorrect relation stated by YCM without proof should not mislead researchers in the future.

In any case, there is also the issue that the theoretical relation that YCM set out to test involves d'_{FC} (2AFC sensitivity) and d'_{YN} (Yes–No sensitivity measured in a Yes–No task). YCM's use of d'_1 in place of a true measure of d'_{YN} (which would have been very easy to obtain experimentally) seems inconsistent with their own stance: By their own admission, the 4-way task demonstrated that "the temporal structure of the [2AFC] task altered sensitivity in one or both intervals and that measured sensitivity is not independent of the psychophysical method used to measure it" (YCM, p. 1848). The logical consequence of this statement is that they should never have regarded $\hat{d}'_1$ from their 4-way task as a valid estimate of what d'_{YN} might have been in an unaltered, stand-alone Yes–No procedure. For this reason, claims (3) and (4) cannot be regarded as properly tested by YCM, and their data cannot be used to test those claims in any reasonable sense.

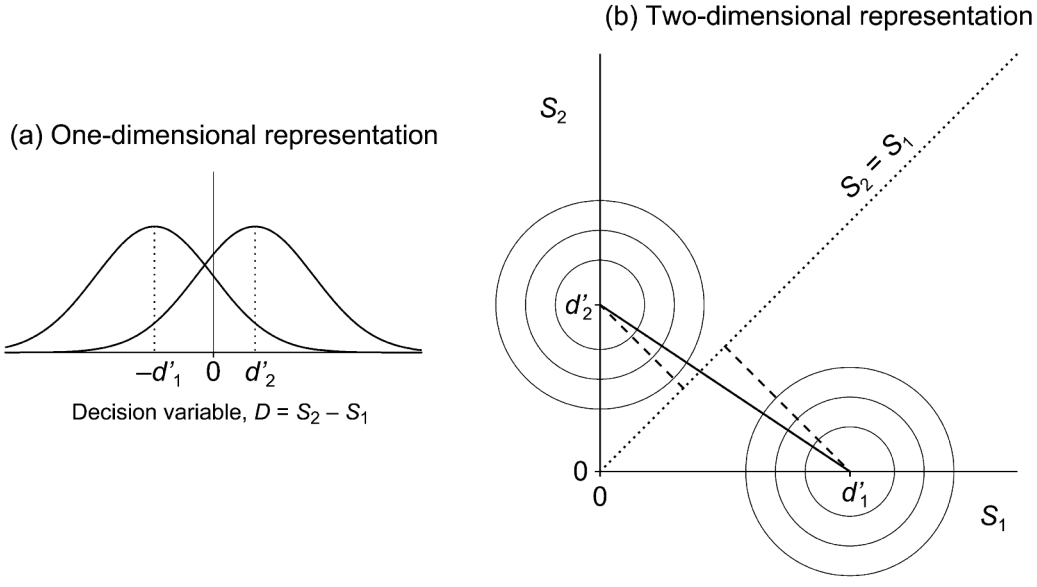

Figure 6. (a) One-dimensional representation of a 2AFC task with different sensitivities across intervals. The Gaussian on the left describes the distribution of the decision variable when the signal is presented in the first interval; the Gaussian on the right describes the same distribution when the signal is presented in the second interval. The means are, respectively, $-d'_1$ and d'_2 , and the variance is 2. (b) Two-dimensional representation of a 2AFC task with different sensitivities across intervals. The axes represent sensory effect S_1 of the signal when presented in the first interval and the sensory effect S_2 of the signal presented in the second interval. Concentric circles represent the bivariate distribution of sensory effects in the two intervals on trials in which the signal is presented in the first interval (lower right) or the second interval (upper left). The means are located at $(d'_1, 0)$ and $(0, d'_2)$, and the variance is unity along each dimension. The diagonal line is the decision boundary under the standard difference model. The distance between the means (indicated by a solid line) is the Pythagorean sum of d'_1 and d'_2. The distances from the mean of each distribution to the decision boundary are indicated with dashed lines, and the sum of these distances is d'_{FC}.

Discussion

Our re-analysis of YCM's data indicate that the difference model of performance in 2AFC tasks cannot be rejected beyond a reasonable doubt. It will be useful to summarize the picture that arises from our analyses.

Our first analysis shows that, on a subject-by-subject basis, claim (1) cannot be rejected for 14 (and not just 12) of YCM's 20 observers, who did not show significantly different proportions of correct responses in the first and second intervals of the 2-way task. Omnibus tests of this claim reject the null, but interval bias is likely an individual characteristic that demands the subject-by-subject analysis that YCM actually performed on their data. The presence of interval bias in some observers rejects the standard difference model (Figure 3a), but it does not reject the difference model in general because two of its versions (illustrated in Figures 3b and 3c) are compatible with interval bias.

Our second analysis reveals that claim (2) cannot be rejected on a subject-by-subject basis for 15 of YCM's 20 observers, who did not show significant differences in sensitivity in the first and second intervals of the 4-way task. Alternative omnibus tests of this claim yield mixed results, but the reasons that justify subject-by-subject analyses for claim (1) also apply to claim (2). Additional characteristics of YCM's data (discussed in our Figures 4 and 5) suggest that sensitivities estimated during each of the intervals of the 4-way task do not apply during the 2-way task. We have also noted that sensitivity during the 4-way task is higher than it is during the 2-way task and we have argued (Appendix A) that these differences may be the result of top-down influences that make the 2-way and 4-way tasks incomparable. This conclusion makes claims (3) and (4) impossible to test with the present data.

Beyond the inadequate use of regression analyses, we have described several other errors in YCM's analyses. While three of them are empirically marginal or have only theoretical importance, the other error has a major bearing on YCM's conclusion. In particular, YCM were aware that observers might not approach the 4-way task as two independent Yes–No tasks, and they contended that their maximum-likelihood approach to obtaining sensitivity estimates would wade through any potential problems in this respect. By deriving closed-form estimators for sensitivity and criterion under YCM's maximum-likelihood approach (Appendix A), we have shown that the estimates actually embody the assumption that observers perform two independent Yes–No tasks. The fact that YCM used numerical methods rather than closed-form expressions to obtain the estimates does not change this characteristic. In addition, the fact that the number of parameters in the model equals the number of independent data sources does not leave room for testing the goodness of the fit. Without evidence as to whether or not the observers actually approached the 4-way task as two independent Yes–No tasks, what YCM's estimates d'_1 and d'_2 actually represent is not at all clear. Further research that bypasses this problem is needed to investigate potential differences in sensitivity across 2AFC intervals (claim (2)) as well as the relation between 2AFC and Yes–No sensitivity (claims (3) and (4)).

The issues that YCM sought to investigate (interval bias, differences in sensitivity across 2AFC intervals, and the validity of the difference model of performance in 2AFC procedures) still require further research. The fact that interval bias is not a universal characteristic and that only some observers show it implies that a realistic model of 2AFC performance must be compatible with the presence and the absence of interval bias. Alternative versions of the difference model depicted in Figures 3b and 3c are compatible with this presence and absence, although none of these two models appear testable (i.e., they can only be fitted to the data and the fit to any possible data will be perfect for both models with no degrees of freedom left to test the inescapably perfect fit).

Conclusion

The need to test equivalence with repeated measures arises in two different contexts. One is in method comparison studies or in the assessment of agreement between instruments, where the underlying latent variables are the same. The second context occurs in theoretical work involving different latent variables in identical or commensurate scales. We have described various complementary approaches to testing equivalence with repeated measures, including omnibus and subject-by-subject tests. We have also shown that these approaches are accurate and powerful, whereas regression analysis (whether unconstrained or through the origin) is inadequate under the circumstances usually surrounding quests for equivalence (bivariate sampling or the measurement model). Finally, we have used dependable procedures in a re-analysis of YCM's data and we have also carried out additional analyses on the data to further complement the quest for equivalence, because a researcher's actual goal is to test experimental hypotheses thoroughly by assessing all of their implications and not just to apply a fixed set of statistical tests (Gigerenzer, 1993, 1998).

The test packages that we have described and applied represent a further step towards a good effort to find an effect if it exists, by looking at different manifestations of equivalence. In addition, we have shown that its application can be further complemented with tests of additional relations or expectations derived from the model under consideration (for a further illustration of this point, see Baguley et al., 2006). Therefore, an inquiry about whether some model is

empirically adequate is not only based on tests of equivalence and the problematic issue of accepting null hypotheses. After all, an experimenter's goal is not to apply a fixed protocol of statistical tests to the data set at hand but rather to explore and scrutinize the data so as to extract as much relevant and useful information as possible that bears on the research questions which prompted the investigation.

References

Alcalá-Quintana, R., & García-Pérez, M. A. (2007). A comparison of fixed-step-size and Bayesian staircases for sensory threshold estimation. *Spatial Vision*, *20*, 197-218.

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician*, *32*, 307-317. doi:10.2307/2987937

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics – Theory and Methods*, *12*, 2663-2692. doi:10.1080/03610928308828634

Astrua, M., Ichim, D., Pennecchi, F., & Pisani, M. (2007). Statistical techniques for assessing agreement between two instruments. *Metrologia*, *44*, 385-392. doi:10.1088/0026-1394/44/5/015

Baguley, T., Lansdale, M. W., Lines, L. K., & Parkin, J. K. (2006). Two spatial memories are not better than one: Evidence of exclusivity in memory for object location. *Cognitive Psychology*, *52*, 243-289. doi:10.1016/j.cogpsych.2005.08.001

Benjamini, Y. (1983). Is the *t* test really conservative when the parent distribution is long-tailed? *Journal of the American Statistical Association*, *78*, 645-654. doi:10.2307/2288133

Blackwelder, W. C. (1982). "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials*, *3*, 345-353. doi:10.1016/0197-2456(81)90059-3

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*, 307-310. doi:10.1016/j.ijnurstu.2009.10.001

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*, 135-160. doi:10.1191/096228099673819272

Bland, J. M., & Altman, D. G. (2003). Applying the right statistics: Analyses of measurement studies. *Ultrasound in Obstetrics and Gynecology*, *22*, 85-93. doi:10.1002/uog.122

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425-440. doi:10.1007/s11336-006-1447-6

Bradley, E. L., & Blackwood, L. G. (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician*, *43*, 234-235. doi:10.2307/2685368

Brink, W. P. van den, & Koele, P. (1980). Item sampling, guessing and decision-making in achievement testing. *British Journal of Mathematical and Statistical Psychology*, *33*, 104-108.

Casella, G. (1983). Leverage and regression through the origin. *The American Statistician*, *37*, 147-152. doi:10.2307/2685876

Chatterjee, S., Hadi, A. S., & Price, B. (2000). *Regression Analysis by Example* (3rd edition). New York, NY: Wiley.

Corina, D. P. (1999). On the nature of left hemisphere specialization for signed language. *Brain and Language*, *69*, 230-240. doi:10.1006/brln.1999.2062

Cox, N. J. (2006). Assessing agreement of measurements and predictions in geomorphology. *Geomorphology*, *76*, 332-346. doi:10.1016/j.geomorph.2005.12.001

Cressie, N. (1980). Relaxing assumptions in the one-sample *t*-test. *Australian Journal of Statistics*, *22*, 143-153. doi:10.1111/j.1467-842X.1980.tb01161.x

Cusack, R., & Carlyon, R. P. (2003). Perceptual asymmetries in audition. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 713-725. doi:10.1037/0096-1523.29.3.713

Diederich, A., & Colonius, H. (2011). Modeling multisensory processes in saccadic responses: Time-window-of-integration model. In M. M. Murray & M. T. Wallace (Eds.), *The Neural bases of multisensory processes*. Boca Raton, FL: CRC Press, in press.

Dierdorff, E. C., & Morgeson, F. P. (2007). Consensus in work role requirements: The influence of discrete occupational context on role expectations. *Journal of Applied Psychology*, *92*, 1228-1241. doi:10.1037/0021-9010.92.5.1228

Dixon, P., & O'Reilly, T. (1999). Scientific versus statistical inference. *Canadian Journal of Experimental Psychology*, *53*, 133-149. doi:10.1037/h0087305

Dunn, G., & Roberts, C. (1999). Modelling method comparison data. *Statistical Methods in Medical Research*, *8*, 161-179. doi:10.1191/096228099668524590

Dunnett, C. W., & Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 × 2 tables. *Biometrics*, *33*, 593-602. doi:10.2307/2529457

Edgell, S. E. (1995). Commentary on "Accepting the null hypothesis." *Memory & Cognition*, *23*, 525. doi:10.3758/BF03197252

Eisenhauer, J. G. (2003). Regression through the origin. *Teaching Statistics*, *25*, 76-80. doi:10.1111/1467-9639.00136

Ferrand, L. (1999). Why naming takes longer than reading? The special case of Arabic numbers. *Acta Psychologica*, *100*, 253-266. doi:10.1016/S0001-6918(98)00021-3

Freund, R. J., Wilson, W. J., & Sa, P. (2006). *Regression Analysis: Statistical Modeling of a Response Variable* (2nd edition). Burlington, MA: Academic Press.

Frick, R. R. (1995a). Accepting the null hypothesis. *Memory & Cognition*, *23*, 132-138. doi:10.3758/BF03210562

Frick, R. R. (1995b). A reply to Edgell. *Memory & Cognition*, *23*, 526. doi:10.3758/BF03197253

García-Pérez, M. A. (1989). Item sampling, guessing, partial information and decision-making in achievement testing. In E. E. Roskam (Ed.), *Mathematical Psychology in Progress* (pp. 249-265). Berlin, Germany: Springer.

García-Pérez, M. A. (2010). *Statistical criteria for parallel tests: A comparison of accuracy and power*. Manuscript submitted for publication.

García-Pérez, M. A., & Alcalá-Quintana, R. (2009). Fixed vs. variable noise in 2AFC contrast discrimination: Lessons from psychometric functions. *Spatial Vision*, *22*, 273-300. doi:10.1163/156856809788746309

García-Pérez, M. A., & Núñez-Antón, V. (2009). Accuracy of power-divergence statistics for testing independence and homogeneity in two-way contingency tables. *Communications in Statistics – Simulation and Computation*, *38*, 503-512. doi:10.1080/03610910802538351

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Methodological issues*. (pp. 311-339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199-200. doi:10.1017/S0140525X98281167

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology*, *63*, 527-537. doi:10.1348/000711009X475853

Good, P. I., & Hardin, J. W. (2006). *Common errors in statistics (and how to avoid them)* (2nd edition). Hoboken, NJ: Wiley.

Goodman, S. N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health*, *78*, 1568-1574. doi:10.2105/AJPH.78.12.1568

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

Hacking, I. (1965). *The logic of statistical inference*. Cambridge, UK: Cambridge University Press.

Hahn, G. J. (1977). Fitting regression models with no intercept term. *Journal of Quality Technology*, *9*, 56-61.

Hawkins, D. M. (2002). Diagnostics for conformity of paired quantitative measurements. *Statistics in Medicine*, *21*, 1913-1935. doi:10.1002/sim.1013

Hays, S., & McCallum, R. S. (2005). A comparison of the pencil-and-paper and computer-administered Minnesota Multiphasic Personality Inventory–Adolescent. *Psychology in the Schools*, *42*, 605-613. doi:10.1002/pits.20106

Hietanen, J. K., & Leppänen, J. M. (2003). Does facial expression affect attention orienting by gaze direction cues? *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 1228-1243. doi:10.1037/0096-1523.29.6.1228

Hollands, J. G., & Spence, I. (1998). Judging proportion with graphs: The summation model. *Applied Cognitive Psychology*, *12*, 173-190. doi:10.1002/(SICI)1099-0720(199804)12:2<173::AID-ACP499>3.0.CO;2-K

Huntsman, L. A. (1998). Testing the direct-access model: GOD does not prime DOG. *Perception & Psychophysics*, *60*, 1128-1140. doi:10.3758/BF03206163

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, *6*, 1307-1322. doi:10.1167/6.11.13

Jordan, P. J., & Troth, A. C. (2004). Managing emotions during team problem solving: Emotional intelligence and conflict resolution. *Human Performance*, *17*, 195-218. doi:10.1207/s15327043hup1702_4

Kane, M. J., Poole, B. J., Tuholski, S. W., & Engle, R. W. (2006). Working memory capacity and the top-down control of visual search: Exploring the boundaries of "executive attention." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 749-777. doi:10.1037/0278-7393.32.4.749

Kirkwood, T. B. L. (1981). Bioequivalence testing – A need to rethink. *Biometrics*, *37*, 589-591. doi:10.2307/2530573

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*, 255-268. doi:10.2307/2532051

Lin, L. I.-K. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, *48*, 599-604. doi:10.2307/2532314

Lin, L. I.-K. (2000). Correction: A note on the concordance correlation coefficient. *Biometrics*, *56*, 324-325.

Lin, L., Hedayat, A. S., Sinha, B., & Yang, M. (2002). Statistical methods for assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association*, *97*, 257-270. doi:10.1198/016214502753479392

Loftus, G. (1985). Johannes Kepler's computer simulation of the universe: Some remarks about theory in psychology. *Behavior Research Methods, Instruments, & Computers*, *17*, 149-156.

Los, S. A. (2004). Inhibition of return and nonspecific preparation: Separable inhibitory control mechanisms in space and time. *Perception & Psychophysics*, *66*, 119-130. doi:10.3758/BF03194866

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A user's guide*. Mahwah, NJ: Erlbaum.

McNicol, D. (2005). *A primer of Signal Detection Theory*. Mahwah, NJ: Erlbaum.

Metzler, C. M. (1974). Bioavailability – A problem in equivalence. *Biometrics*, *30*, 309-317. doi:10.2307/2529651

Miller, J. (1996). The sampling distribution of *d'*. *Perception & Psychophysics*, *58*, 65-72. doi:10.3758/BF03205476

Mukherjee, C., White, H., & Wuyts, M. (1998). *Econometrics and data analysis for developing countries*. New York, NY: Routledge.

Myers, R. H. (1990). *Classical and modern regression with applications* (2nd edition). Boston, MA: PWS-KENT.

Neter, J., Kutner, M. H., Wasserman, W., & Nachtsheim, C. J. (1996). *Applied linear statistical models* (4th edition). Chicago, IL: Irwin.

Perea, M., & Rosa, E. (2002). Does the proportion of associatively related pairs modulate the associative priming effect at very brief stimulus-onset asynchronies? *Acta Psychologica*, *110*, 103-124. doi:10.1016/S0001-6918(01)00074-9

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*, 553-565. doi:10.1037//0033-2909.113.3.553

Rorden, C., Karnath, H.O., & Driver, J. (2001). Do neck-proprioceptive and caloric-vestibular stimulation influence covert visual attention in normals, as they influence visual neglect? *Neuropsychologia*, *39*, 364-375. doi:10.1016/S0028-3932(00)00126-3

Russo, R., Fox, E., & Bowles, R. J. (1999). On the status of implicit memory bias in anxiety. *Cognition and Emotion*, *13*, 435-456. doi:10.1080/026999399379258

Saint-Aubin, J., & Poirier, M. (1999). Semantic similarity and immediate serial recall: Is there a detrimental effect on order

information? *Quarterly Journal of Experimental Psychology*, *52*(A), 367-394. doi:10.1080/027249899391115

Segrin, C. (2004). Concordance on negative emotion in close relationships: Transmission of emotion or assortative mating? *Journal of Social and Clinical Psychology*, *23*, 836-856. doi:10.1521/jscp.23.6.836.54802

Selwyn, M. R., Demptster, A. P., & Hall, N. R. (1981). A Bayesian approach to bioequivalence for the 2 × 2 changeover design. *Biometrics*, *37*, 11-21. doi:10.2307/2530518

Selwyn, M. R., & Hall, N. R. (1984). On Bayesian methods for bioequivalence. *Biometrics*, *40*, 1103-1108. doi:10.2307/2531161

Sen, A., & Srivastava, M. (1990). *Regression analysis. Theory, methods, and applications*. New York, NY: Springer.

Smith, R. W., & Kounios, J. (1996). Sudden insight: All-or-none processing revealed by speed–accuracy decomposition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1443-1462. doi:10.1037//0278-7393.22.6.1443

Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics*, *59*, 1-22. doi:10.3758/BF03206843

Spence, C., & Driver, J. (1998). Auditory and audiovisual inhibition of return. *Perception & Psychophysics*, *60*, 125-139. doi:10.3758/BF03211923

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning*, *19*, 193-198. doi:10.1016/0149-7189(96)00011-0

Van Berkum, J. J. A. (1997). Syntactic processes in speech production: The retrieval of grammatical gender. *Cognition*, *64*, 115-152. doi:10.1016/S0010-0277(97)00026-7

van Stralen, K. J., Jager, K. J., Zoccali, C., & Dekker, F. W. (2008). Agreement between methods. *Kidney International*, *74*, 1116-1120. doi:10.1038/ki.2008.306

Tipples, J., & Sharma, D. (2000). Orienting to exogenous cues and attentional bias to affective pictures reflect separate processes. *British Journal of Psychology*, *91*, 87-97. doi:10.1348/000712600161691

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371-386.

Tryon, W. W., & Lewis, C. (2008). An inferential confidence interval method for establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*, *13*, 272-277. doi:10.1037/a0013158

Turner, M. E. (1960). Straight line regression through the origin. *Biometrics*, *16*, 483-485. doi:10.2307/2527698

Vatakis, A., & Spence, C. (2008). Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica*, *127*, 12-23. doi:10.1016/j.actpsy.2006.12.002

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision*, *8*(9), 1-11. doi:10.1167/8.9.14

Wang, C. M., & Iyer, H. K. (2008). Fiducial approach for assessing agreement between two instruments. *Metrologia*, *45*, 415-421. doi:10.1088/0026-1394/45/4/006

Westgard, J. O., & Hunt, M. R. (1973). Use and interpretation of common statistical tests in method-comparison studies. *Clinical Chemistry*, *19*, 49-57. doi:10.1373/clinchem.2007.094060

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, *32*, 741-744. doi:10.2307/2529259

Westlake, W. J. (1979). Statistical aspects of comparative bioavailability trials. *Biometrics*, *35*, 273-280. doi:10.2307/2529949

Westlake, W. J. (1981). Bioequivalence testing – A need to rethink (Reader reaction response). *Biometrics*, *37*, 591-593.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. New York, NY: Oxford.

Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, *48*, 1837-1851. doi:10.1016/j.visres.2008.05.008

Zampini, M., Brown, T., Shore, D. I., Maravita, A., Röder, B., & Spence, C. (2005). Audiotactile temporal order judgments. *Acta Psychologica*, *118*, 277-291. doi:10.1016/j.actpsy.2004.10.017

Received June 17, 2010
Revision received February 08, 2011
Accepted March 21, 2011

APPENDIX A

Although YCM regarded their 4-way task as two independent Yes–No tasks, they admitted that observers could perform it differently and they noted that "if we were certain that the observer is carrying out the actual task as two independent Yes–No tasks, then the data could be fit as data from two signal detection tasks using the methods described in Appendix B.2. However, we cannot assume that, for example, the observer's judgment in one interval affects the judgment in the other and consequently we fit the data by the method of maximum likelihood. For example, the human observer may be biased against responding NN and an N response in either interval would affect the probability of an N response in the other" (YCM, p. 1850).[7] As we will formally prove here, YCM's maximum-likelihood approach did not protect their estimates against bias of this or other forms in the 4-way task, and the resultant estimates actually embody the untested and suspect assumption that observers perform two independent Yes–No tasks.

Although YCM obtained maximum-likelihood estimates of d'_1, c_1, d'_2, and c_2 numerically (see their Appendix B.3), closed-form expressions for the estimators can actually be derived which are identical to those that apply to data from two independent Yes–No tasks. To demonstrate it, we start with the log-likelihood equation in YCM's Equation (9), namely,

$$\lambda(d'_1, c_1, d'_2, c_2) = \sum_{i=1}^{4}\sum_{j=1}^{4} n_{ij} \log\left[p_{ij}(d'_1, c_2, d'_2, c_2)\right], \tag{A1}$$

where i and j respectively denote the rows and columns of the 4 × 4 table describing the trial types and possible outcomes of the 4-way task (see Figure A1, and note that the table therein has the last two rows and columns labeled differently with respect to an analogous table in Figure 6 of YCM; the reasons will become apparent in Appendix B), n_{ij} denotes the observed number of responses in cell (i, j) of the table, and p_{ij} denotes the probability of an observation falling in cell (i, j). For simplicity, the hit and false-alarm probabilities in the Yes–No task of interval k will be respectively denoted h_k and f_k and, according to YCM's Equation (10),

		Response			
		yn	ny	yy	nn
Trial type	YN	$h_1\,(1–f_2)$	$(1–h_1)\,f_2$	$h_1\,f_2$	$(1–h_1)(1–f_2)$
	NY	$f_1\,(1–h_2)$	$(1–f_1)\,h_2$	$f_1\,h_2$	$(1–f_1)(1–h_2)$
	YY	$h_1\,(1–h_2)$	$(1–h_1)\,h_2$	$h_1\,h_2$	$(1–h_1)(1–h_2)$
	NN	$f_1\,(1–f_2)$	$(1–f_1)\,f_2$	$f_1\,f_2$	$(1–f_1)(1–f_2)$

Figure A1. Table representing trial types (rows) and responses (columns) in YCM's 4-way task. Contents indicate cell probabilities according to YCM's model for the 4-way task.

$$h_k = 1 - \Phi(c_k - d'_k), \tag{A2a}$$

$$f_k = 1 - \Phi(c_k). \tag{A2b}$$

The probabilities p_{ij} to be inserted in Equation (A1) are thus given by combinations of these two expressions, as illustrated in YCM's Equation (11). The particular combination that holds in each cell is shown in Figure A1 as the content of that cell. Inserting these 16 probabilities into Equation (A1), taking partial derivatives with respect to each of the four parameters and simplifying the results yields

[7] The middle sentence in this three-sentence excerpt expresses an idea that contradicts those in the two other sentences and that is inconsistent with the thread of the paragraph from which these three sentences are extracted. We are inclined to think that what YCM meant to say in the second sentence is that "we cannot assume that the observer's judgment in one interval *does not affect* the judgement in the other."

$$\frac{\partial \lambda}{\partial d'_1} = (n_{11}+n_{13}+n_{31}+n_{33})\frac{h'_1}{h_1} - (n_{12}+n_{14}+n_{32}+n_{34})\frac{h'_1}{1-h_1}, \tag{A3a}$$

$$\frac{\partial \lambda}{\partial c_1} = (n_{11}+n_{13}+n_{31}+n_{33})\frac{h'_1}{h_1} - (n_{12}+n_{14}+n_{32}+n_{34})\frac{h'_1}{1-h_1} + (n_{21}+n_{23}+n_{41}+n_{43})\frac{f'_1}{f_1} - (n_{22}+n_{24}+n_{42}+n_{44})\frac{f'_1}{1-f_1}, \tag{A3b}$$

$$\frac{\partial \lambda}{\partial d'_2} = (n_{22}+n_{23}+n_{32}+n_{33})\frac{h'_2}{h_2} - (n_{21}+n_{24}+n_{31}+n_{34})\frac{h'_2}{1-h_2}, \tag{A3c}$$

$$\frac{\partial \lambda}{\partial c_2} = (n_{22}+n_{23}+n_{32}+n_{33})\frac{h'_2}{h_2} - (n_{21}+n_{24}+n_{31}+n_{34})\frac{h'_2}{1-h_2} + (n_{12}+n_{13}+n_{42}+n_{43})\frac{f'_2}{f_2} - (n_{11}+n_{14}+n_{41}+n_{44})\frac{f'_2}{1-f_2}, \tag{A3d}$$

where h'_k and f'_k in each equation respectively denote the partial derivatives of h_k and f_k with respect to the parameter in the corresponding equation. Note that the partial derivative of λ with respect to d'_1 in Equation (A3a) only includes eight terms involving the hit probability h_1. This is because terms involving the false-alarm probability f_1 or the hit and false-alarm probabilities h_2 and f_2 are not functions of d'_1, as is clear from Equations (A2a) and (A2b). Similarly, the partial derivative of λ with respect to c_1 in Equation (A3b) includes eight terms involving the hit probability h_1 (which are identical to those in the partial derivative with respect to d'_1) and eight more terms involving the false-alarm probability f_1, but it does not include any term involving the hit and false-alarm probabilities h_2 and f_2 because these are not functions of c_1. The same holds for partial derivatives of λ with respect to d'_2 and c_2 in Equations (A3c) and (A3d). Equating each of these partial derivatives to zero, inserting in them the expressions for h_k, f_k, h'_k, and f'_k, and simplifying the results yields the following system of four equations in four unknowns:

$$\left(\sum_{j=1}^{4} n_{1j} + \sum_{j=1}^{4} n_{3j}\right)\Phi(c_1 - d'_1) - (n_{12}+n_{14}+n_{32}+n_{34}) = 0, \tag{A4a}$$

$$\left[(n_{22}+n_{24}+n_{42}+n_{44}) - \left(\sum_{j=1}^{4} n_{2j} + \sum_{j=1}^{4} n_{4j}\right)\Phi(c_1)\right]\Phi(c_1 - d'_1)\left[1-\Phi(c_1 - d'_1)\right]\varphi(c_1) + \left[(n_{12}+n_{14}+n_{32}+n_{34}) - \left(\sum_{j=1}^{4} n_{1j} + \sum_{j=1}^{4} n_{3j}\right)\Phi(c_1 - d'_1)\right]\Phi(c_1)\left[1-\Phi(c_1)\right]\varphi(c_1 - d'_1) = 0, \tag{A4b}$$

$$\left(\sum_{j=1}^{4} n_{2j} + \sum_{j=1}^{4} n_{3j}\right)\Phi(c_2 - d'_2) - (n_{21}+n_{24}+n_{31}+n_{34}) = 0, \tag{A4c}$$

$$\left[(n_{11}+n_{14}+n_{41}+n_{44}) - \left(\sum_{j=1}^{4} n_{1j} + \sum_{j=1}^{4} n_{4j}\right)\Phi(c_2)\right]\Phi(c_2 - d'_2)\left[1-\Phi(c_2 - d'_2)\right]\varphi(c_2) + \left[(n_{21}+n_{24}+n_{31}+n_{34}) - \left(\sum_{j=1}^{4} n_{2j} + \sum_{j=1}^{4} n_{3j}\right)\Phi(c_2 - d'_2)\right]\Phi(c_2)\left[1-\Phi(c_2)\right]\varphi(c_2 - d'_2) = 0, \tag{A4d}$$

where φ is the unit-normal probability density function. Note that the first pair of equations has only d'_1 and c_1 as unknowns, whereas the second pair has only d'_2 and c_2 as unknowns. These two pairs of equations each in two unknowns can easily be solved algebraically to arrive at

$$\hat{d}'_1 = \Phi^{-1}\left[\frac{n_{11}+n_{13}+n_{31}+n_{33}}{\sum_{j=1}^{4} n_{1j} + \sum_{j=1}^{4} n_{3j}}\right] - \Phi^{-1}\left[\frac{n_{21}+n_{23}+n_{41}+n_{43}}{\sum_{j=1}^{4} n_{2j} + \sum_{j=1}^{4} n_{4j}}\right], \tag{A5a}$$

$$\hat{c}_1 = \Phi^{-1}\left[1 - \frac{n_{21}+n_{23}+n_{41}+n_{43}}{\sum_{j=1}^{4} n_{2j} + \sum_{j=1}^{4} n_{4j}}\right], \tag{A5b}$$

$$\hat{d}'_2 = \Phi^{-1}\left[\frac{n_{22}+n_{23}+n_{32}+n_{33}}{\sum_{j=1}^{4} n_{2j}+\sum_{j=1}^{4} n_{3j}}\right] - \Phi^{-1}\left[\frac{n_{12}+n_{13}+n_{42}+n_{43}}{\sum_{j=1}^{4} n_{1j}+\sum_{j=1}^{4} n_{4j}}\right], \tag{A5c}$$

$$\hat{c}_2 = \Phi^{-1}\left[1-\frac{n_{12}+n_{13}+n_{42}+n_{43}}{\sum_{j=1}^{4} n_{1j}+\sum_{j=1}^{4} n_{4j}}\right]. \tag{A5d}$$

Consider each of the two terms in the estimator for d'_1 in Equation (A5a), and recall that subscripts for n refer to the ordinal position of the labeled rows and columns in Figure A1. The numerator of the argument of the first term is the sum of occasions in which the observer responded 'y' on the first interval when a signal was actually present, and the denominator is the number of trials in which a signal was present in the first interval. In other words, the argument of the first term of Equation (A5a) is the hit rate in the first interval. The argument of the second term is analogously seen to represent the false-alarm rate in the first interval, so that Equation (A5a) is the typical Yes–No estimate of sensitivity, and Equation (A5b) is also the typical Yes–No estimate of criterion. The same holds for Equations (A5c) and (A5d), but applied to the Yes–No task in the second interval. Then, these expressions yield sensitivity and criterion estimates for Yes–No tasks as if they had been carried out independently during each interval of the 4-way task.

It should not come as a surprise that these estimates turn up as if the two Yes–No tasks were independent. Independence was actually built into YCM's likelihood equation: The cell probabilities in the table of Figure A1 were defined by YCM to be the product of the probability of the given outcome in the Yes–No task of the first interval and the probability of the given outcome in the Yes–No task of the second interval, which implies the assumption of independence. As a result, the 4 × 4 table in which YCM presented their results (see their Figure 6) is actually two nested 2 × 2 tables (see also Figure B1 below). The table at the upper level represents {Y, N} × {y, n} in the first interval, and each cell in this table contains in turn a 2 × 2 table representing {Y, N} × {y, n} in the second interval. These four 'inside' tables are identical in structure and cell probabilities. There are, therefore, only four independent pieces of data: The overall count in each 'inside' table gives the count in the corresponding cell of the 2 × 2 table at the upper level (with only two independent data sources), and the cell-by-cell sum of the four 'inside' tables gives the count in each cell of the 2 × 2 table at the lower level (which also has only two independent sources of data). With four parameters to estimate and only four independent data sources, there is always a single solution, no degrees of freedom left to test the fit, and a one-to-one mapping of data to parameter estimates, which is given by the expressions in Equations (A5a)–(A5d). We have checked that these expressions actually yield the estimates $\hat{d}'_1$, $\hat{c}_1$, $\hat{d}'_2$, and $\hat{c}_2$ reported in Table 1 (which were obtained by YCM using numerical methods) and we have noted only occasional discrepancies at or beyond the fourth decimal place, which are due to the numerical approximation used by YCM.

This demonstration has the additional implication of removing the ground beneath YCM's contention that their estimates of d'_1 and d'_2 are free of the response biases that they described. Without solid evidence that observers actually approached the 4-way task as two independent Yes–No tasks, estimates $\hat{d}'_1$ and $\hat{d}'_2$ obtained under the assumption of independence cannot be taken to reflect "pure" measures of sensitivity in the first and second 2AFC intervals. Furthermore, if $\hat{d}'_1$ and $\hat{d}'_2$ differ significantly (as YCM considered proved by their analyses), which of these estimates (if any) would represent what d' might have been in a stand-alone Yes–No task? YCM seemed to assume that it is $\hat{d}'_1$, because events in a Yes–No task are exactly those that take place up to the end of the first interval of a 4-way trial. However, by their own admission (stated in the quotation from their p. 1850 reproduced at the beginning of this appendix), the second interval in a 4-way trial may affect the observer's response to the first interval, breaking the equivalence with a stand-alone Yes–No task. In sum, it is not at all clear that $\hat{d}'_1$ and $\hat{d}'_2$ are adequate estimates of sensitivity in purportedly independent Yes–No tasks carried out in the first and second intervals of the 4-way task.

APPENDIX B

Differences in sensitivity across two Yes–No tasks are significant when zero is not in the confidence interval for the difference in *d'*. Computing this confidence interval requires estimating the variances of $\hat{d}'_1$ and $\hat{d}'_2$ (i.e., the estimated sensitivity *d'* in each of the two tasks), which are given by

$$\text{var}(\hat{d}'_i) = \frac{\hat{H}_i(1-\hat{H}_i)}{N_{\text{s},i}\left[\varphi(\Phi^{-1}(\hat{H}_i))\right]^2} + \frac{\hat{F}_i(1-\hat{F}_i)}{N_{\text{n},i}\left[\varphi(\Phi^{-1}(\hat{F}_i))\right]^2} \tag{B1}$$

(Macmillan & Creelman, 2005, p. 325), where $\hat{H}$ and $\hat{F}$ respectively stand for the observed hit and false-alarm rates, N_s and N_n respectively denote the number of signal and noise trials, and φ is the unit-normal probability density function. When these variances have been estimated, the 95% confidence interval for differences in sensitivity is obtained as $(\hat{d}'_1 - \hat{d}'_2) \pm 1.96\sqrt{\text{var}(\hat{d}'_1) + \text{var}(\hat{d}'_2)}$.

4-way task

Trial type	Response: yn	ny	yy	nn	
YN	76	1	12	13	102
NY	4	80	6	12	102
YY	19	10	66	6	101
NN	11	17	2	72	102

Interval 1

Signal	Response: y	n	
Y	173	30	203
N	23	181	204

Interval 2

Signal	Response: y	n	
Y	162	41	203
N	32	172	204

Figure B1. Sample data table (left panel) for observer #2 in Table 1, replotted from YCM's Figure 6, and the two tables (center and right panels) representing the outcomes in the Yes–No tasks of intervals 1 and 2 of the 4-way task.

In the 4-way task of YCM, the number of signal and noise trials in each implied Yes–No task and the hit and false-alarm rates in each of them can be easily determined from response tables such as that in YCM's Figure 6, which pertains to observer #2 in Table 1 and which we reproduce here on the left side of Figure B1. Note that the labeling of NN and YY trial types and nn and yy responses was swapped in Figure 6 of YCM, and the correct labeling is used in our Figure B1. (This is the reason that we used this labeling in Figure A1, discussed in Appendix A.) The two tables on the right of Figure B1 show data for the Yes–No tasks in intervals 1 and 2 of the 4-way task, which are obtained from the table on the left as exemplified by the shaded and hatched cells. (Justification for this rearrangement of data was provided in Appendix A.) For instance, the number of hits in the Yes–No task of interval 1 is given by the number of occasions in which the 4-way task presented a signal in interval 1 (the YY and YN trial types) and the observer also responded 'yes' on interval 1 (the yy and yn responses). Thus, 76 + 12 + 19 + 66 = 173, as indicated through shaded cells. Similar considerations yield the number of hits in the Yes–No task of interval 2 (80 + 6 + 10 + 66 = 162, as indicated through hatched cells), and yield also the remaining cases making up the two tables on the right of Figure B1. Thus, for this observer, $N_{s,1} = N_{s,2} = 203$, $N_{n,1} = N_{n,2} = 204$, $\hat{H}_1 = 173/203 = .8522$, $\hat{F}_1 = 23/204 = .1127$, $\hat{H}_2 = 162/203 = .7980$, and $\hat{F}_2 = 32/204 = .1569$.

As demonstrated in Appendix A, estimated sensitivity and criterion in each Yes–No task can be obtained from these alternative tables to yield the results that YCM obtained using a more elaborate numerical approach discussed in their Appendix B.3. Thus, for the Yes–No trial in the first interval of the 4-way task for this observer, $\hat{d}'_1 = \Phi^{-1}(173/203) - \Phi^{-1}(23/204) = 2.2580$ and $\hat{c}_1 = \Phi^{-1}(1 - 23/204) = 1.2121$; similar calculations for the Yes–No task in the second interval yield $\hat{d}'_2 = \Phi^{-1}(162/203) - \Phi^{-1}(32/204) = 1.8421$ and $\hat{c}_2 = \Phi^{-1}(1 - 32/204) = 1.0074$. Note that these estimates equal those reported in Table 1 for observer #2, which were obtained by YCM using numerical methods.

Back to our main point, and to illustrate the computation of confidence intervals for differences in sensitivity for our sample observer, inserting into Equation (B1) the observed hit and false alarm rates computed earlier and the numbers of signals and noise trials in the Yes–No task of each interval of the 4-way task yields $\text{var}(\hat{d}'_1) = 0.0250$ and $\text{var}(\hat{d}'_2) = 0.0213$. The confidence interval is then $(2.2580 - 1.8421) \pm 1.96\sqrt{0.0250 + 0.0213} = 0.4159 \pm 0.4217$. Because zero is in the interval, sensitivities do not differ significantly. Application of this strategy to all observers in YCM's 4-way task yields the results reported on the left of the column for $\hat{d}'_2$ in Table 1, where a star indicates that $\hat{d}'_1$ and $\hat{d}'_2$ differed significantly for that observer.

We should finally stress that confidence intervals thus computed, as well as the ensuing conclusions regarding the significance of differences in sensitivity, are based on variances estimated through Equation (B1), which only yield an approximation to the true variance of *d'*. Miller (1996; see his Table 2) has shown that this approximation underestimates the true variance of *d'*, particularly when the number of signal and noise trials is small (below 64 for each type of trial), *d'* is high (above 3.5), and the data include cases of perfect or null hit or false alarm rates. Yet, Miller (1996) also noted that the sampling distribution of *d'* is not exactly normal and that confidence intervals obtained through the approximation used in this appendix (which underestimates the variance of *d'* and assumes that *d'* is normally distributed with its mean at the true value of *d'*) can nevertheless be very accurate in comparison with those obtained through direct computation based on intensive simulations to obtain the actual sampling distribution of *d'* instead of a normal approximation with fixed mean (at the true *d'*) and variance given by Equation (B1). Moreover, Miller's Table 3 shows that approximate confidence intervals obtained under the normality assumption are often narrower than they should be according to exact computations based on the actual sampling distribution of *d'*. Then, the important implication of Miller's results is that decisions on the significance of differences in sensitivity such as those made in this appendix are generally slightly liberal in the sense that the null hypothesis $d'_1 = d'_2$ is rejected more often than it should be at any given confidence level. In other words, these confidence intervals yield a test that cannot be charged of bias towards accepting the null hypothesis $d'_1 = d'_2$.

APPENDIX C

Since $\hat{d}'_1$ and $\hat{d}'_2$ as estimated from the 4-way task differ (whether significantly or not) for all observers, predictions of performance in 2AFC tasks based on these estimates requires a recourse to the version of the difference model illustrated in Figure 3b, which allows for differences in sensitivity across intervals. Obtaining the predicted proportions of correct responses in each interval under this model requires specifying the mean of each of the distributions drawn in Figure 3b and, then, finding areas under Gaussian distributions (as given by the numerals at the bottom of Figure 3b for the particular means used in that illustration). YCM provided the information that is needed to resolve the question of what the means are: On describing the assumptions of the 4-way task in the first paragraph of their Appendix B.3, they stated that "when a signal is present in Interval j, the distribution of the random variable S_j is Gaussian with mean d'_j and variance 1. When a signal is absent, the distribution is Gaussian with mean 0, variance 1." Thus, under the model in Figure 3b, the mean of $D = S_2 - S_1$ for signals presented in the first interval is $-\hat{d}'_1$, whereas the mean for signals presented in the second interval is $\hat{d}'_2$, and in either case the variance is 2 because $D = S_2 - S_1$. Then, the probability that $D < 0$ for signal presentations in the first interval and the probability that $D > 0$ for signal presentations in the second interval are the 4-way predictions (because the location of these distributions has been determined from 4-way data) of proportion correct in each interval of the 2-way task (because the model applies to 2AFC tasks). From the illustration in Figure 3b, these predicted proportions of correct responses in the first and second intervals of the 2-way task are, then, $\Phi(\hat{d}'_1/\sqrt{2})$ and $1 - \Phi(-\hat{d}'_2/\sqrt{2}) = \Phi(\hat{d}'_2/\sqrt{2})$, which are straightforward generalizations of the expressions that are used when the means of the two distributions differ only in sign (see Equation 6.2 in Wickens, 2002).

For an illustration, consider observer #1 in Table 1, for whom $\hat{d}'_1 = 3.6247$ and $\hat{d}'_2 = 2.9144$. The predicted proportion correct in the first 2AFC interval is $\Phi(3.6247/\sqrt{2}) = .9948$ and the predicted proportion correct in the second 2AFC interval is $\Phi(2.9144/\sqrt{2}) = .9803$. Note also in Table 1 that the actual proportions correct for this observer in the 2-way task were, respectively, .9755 and .9902.

APPENDIX D

YCM regressed d'_2 on d'_1 through the origin, which yielded $d'_2 = 0.908\,d'_1$ in their notation. Then, algebraic manipulation of this regression equation led them to conclude that $\hat{\rho} = d'_2/d'_1 = 0.908$. This appendix discusses the validity of regression through the origin to estimate the ratio of the regressand to the regressor. To simplify our notation, we will rename $Y \equiv d'_2$ and $X \equiv d'_1$ so that we can use the less cumbersome notation for regression of Y on X.

Every source that discusses regression through the origin (e.g., Chatterjee, Hadi, & Price, 2000; Eisenhauer, 2003; Freund et al., 2006; Hahn, 1977; Myers, 1990; Neter, Kutner, Wasserman, & Nachtsheim, 1996; Sen & Srivastava, 1990) shows that this regression implies the model $Y = \beta_1 X + \varepsilon$, where ε is the residual. Unlike in ordinary least-squares regression with an intercept, the residuals ε do not have a zero mean because forcing the regression through the origin is generally inconsistent with the best fit. Whichever their mean is, residuals ε remain an integral part of the regression equation. To avoid carrying this term, the regression equation is typically written as $Y' = \beta_1 X$, where the apostrophe on the left-hand side indicates that this is a regression equation in which the residual has been omitted for simplicity (but not because it should not be there), and implicit in this notation is the fact that $Y' = Y - \varepsilon$. Most users finally write the regression equation improperly as $Y = \beta_1 X$, which only confounds the unwary by making the regression equation look like an ordinary equation that can easily be manipulated to render $\beta_1 = Y/X$.

But a least-squares regression equation is not an ordinary equation and it cannot be manipulated in this way. Making β_1 the subject of the true form of the regression equation yields $\beta_1 = (Y - \varepsilon)/X$, which is useless for YCM's purposes and still neglects the fact that ε is a random variable, not a fixed quantity. Making β_1 the subject of the simplified form of the regression equation renders $\beta_1 = Y'/X$, but this could by no means be taken for the ratio Y/X of the two variables, even if the apostrophe that marks the Y variable were omitted. Thus, YCM's estimate $\hat{\rho}$ is in error because it arises from illegal manipulation of the regression equation.

That the estimate $\hat{\rho}$ obtained by YCM is erroneous is also appreciated by noting that, with least-squares methods, the X and Y variables have different status and, then, regressing Y on X yields results that are non-trivially different from those of regressing X on Y. This is clearly noted by considering the closed-form expression for the slope β_1 of the regression line through the origin, which is $\dfrac{r_{xy}s_x s_y + \overline{X}\,\overline{Y}}{s_x^2 + \overline{X}^2}$ in the former case and $\dfrac{r_{xy}s_x s_y + \overline{X}\,\overline{Y}}{s_y^2 + \overline{Y}^2}$ in the latter. Thus, the slope of the regression line of Y on X is not the inverse of the slope of the regression line of X on Y, which further clarifies that one cannot move from one regression equation to the other (or elsewhere, for that matter) by algebraic manipulation. Actually, one cannot go anywhere by algebraic manipulation (as shown in the preceding paragraph) except to the trivial and otherwise useless result that $Y' = \beta_1 X$ yields $X = Y'/\beta_1$ (and note the apostrophe on Y in either expression). The implication is that an estimate of ρ obtained with the method of YCM would have been different (and not just by inversion) if they had regressed d'_1 on d'_2 instead. In any case, neither of these regression slopes actually estimates parameter ρ or its inverse, for a further reason discussed next.

It is convenient to remember the context in which parameter ρ arises, namely, the expression $d'_{FC} = \sqrt{d'^2_1 + d'^2_2} = d'_1\sqrt{1+\rho^2}$, where $\rho = d'_2/d'_1$ (YCM, p. 1843). If the true values of d'_1 and d'_2 were known, their ratio would yield ρ immediately; otherwise, ρ must be estimated as the ratio of estimates of d'_1 and d'_2. But note that not all observers should be differently sensitive in the two 2AFC intervals, and results presented earlier in this paper indicate that only 5 out of 20 observers showed significantly different sensitivities across intervals. Then, why estimate ρ once for all observers instead of individually for each of them? In other words, why should the data from an observer for whom $\hat{d}'_2/\hat{d}'_1 = 1$ be expected to show characteristics that hold under non-unit ratios estimated using data from other observers? From this point of view, $\hat{\rho} = \hat{d}'_2/\hat{d}'_1$ would better be separately estimated for each observer and, if needed, the average $\hat{\rho}$ could be obtained as usual. From the data in Table 1, $\hat{\rho}$ ranges across observers from 0.706 to 1.420, with an average of 0.920. However, no such group average would be needed in a subject-by-subject analysis of data. And there is also the issue that the relation $d'_{FC} = \sqrt{d'^2_1 + d'^2_2}$ is not correct and that an overall estimate of ρ is unnecessary.