Behav Res (2018) 50:2226–2255
https://doi.org/10.3758/s13428-017-0989-0

# Nonparametric tests for equality of psychometric functions

**Miguel A. García-Pérez**[1] · **Vicente Núñez-Antón**[2]

Published online: 7 December 2017


**Abstract** Many empirical studies measure psychometric functions (curves describing how observers' performance varies with stimulus magnitude) because these functions capture the effects of experimental conditions. To assess these effects, parametric curves are often fitted to the data and comparisons are carried out by testing for equality of mean parameter estimates across conditions. This approach is parametric and, thus, vulnerable to violations of the implied assumptions. Furthermore, testing for equality of means of parameters may be misleading: Psychometric functions may vary meaningfully across conditions on an observer-by-observer basis with no effect on the mean values of the estimated parameters. Alternative approaches to assess equality of psychometric functions per se are thus needed. This paper compares three nonparametric tests that are applicable in all situations of interest: The existing generalized Mantel–Haenszel test, a generalization of the Berry–Mielke test that was developed here, and a split variant of the generalized Mantel–Haenszel test also developed here. Their statistical properties (accuracy and power) are studied via simulation and the results show that all tests are indistinguishable as to accuracy but they differ non-uniformly as to power. Empirical use of the tests is illustrated via analyses of published data sets and practical recommendations are given. The computer code in MATLAB and R to conduct these tests is available as Electronic Supplemental Material.

**Keywords** Psychometric function · Nonparametric methods · Equality tests · Homogeneity of distributions

A large number of empirical studies in diverse areas of research require measuring observers' performance on some task as a function of stimulus magnitude. Most often, performance is expressed as proportion correct across a set of trials at each stimulus level and such data describe what is known as a psychometric function: A curve indicating how proportion correct varies with stimulus level. In other cases, observers' responses on each trial are judgments in three or more categories which are not (or cannot be) classified as correct or incorrect. Nevertheless, a set of psychometric functions still describes performance by indicating how the proportion of responses in each category varies with stimulus level. Most studies aim at assessing how performance varies across experimental conditions (using within-subjects or between-subjects designs) or across groups defined according to subject variables (in *ex post facto* designs). To serve these goals, psychometric functions need to be compared across groups or experimental conditions and several options are available for this purpose.

One option consists of fitting model curves to summarize each observer's performance via model parameters (usually location and slope of the psychometric function). Once this is done, parameter estimates (or transformations thereof) are subjected to comparison across groups or experimental conditions via *t* tests or ANOVAs (see, e.g., Donohue, Woldorff, & Mitroff, 2010; Gil, Rousset, & Droit-Volet, 2009; Lee & Noppeney, 2014; Tipples, 2010; Vroomen & Stekelenburg,

**Electronic supplementary material** The online version of this article (https://doi.org/10.3758/s13428-017-0989-0) contains supplementary material, which is available to authorized users.

✉ Miguel A. García-Pérez
miguel@psi.ucm.es

[1] Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain

[2] Departamento de Econometría y Estadística (E.A. III), Universidad del País Vasco UPV/EHU, Avda. Lehendakari Aguirre 83, 48015 Bilbao, Spain

2011). The validity of this parametric approach rests on the adequacy of the selected model curves and on the good fit to each observer's data; if these conditions do not hold, comparisons are compromised. A further problem with this approach is that it does not test equality of psychometric functions per se: It only tests for equality of the mean of the estimated parameters, which may hold true even when the psychometric functions differ systematically across conditions on an observer-by-observer basis.

A still parametric but less stringent option consists of defining the $K$ stimulus magnitudes at which data were collected as the levels of a repeated-measures factor for an ANOVA. When each trial only allows a binary response (e.g., correct or incorrect) the dependent variable is the proportion of, say, correct responses. These ANOVAs usually involve other repeated-measures or grouping factors, as needed by the design of the study (see, e.g., Capa, Duval, Blaison, & Giersch, 2014; Droit-Volet, Bigand, Ramos, & Oliveira Bueno, 2010; Gable & Poole, 2012; Wilbiks & Dyson, 2013; references listed in the preceding paragraph also reported ANOVAs of this type). This strategy allows testing for equality of psychometric functions directly across those other factors and their interaction, as it is clear beforehand that proportion correct will vary across stimulus levels. However, the use of this strategy is limited to cases in which only two response categories are allowed. On the other hand, the parametric assumptions of ANOVA do not hold when data are proportions, besides the almost sure violation of the assumptions of homoscedasticity and sphericity in such conditions.

There are situations in which these parametric approaches are either inapplicable or unadvisable. For instance, in within-subjects designs, psychometric functions are measured for each observer under several experimental conditions. Given that performance generally varies greatly across observers, aggregating data across them for a comparison of conditions adds unnecessary error variance and, thus, tests of equality of psychometric functions across conditions are needed on an observer-by-observer basis. The same holds when data for each condition need to be collected across several sessions with each observer, which calls for an analogous observer-by-observer test of equality of psychometric functions across sessions before data from them all are aggregated. Parametric approaches are inapplicable in all these cases and an ANOVA for categorical variables (referred to as CATANOVA; Anderson & Landis, 1980; Onukogu, 1985a, b) might seem appropriate, but we will show that CATANOVA does not measure up to its expected performance.

The work described in this paper set out to develop three fully nonparametric tests of equality of psychometric functions and to assess their statistical properties (accuracy and power). The tests were designed to be applicable for data collected at $K \geq 2$ stimulus levels in each of $I \geq 2$ conditions with a task that allows for $J \geq 2$ response categories in each trial. These tests are more general than that proposed by Logvinenko, Tyurin, and Sawey (2012), the applicability of which is limited to situations in which $I = J = 2$, and which is insensitive to certain differences between psychometric functions. The three tests are presented in the next section, which is followed by a description of the simulation study that assessed the accuracy and power of each test. Results are presented and discussed immediately afterwards, followed by a brief section documenting the unsuitability of CATANOVA. Examples of the application of these tests are next given using published data from several studies, including comparative examples of the outcomes of our nonparametric tests and a conventional parametric approach. Practical recommendations are presented before our final discussion. A computer code to conduct these tests in MATLAB and R is made available as Electronic Supplementary Material.

## Three nonparametric tests of equality of psychometric functions

To accompany our presentation with a suitable referent, consider the sample case in Fig. 1, involving two populations ($I = 2$) and psychometric functions reflecting the distribution of responses in a task with $J = 3$ response categories (so that there are three psychometric functions per population) at each of $K = 6$ stimulus levels. ("Population" is used here with its statistical meaning to refer to the conditions under which the data were collected; these might be, for instance, the two sessions over which data had been collected from an observer.) The left and center panels at the top of Fig. 1 plot the data and the bottom part tabulates them as indicated in the sketch at the top right, where $f_{ijk}$ is the observed frequency of responses by population $i$ in category $j$ at stimulus level $k$, $f_{i\cdot k} = \sum_{j=1}^{J} f_{ijk}$ is the $i$-th row marginal, $f_{\cdot jk} = \sum_{i=1}^{I} f_{ijk}$ is the $j$-th column marginal, and $N_k = f_{\cdot\cdot k} = \sum_{i=1}^{I}\sum_{j=1}^{J} f_{ijk}$ is the total number of observations at stimulus level $k$. The row marginal frequencies $f_{i\bullet k}$ represent the number of trials placed at stimulus level $k$ in population $i$. These numbers may have been fixed beforehand (e.g., for data collected with the method of constant stimuli) so that non-zero row marginal frequencies are guaranteed for all $i$ at each $k$. But data may also be collected with adaptive methods so that row marginal frequencies will generally vary across $i$ and $k$, potentially producing zero row marginal frequencies for some $i$ at some $k$. All $f_{i\bullet k}$ in the example of Fig. 1 are equal but the tests are applicable with arbitrarily different $f_{i\bullet k}$ as well as with arbitrary numbers of populations ($I \geq 2$), response categories ($J \geq 2$), and stimulus levels ($K \geq 2$), without constraints as to the aspects on which psychometric functions might differ across populations. The tests are thus more general than that developed by Logvinenko et al. (2012),

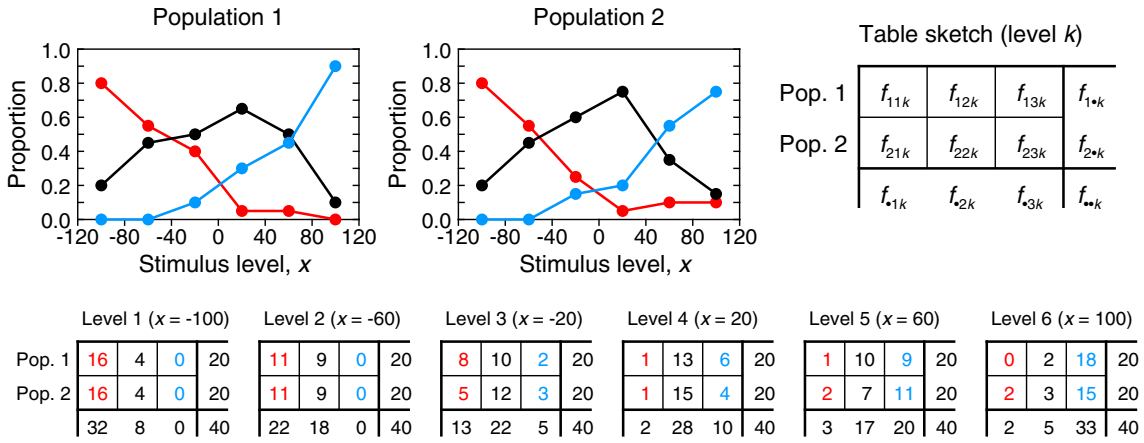



**Fig. 1** Sample data for application of the tests of equality of psychometric functions. The left and center panels at the top show empirical psychometric functions for two populations ($I = 2$) probed at $K = 6$ stimulus levels with a task that allows $J = 3$ response categories. Symbols indicate the proportion of responses in each category (red: first category; black: second category; blue: third category) at each stimulus level in each population. The table sketch at the top right indicates the notation used to refer to the counts of responses in each category (columns) by each population (rows) at stimulus level $k$. The $K = 6$ tables at the bottom depict the data at each stimulus level, with cell counts printed in the color used for the psychometric functions at the top

which was designed to test only for lateral displacement (with no other differences) when $I = 2$ and $J = 2$. Later we will come back to these sample data to discuss how a parametric approach would address the assessment of equality of psychometric functions.

Equality of psychometric functions implies homogeneity of the distributions of responses across the $J$ categories in all $I$ populations, although these distributions naturally vary across the $K$ stimulus levels. Under the null hypothesis of homogeneity, expected cell frequencies at stimulus level $k$ are given by $F_{ijk} = f_{i\cdot k} f_{\cdot jk}/N_k$ and Pearson's statistic

$$X_k^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(f_{ijk} - F_{ijk}\right)^2}{F_{ijk}} \tag{1}$$

measures deviation from homogeneity at each $k$. The tests described next differ as to how some variant of Eq. 1 combines deviations across $K$ tables to render an omnibus test statistic.

### The generalized Mantel–Haenszel test of homogeneity

The generalized Mantel–Haenszel statistic $Q_{\mathrm{GMH}}$ for $I \times J \times K$ tables (Birch, 1965; Landis, Heyman, & Koch, 1978; Mantel & Haenszel, 1959; Somes, 1986; see also Agresti, 2002, section 7.5.3) is defined by taking at each $k$ the subtable of pivotal cells that results from eliminating the last row and the last column, rendering $K$ subtables of size $(I - 1) \times (J - 1)$. (Incidentally, which row and column are eliminated is inconsequential.) Then, for each such subtable, place the pivotal observed frequencies $f_{ijk}$ (for $1 \le i \le I - 1$ and $1 \le j \le J - 1$) columnwise in a row vector $\mathbf{O}_k$ of size $1 \times (I-1)(J-1)$ so that $f_{ijk}$ maps onto the $m$-th element of $\mathbf{O}_k$, with $m = i + (j - 1)(I - 1)$ and similarly arrange the corresponding expected frequencies $F_{ijk}$ into a row vector $\mathbf{E}_k$ of the same size. Now define the variance–covariance matrix $\mathbf{V}_k$ of size $(I-1)(J-1) \times (I-1)(J-1)$ so that $v_{k_{mn}}$ relates to the pivotal cells $i_1 j_1$ and $i_2 j_2$, with $m = i_1 + (j_1 - 1)(I-1)$ and $n = i_2 + (j_2 - 1)(I-1)$. This matrix is readily computed as

$$\mathbf{V}_k = \frac{\left(N_k \mathrm{diag}(\mathbf{C}_k) - \mathbf{C}_k \mathbf{C}_k'\right) \otimes \left(N_k \mathrm{diag}(\mathbf{R}_k) - \mathbf{R}_k \mathbf{R}_k'\right)}{N_k N_k (N_k - 1)}, \tag{2}$$

where $\mathbf{C}_k = [f_{\cdot jk}]$ is a $(J-1) \times 1$ vector with the marginal frequencies of the pivotal columns, $\mathbf{R}_k = [f_{i\cdot k}]$ is a $(I-1) \times 1$ vector with the marginal frequencies of the pivotal rows, $\mathrm{diag}(\mathbf{X})$ is a diagonal matrix with elements $\mathbf{X}$, $\otimes$ indicates Kronecker product, and the apostrophe indicates transposition. Unpacking Eq. 2, the general element of $\mathbf{V}_k$ is

$$v_{k_{mn}} = \begin{cases} \dfrac{f_{i_1\cdot k}\left(N_k - f_{i_1\cdot k}\right) f_{\cdot j_1 k}\left(N_k - f_{\cdot j_1 k}\right)}{N_k N_k (N_k - 1)} & \text{if } i_1 = i_2 \wedge j_1 = j_2 \\ \dfrac{-f_{i_1\cdot k}\left(N_k - f_{i_1\cdot k}\right) f_{\cdot j_1 k} f_{\cdot j_2 k}}{N_k N_k (N_k - 1)} & \text{if } i_1 = i_2 \wedge j_1 \neq j_2 \\ \dfrac{-f_{i_1\cdot k} f_{i_2\cdot k} f_{\cdot j_1 k}\left(N_k - f_{\cdot j_1 k}\right)}{N_k N_k (N_k - 1)} & \text{if } i_1 \neq i_2 \wedge j_1 = j_2 \\ \dfrac{f_{i_1\cdot k} f_{i_2\cdot k} f_{\cdot j_1 k} f_{\cdot j_2 k}}{N_k N_k (N_k - 1)} & \text{if } i_1 \neq i_2 \wedge j_1 \neq j_2 \end{cases}. \tag{3}$$

Then,

$$Q_{\mathrm{GMH}} = \mathbf{D}\mathbf{V}^{-1}\mathbf{D}', \tag{4}$$

where $\mathbf{D} = \sum_{k=1}^{K} \mathbf{O}_k - \mathbf{E}_k$, $\mathbf{V} = \sum_{k=1}^{K} \mathbf{V}_k$, and $\mathbf{V}^{-1}$ is the inverse of $\mathbf{V}$. The $Q_{\mathrm{GMH}}$ statistic has asymptotically a $\chi^2$ distribution with $(I-1)(J-1)$ degrees of freedom. For the sample data in Fig. 1, $Q_{\mathrm{GMH}} = 0.146$ and the associated $p$ value from a $\chi^2$ distribution with 2 degrees of freedom is .930, so that equality of psychometric functions is not rejected at $\alpha = .05$.

Because all $K$ tables are aggregated into $\mathbf{D}$ and $\mathbf{V}$ before $Q_{GMH}$ is computed via Eq. 4, the statistic is well defined even when one or more of the individual tables has some empty columns or rows. Nevertheless, rows or columns that are empty in all $K$ tables should be discarded because they do not contribute data and degrees of freedom. It should also be noted that tables with $J - 1$ empty columns (or $I - 1$ empty rows) make no contribution to $Q_{GMH}$ because, for any such table, $\mathbf{O_k} - \mathbf{E_k} = \mathbf{0}$ and $\mathbf{V_k} = \mathbf{0}$, where $\mathbf{0}$ is a null matrix of appropriate size. This eventuality does not have any major consequence, except that the total sample size for the test reduces from the nominal $N = \sum_{k=1}^{K} N_k$ to the sum of $N_k$ across the remaining tables. If all but one of the $K$ tables has to be discarded for this reason, $Q_{GMH}$ degenerates to Pearson's statistic multiplied by $N_k/(N_k - 1)$ (Birch, 1965).

It should also be noted that tables for which $N_k = 1$ must also be discarded because then $\mathbf{V_k}$ is undefined: Note in Eq. 3 that computation of $v_{k_{mn}}$ requires division by $N_k - 1$. This will never occur for data collected with the method of constant stimuli but it may occur at some $k$ for data collected with adaptive methods. Removing such tables is clearly justifiable: $N_k = 1$ means that a single trial was placed at stimulus level $k$ in only one of the populations and, hence, there are actually no distributions to compare at this stimulus level. Data collected with the class of adaptive methods that place each trial at a unique stimulus level are likely to result in large $K$ with $N_k = 1$ for most $k$, which again precludes application of this test for lack of distributions to compare. However, this class of adaptive methods is unadvisable (and, actually, rarely used) to measure psychometric functions; dependable adaptive methods for measuring psychometric functions (see García-Pérez & Alcalá-Quintana, 2005; García-Pérez, 2014a) always place trials on a lattice so that $N_k = 1$ is a rare event.

It is important to stress that the appeal of $Q_{GMH}$ lies in the aggregation of deviations $\mathbf{O_k} - \mathbf{E_k}$ into $\mathbf{D}$ across tables. When psychometric functions are identical in all populations, the sign of these deviations will only vary randomly at each $j$ across the $K$ tables and their aggregation will result in $\mathbf{D} \approx \mathbf{0}$. On the other hand, when (monotonic) psychometric functions differ only in lateral position across populations (see Fig. 2a for an example with $I = 2$, $J = 2$, and $K = 13$), deviations will consistently have the same sign at each $j$ across the $K$ tables and aggregation will strengthen $\mathbf{D}$ by capitalizing on this systematic pattern of deviations. However, this strength turns into a serious weakness when monotonic psychometric functions differ only in slope across populations (see Fig. 2b) or when non-monotonic psychometric functions differ in lateral position across populations (see Fig. 2c). In both cases, systematic deviations of one sign will occur at stimulus levels below the crossing point of the psychometric functions, whereas systematic deviations of the opposite sign will occur above the crossing point. Then, aggregation across the $K$ tables annihilates these systematic patterns and renders $\mathbf{D} \approx \mathbf{0}$ as if psychometric functions did not differ across populations, a form of Simpson's paradox (Blyth, 1972; Simpson, 1951). The $Q_{GMH}$ statistic will be affected by these problems in a somewhat more complex manner when $J > 2$ (as in the example of Fig. 1) because such cases include a mixture of monotonic and non-monotonic psychometric functions in each population.

Simulation results presented below document the failure of the $Q_{GMH}$ statistic to detect differences such as those illustrated in Fig. 2b and c. A satisfactory solution to this problem is not immediately obvious but the two tests described next circumvent it in different ways.

## The generalized Berry–Mielke test of homogeneity

Deviations from homogeneity in each of the $K$ tables can be separately assessed via Pearson's statistic in Eq. 1. The significance of each of these individual deviations could be assessed with respect to the asymptotic $\chi^2$ distribution with $(I - 1)(J - 1)$ degrees of freedom, arguably with a Bonferroni correction for multiple testing. Alternatively, an omnibus test of homogeneity across the $K$ tables can be defined via

$$X^2 = \sum_{k=1}^{K} X_k^2, \tag{5}$$

which has an asymptotic $\chi^2$ distribution with $K(I - 1)(J - 1)$ degrees of freedom. However, Berry and Mielke (1988; see also Lewis, Saunders, & Westcott, 1984) showed that the small-sample significance of Pearson's statistic is more accurately assessed via a non-asymptotic Pearson Type III distribution (for an in-depth analysis of this superiority, see García-Pérez & Núñez-Antón, 2009). Unfortunately, the parameters of the Pearson Type III distribution depend on the marginal distributions in each of the $K$ tables and an omnibus test statistic different from that in Eq. 5 is needed. Such a generalized Berry–Mielke statistic is derived next.

Berry and Mielke (1988) showed that the adjusted Pearson's statistic

$$T_k = \frac{N_k - 1}{N_k} X_k^2 \tag{6}$$

has a conditional permutation distribution with exact mean $\mu_{T_k}$, exact variance $\sigma^2_{T_k}$, and exact skewness $\gamma_{T_k}$ (for the computation of these moments, see Berry & Mielke, 1988; Mielke & Berry, 1985; see also Supplementary Appendix A). Then, the standardized statistic

$$Z_k = \frac{T_k - \mu_{T_k}}{\sigma_{T_k}} \tag{7}$$

has approximately the Pearson Type III distribution

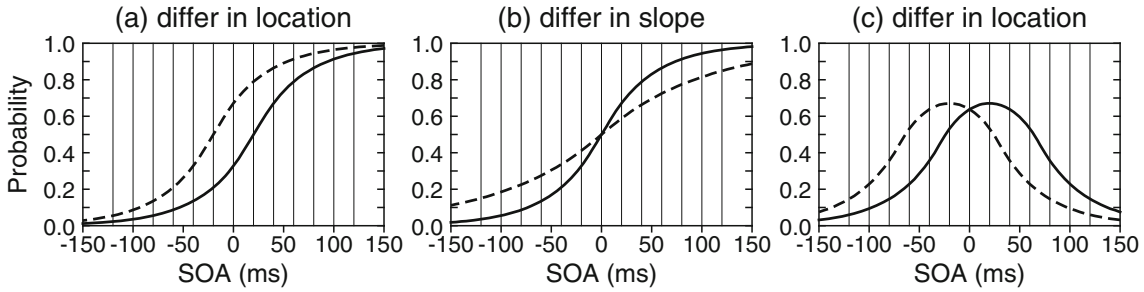



**Fig. 2** Hypothetical psychometric functions in two populations ($I = 2$; solid and dashed curves in each panel) probed at $K = 13$ stimulus levels (thin vertical lines in each panel) with a task that only allows two response categories ($J = 2$) so that each panel shows only the psychometric function for one of the categories in each population. (**a**) Monotonic functions that differ only in location across populations. (**b**) Monotonic functions that differ only in slope across populations. (**c**) Non-monotonic functions that differ only in location across populations. *SOA* stimulus-onset asynchrony

$$f(z) = \frac{(2/\gamma)^{4/\gamma^2}}{\Gamma(4/\gamma^2)}(z + 2/\gamma)^{4/\gamma^2-1}\exp\left[-(2z/\gamma + 4/\gamma^2)\right] \quad (8)$$

with $\gamma = \gamma_{T_k}$, where $z > -2/\gamma$ (Mielke & Berry, 1985) and $\Gamma$ denotes the gamma function. This is indeed the generalized gamma distribution

$$f(z) = \frac{d(z-a)^{cd-1}}{b^{cd}\Gamma(c)}\exp\left[-\left(\frac{z-a}{b}\right)^d\right] \quad (9)$$

with location parameter $a = -2/\gamma$, scale parameter $b = \gamma/2$, and shape parameters $c = 4/\gamma^2$ and $d = 1$, where $z > a$ (Forbes, Evans, Hastings, & Peacock, 2011, p. 113).

To arrive at an omnibus test, first define the standardized statistic

$$G_k = \frac{Z_k - a_k}{b_k}, \quad (10)$$

where $a_k = -2/\gamma_{T_k}$ and $b_k = \gamma_{T_k}/2$. Because $Z_k$ has a generalized gamma distribution (i.e., $Z_k \sim \mathrm{G}(a_k, b_k, c_k, 1)$, with $c_k = 4/\gamma^2_{T_k}$ ), $G_k$ is distributed G(0, 1, $c_k$, 1), which is the standard gamma distribution with scale parameter $b = 1$ and shape parameter $c = c_k$ (Forbes et al., 2011, p. 113). Then, the omnibus statistic

$$G = \sum_{k=1}^{K} G_k \quad (11)$$

has a standard gamma distribution with scale parameter $b = 1$ and shape parameter $c = \sum_{k=1}^{K} c_k$ (Forbes et al., 2011, p. 111). The null hypothesis of homogeneity is thus rejected when the $p$ value associated with the sample value of $G$ is lower than $\alpha$. Table 1 lists the necessary magnitudes computed at each $k$ for the data in Fig. 1, which yield $G = 7.237$ and $c = 10.448$. The associated $p$ value is .844 and the null is not rejected either by this test at $\alpha = .05$. But computation of the omnibus $G$ statistic also requires a number of precautions.

It should first be stressed that all the necessary computations must use the actual $I$ and $J$ at each $k$, which may differ from their nominal values. For instance, in the example of Fig. 1, the last column in the tables for stimulus levels 1 and 2 is filled with sampling zeros and, hence, these must be treated as 2 × 2 tables instead of 2 × 3 tables. Values reported in Table 1 were computed accordingly (e.g., note that $\mu_{T_k} = 1$ for $k \in \{1, 2\}$, whereas $\mu_{T_k} = 2$ for $k \in \{3, 4, 5, 6\}$). It should also be noted that sampling zeros may render a table with a single row or a single column at some $k$. This was not a problem for computation of the $Q_{\mathrm{GMH}}$ statistic but, in such cases, $G_k$ is undefined and the table must be discarded. (Note that a table for which $N_k = 1$ is also in the class of tables with a single row and a single column.)

There is another case in which the table at stimulus level $k$ must be discarded, namely, when $\sigma^2_{T_k} = 0$ so that $G_k$ is also undefined. This can only occur in $I \times 2$ tables in which all row marginal frequencies are equal and either $f_{\bullet 1k} = 1$ or $f_{\bullet 2k} = 1$ or, analogously, in $2 \times J$ tables in which all column marginal frequencies are equal and either $f_{1\bullet k} = 1$ or $f_{2\bullet k} = 1$. In either case, the permutation set consists of tables that differ only in how rows (or columns) are arranged, which does not alter the value of $T_k$. This situation may additionally occur when $J > 2$ if $J - 2$ columns are discarded due to sampling zeros (or, analogously, when $I > 2$ if $I - 2$ rows are discarded due to sampling zeros).

A final and less obvious precaution is that tables for which $\gamma_{T_k} < 0.5$ should also be discarded. The reason is an anomaly in the permutation distribution, whose description and analysis are somewhat technical, and is deferred to Supplementary Appendix B. Suffice it to say here that such small values of $\gamma_{T_k}$ arise only in $2 \times J$ (or $I \times 2$) tables with equal and small row (or column) marginal frequencies and uneven column (or row) marginal frequencies.

The distinctly different principles under which the omnibus $G$ statistic is derived demand a comparison with the $Q_{\mathrm{GMH}}$ statistic in terms of their expected performance. In principle, there is no reason to think that one or the other will be more accurate in terms of Type-I error rates. However, there are good reasons to think that they will perform differently in

**Table 1** Magnitudes required at each $k$ for the computation of the omnibus $G$ statistic for data in Fig. 1. Values at the bottom of the two rightmost columns are sums across $K$

| Level, $k$ | $T_k$ | $\mu_{T_k}$ | $\sigma_{T_k}$ | $\gamma_{T_k}$ | $Z_k$ | $G_k$ | $c_k$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.000 | 1.000 | 1.354 | 2.445 | −0.739 | 0.065 | 0.669 |
| 2 | 0.000 | 1.000 | 1.396 | 2.719 | −0.716 | 0.014 | 0.541 |
| 3 | 1.047 | 2.000 | 1.896 | 1.688 | −0.503 | 0.808 | 1.404 |
| 4 | 0.529 | 2.000 | 1.713 | 1.330 | −0.859 | 0.970 | 2.262 |
| 5 | 1.036 | 2.000 | 1.828 | 1.544 | −0.527 | 0.996 | 1.679 |
| 6 | 2.411 | 2.000 | 1.654 | 1.014 | 0.248 | 4.384 | 3.893 |
| | | | | | | 7.237 | 10.448 |

power studies. Because the omnibus $G$ statistic assesses deviation independently in each of the $K$ tables, it should be free of the problems that aggregation creates for the $Q_{GMH}$ statistic in cases such as those illustrated in Fig. 2b and c and, thus, it should have more power. At the same time, and because of that aggregation, the $Q_{GMH}$ statistic should have more power in cases such as those illustrated in Fig. 2a. Nevertheless, the true reality that generated the data is always unknown and, hence, it is practically impossible to decide in advance (i.e., before seeing the data) whether one or the other statistic will be more appropriate. We will come back to this issue in the section on "Practical recommendations" at the end of the paper.

### The split Mantel–Haenszel test of homogeneity

Cases such as those in Fig. 2b and c cause problems to the $Q_{GMH}$ statistic because the sign of the differences between psychometric functions varies across stimulus levels, and these signed differences cancel each other out upon aggregation into **D** across $K$. In the scenarios of Fig. 2b and c, it is obvious that the $Q_{GMH}$ statistic computed for data from only the lower half of stimulus levels will detect the differences (which always have the same sign) and that an analogous computation using only the upper half of stimulus levels will also detect the differences (which also have the same sign, although opposite to that in the lower half). This split computation renders two $Q_{GMH}$ statistics (say, $Q_{GMH}^{(L)}$ and $Q_{GMH}^{(U)}$, where superscripts denote the Lower and Upper sets of stimulus levels) each of which has an asymptotic $\chi^2$ distribution with $(I-1)(J-1)$ degrees of freedom. Then, the split statistic

$$S\text{-}Q_{GMH} = Q_{GMH}^{(L)} + Q_{GMH}^{(U)} \tag{12}$$

has an asymptotic $\chi^2$ distribution with $2(I-1)(J-1)$ degrees of freedom. Application of this statistic to the data in Fig. 1 with an even split (i.e., stimulus levels 1–3 contribute to $Q_{GMH}^{(L)}$ and stimulus levels 4–6 contribute to $Q_{GMH}^{(U)}$) renders $Q_{GMH}^{(L)} = 0.4981$ and $Q_{GMH}^{(U)} = 1.4163$ so that $S\text{-}Q_{GMH} = 1.9145$. The $p$ value is .752 and the null is not rejected at $\alpha = .05$.

In principle, the $S\text{-}Q_{GMH}$ statistic should be as accurate as the overall $Q_{GMH}$ statistic and it should only be slightly less powerful than the overall $Q_{GMH}$ statistic when the latter performs well (e.g., in cases such as that illustrated in Fig. 2a). However, in the scenarios of Fig. 2b and c, $S\text{-}Q_{GMH}$ should be meaningfully more powerful than $Q_{GMH}$.

It should also be noted that the scenarios in Fig. 2b and c represent ideal cases in which the $S\text{-}Q_{GMH}$ statistic will clearly outperform the $Q_{GMH}$ statistic. The reason is that the stimulus levels at which the psychometric functions are probed (thin vertical lines in the panels) are placed symmetrically about the crossing point of the psychometric functions. In experimental practice, data are collected at stimulus levels placed without knowledge of the slope and location of the psychometric functions. Psychometric functions may also differ across populations in all respects (i.e., location, slope, and symmetry, not just the single-aspect difference illustrated in Fig. 2). In general, the optimal split is not in two equal halves; the optimal split is instead that which separates regions where differences between psychometric functions have opposite signs (i.e., at the point where the data suggest that the psychometric functions cross). The power of the $S\text{-}Q_{GMH}$ statistic will be obviously reduced for suboptimal splits but an adequate split can always be judged by eye in empirical applications.

Splitting the computation into these two components may have a consequence that is worth commenting on. Because $Q_{GMH}^{(L)}$ uses data at the lower stimulus levels and $Q_{GMH}^{(U)}$ uses data at the higher stimulus levels, columns filled with sampling zeros are more likely to occur in only one of these two components (something that would have happened for data in Fig. 1 if the last column of the table for stimulus level 3 had also been filled with zeros). This eventuality depends on the number of stimulus levels in each component, their location relative to the psychometric functions, the number $I$ of populations, and the number $J$ of response categories. Its only implication

is that removal of the incumbent columns in one or the other component will alter the degrees of freedom of $Q_{GMH}^{(L)}$ or $Q_{GMH}^{(U)}$, something that should be kept in mind. These considerations also hold when it is instead (or additionally) the rows that are filled with sampling zeros (e.g., when data are collected with adaptive methods that place stimulus levels within different regions of a lattice for different populations).

## Simulation method

The accuracy and power of the three tests were assessed in a series of simulations involving several numbers of populations ($I \in \{2, 3, 4\}$), response categories ($J \in \{2, 3, 4\}$), stimulus levels ($K \in \{7, 13\}$), and row marginal frequencies $f_{i\bullet k}$ (from 5 to 50 in steps of 5). These combinations cover the situations in the vast majority of empirical studies. Each simulation condition generated 30,000 replicates (i.e., sets of $I \times J \times K$ tables such as those in Fig. 1) for which all statistics ($Q_{GMH}$, $S$-$Q_{GMH}$, and $G$) were computed and their $p$ values determined. The $S$-$Q_{GMH}$ statistic was computed with an almost-even split (i.e., lower 3 vs. upper 4 stimulus levels when $K = 7$ and lower 6 vs. upper 7 when $K = 13$) that is optimal or near-optimal under the simulated conditions. In conditions with identical psychometric functions in the $I$ populations, accuracy at $\alpha \in \{.01, .05\}$ was assessed via the proportion of times that the (true) null was rejected; in conditions with different psychometric functions across populations, power was assessed via the proportion of times that the (false) null was rejected.

Without loss of generality, data were generated with psychometric functions from a model of performance in timing tasks that allow from $J = 2$ to $J = 4$ response categories (García-Pérez & Alcalá-Quintana, 2012a, 2012b, 2015a, 2015b, 2015c, 2017a). A brief description of these tasks and the model is useful for later references. In timing tasks, two stimuli (A and B) are presented with some stimulus-onset asynchrony (SOA) that varies across trials within a set of $K$ levels. In the binary *temporal-order judgment* (TOJ) task, observers must report whether A or B was subjectively presented first and, hence, $J = 2$; in the *ternary synchrony judgment* (SJ3) task, observers are additionally allowed to report that both stimuli seemed to be presented simultaneously and, hence, $J = 3$; in the *4-ary synchrony judgment* (SJ4) task, observers are additionally allowed to report that they cannot tell temporal order even though presentations appeared to be non-simultaneous and, hence, $J = 4$. Under the model, the psychometric functions describing how the probabilities of "A first" (AF) and "B first" (BF) responses vary with SOA ($\Delta t$) in the TOJ task are

$$\begin{cases} \Psi_{AF}^{(TOJ)}(\Delta t) = 1-(1-\xi)F(\delta_4;\Delta t)-\xi F(\delta_1;\Delta t) \\ \Psi_{BF}^{(TOJ)}(\Delta t) = (1-\xi)F(\delta_4;\Delta t) + \xi F(\delta_1;\Delta t) \end{cases} \quad (13)$$

where

$$F(d;\Delta t) = \begin{cases} \dfrac{\lambda_A}{\lambda_A+\lambda_B}\exp[\lambda_A(d-\Delta t-\tau)] & \text{if } d \le \Delta t + \tau \\ 1-\dfrac{\lambda_B}{\lambda_A+\lambda_B}\exp[-\lambda_B(d-\Delta t-\tau)] & \text{if } d > \Delta t + \tau \end{cases} \quad (14)$$

and $\lambda_A$, $\lambda_B$, $\tau$, $\delta_1$, $\delta_4$, and $\xi$ are model parameters described below. In the SJ3 task, the probabilities of AF, BF, and "simultaneous" (S) responses vary with SOA as

$$\begin{cases} \Psi_{AF}^{(SJ3)}(\Delta t) = 1 - F(\delta_4;\Delta t) \\ \Psi_{S}^{(SJ3)}(\Delta t) = F(\delta_4;\Delta t)-F(\delta_1;\Delta t) \\ \Psi_{BF}^{(SJ3)}(\Delta t) = F(\delta_1;\Delta t) \end{cases} \quad (15)$$

and note that $\xi$ is not a parameter in the model for this task. Finally, in the SJ4 task, the probabilities of AF, BF, S, and "unknown order" (U) responses vary with SOA as

$$\begin{cases} \Psi_{AF}^{(SJ4)}(\Delta t) = 1 - F(\delta_4;\Delta t) \\ \Psi_{S}^{(SJ4)}(\Delta t) = F(\delta_3;\Delta t) - F(\delta_2;\Delta t) \\ \Psi_{BF}^{(SJ4)}(\Delta t) = F(\delta_1;\Delta t) \\ \Psi_{U}^{(SJ4)}(\Delta t) = F(\delta_4;\Delta t)-F(\delta_3;\Delta t) + F(\delta_2;\Delta t)-F(\delta_1;\Delta t) \end{cases} \quad (16)$$

where $\delta_2$ and $\delta_3$ are additional model parameters for this task.

The parameters just mentioned describe a process model and they affect the shape, location, and symmetry of the resultant psychometric functions in ways that will be described when we assess power as a function of differences in parameter values. The model posits that an observer's judgment and the subsequent response is based on the difference between the arrival times of sensory signals from stimuli A and B at a central mechanism. Arrival times are assumed to have shifted exponential distributions with rates $\lambda_A$ and $\lambda_B$ and delays $\tau_A$ and $\tau_B$, respectively, for stimuli A and B. The difference in arrival times is the decision variable and has a bilateral exponential distribution whose cumulative distribution function is given by Eq. 14, where $\tau = \tau_B - \tau_A$. The decision rule partitions the domain of the decision variable into five regions with boundaries at $\delta_1$, $\delta_2$, $\delta_3$, and $\delta_4$. Then, BF judgments are associated with the range $(-\infty, \delta_1)$, AF judgments are associated with the range $(\delta_4, \infty)$, S judgments are associated with the range $(\delta_2, \delta_3)$, and U judgments are associated with the ranges $(\delta_1, \delta_2)$ and $(\delta_3, \delta_4)$. In the absence of response errors (which are not considered here), these judgments are directly expressed as the

corresponding responses in the SJ4 task, rendering the psychometric functions in Eq. 16 (for empirical evidence to this effect, see García-Pérez & Alcalá-Quintana, 2017a). In SJ3 tasks where U responses are not allowed, Eq. 15 portrays that U judgments are reported as S responses, with no implicit claim that this yet untested assumption is empirically tenable. Finally, in TOJ tasks where only AF or BF responses are allowed, Eq. 13 implies that U and S judgments yield random guesses whereby AF responses occur with probability $\xi$ and BF responses occur with probability $1 - \xi$ (for empirical evidence to this effect, see García-Pérez & Alcalá-Quintana, 2012a, 2015a, b). Parameter values used in our simulations are realistic and within the broad range observed in the empirical studies that tested the model.

Artificial data displayed in the illustration of Fig. 1 were generated for a condition with $J = 3$ (i.e., the SJ3 task), $K = 6$ (i.e., six SOAs), $f_{i\bullet k} = 20$ for all $i$ and $k$ (i.e., 20 trials per SOA), and equal psychometric functions in $I = 2$ populations (i.e., true null) with parameters ($\lambda_A$, $\lambda_B$, $\tau$, $\delta_1$, $\delta_4$) = (1/45, 1/45, 0, −50, 50). Along with $\xi = 0.5$ (needed when $J = 2$ in the TOJ task) and ($\delta_2$, $\delta_3$) = (−40, 40) (needed when $J = 4$ in the SJ4 task), these are the parameter values used without loss of generality in simulations when the null is true. The top row in Fig. 3 below shows the resultant psychometric functions for $J \in \{2, 3, 4\}$. In simulations when the null is false, parameter values differed across populations as described later.

The $K$ stimulus levels were always placed so as to cover either a broad or a narrow region that sampled the underlying psychometric functions centrally (see the thin vertical lines within each panel in the top row of Figs. 3, 4, and 5 below). Stimulus levels that are too extreme were avoided because data collected at them are uninformative: Even when psychometric functions differ, they have the same upper- and lower-asymptotic regimes and they come together at extreme stimulus levels. Tables at such levels are thus likely to contain $J - 1$ empty columns and, as discussed earlier, they make no contribution to the $Q_{GMH}$ or $S$-$Q_{GMH}$ statistics and they have to be discarded for computation of the $G$ statistic. Extreme stimulus levels are also avoided in empirical studies that measure psychometric functions to assess differences, although they may be used for other purposes (e.g., to estimate lapse rates).

## Results

Simulation results are presented next for each of the three tests. Complete results for the generalized Mantel–Haenszel test are presented first. This set of results is used as a reference for comparison with those for the two other tests, which are presented more succinctly.

### Empirical Type-I error rate and power of the generalized Mantel–Haenszel test

The top row of Fig. 3 shows the psychometric functions used to generate data for $J \in \{2, 3, 4\}$ and the thin vertical lines indicate the $K = 13$ stimulus levels at which data were generated in one of the conditions for analyses of accuracy, where psychometric functions are identical in all $I$ populations. The rows underneath display, for $I \in \{2, 3, 4\}$, the empirical Type-I error rates at nominal $\alpha \in \{.01, .05\}$ as a function of number of trials per level (i.e., the size of the row marginal frequencies $f_{i\bullet k}$, identical for all $i$ and $k$). By Bradley's (1978) stringent criterion of accuracy (namely, that empirical Type-I error rates be within 10% of the nominal rates), the test is remarkably accurate in all cases. Accuracy is analogous for data collected at the $K = 7$ central stimulus levels (see Fig. 4) or when the $K = 13$ stimulus levels are packed in the central region of the psychometric functions (see Fig. 5).

Two additional simulations were run in which the row marginal frequencies $f_{i\bullet k}$ varied across populations or across stimulus levels. In both cases, $f_{1\bullet 1}$ always spanned the range of values used in the preceding simulations (i.e., from 5 to 50 in steps of 5). In one of these additional simulations, row marginal frequencies for $i > 1$ increased as $f_{i\bullet k} = 2^i f_{1\bullet 1}$ for all $k$; in the other simulation, row marginal frequencies remained identical across populations but increased across stimulus levels (i.e., for $k > 1$) as $f_{i\bullet k} = f_{1\bullet 1} + 10(k - 1)$ for all $i$. A graphic presentation of these results is omitted but differences in the size of row marginal frequencies across populations or stimulus levels did not alter the accuracy documented in Figs. 3, 4, and 5.

In sum, the generalized Mantel–Haenszel test of equality of psychometric functions is very accurate even when the number of trials per stimulus level is small (e.g., when $f_{i\bullet k} = 5$ for all $i$ and $k$). A second aspect that is also relevant empirically is the capability of the test to reject a false null (i.e., its power), something that is unlikely to occur when $f_{i\bullet k}$ is small.

The power of a statistical test is usually expressed as a function of effect size, a suitable measure of the relevant difference between (two) populations. Effect size is well defined for parameters such as means, variances, proportions, or correlations (see, e.g., Cohen, 1992; Faul, Erdfelder, Lang, & Buchner, 2007; Faul, Erdfelder, Buchner, & Lang, 2009; Fritz, Morris, & Richler, 2012; Lakens, 2013). However, differences between distributions may occur in a variety of forms that do not lend themselves to quantification along a one-dimensional metric. This is surely the reason why the power of tests of homogeneity (and, more generally, tests of the chi-square type) is often computed against arbitrarily defined alternatives, sometimes characterized by the largest point difference (see, e.g., Agresti, 1983; Cressie & Read, 1984; Eubank, 1997; Nisen & Schwertman, 2008; Pardo, 1998;

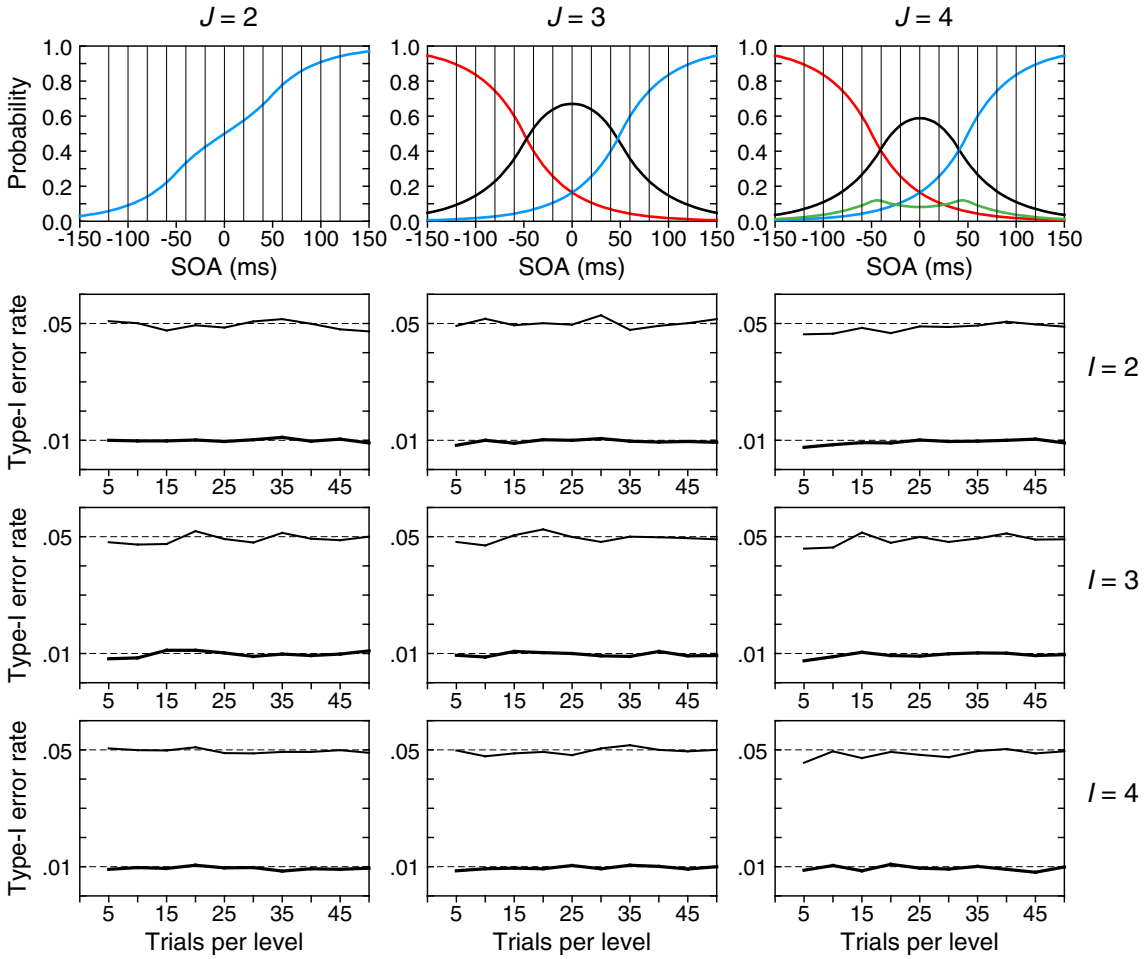



**Fig. 3** Accuracy of the generalized Mantel–Haenszel test of equality of psychometric functions at nominal $\alpha \in \{.05, .01\}$ as a function of trials per level (i.e., the size of row marginal frequencies $f_{i\bullet k}$, identical for all $i$ and $k$). The $Q_{GMH}$ statistic was computed from $K = 13$ tables sampling the psychometric functions at locations indicated by the thin vertical lines in the top row, which shows also the shape of the psychometric functions used to generate the data for the combinations of $I$ and $J$ indicated at the top of each column and at the right of each row. *SOA* stimulus-onset asynchrony

Read, 1984). This approach is unsatisfactory in the present context because the largest point difference is not a reasonable measure of differences between psychometric functions. Then, power is documented in two ways: Firstly, as a function of number of trials (defined as the size of $f_{i\bullet k}$) for fixed differences in selected parameters of the psychometric functions; secondly, as a function of the magnitude of the differences in selected parameters for a fixed number of trials. In all analyses, $K \in \{7, 13\}$ with stimulus levels placed as in Figs. 3, 4, and 5.

The center and bottom rows of Fig. 6 show power functions for each set of $K$ levels (black, green, and magenta curves) as a function of number of trials when two ($I = 2$) or three ($I = 3$) psychometric functions differ only by translation as shown in the top row. In terms of the model parameters (see Eqs. 13–16 above), the only difference was that $\tau = 0$ in population 1 (solid curves in the top row of Fig. 6), $\tau = 20$ in population 2 (dashed curves), and $\tau = -20$ in population 3 (dotted curves). Naturally, power varies with the number and location of the $K$ stimulus levels. Comparatively, power functions are lowest when $K = 7$ stimulus levels are placed in the central region (see the green sampling points above the top panels), they increase when additional stimulus levels (up to $K = 13$) are placed further out on each side (black sampling points above the top panels), and increase further when the $K = 13$ stimulus levels are packed in the central region where psychometric functions differ most (magenta sampling points above the top panels).

Also naturally, power increases as a sigmoidal function of sample size for all $I$, $J$, and $K$, although there are clear differences across conditions. Power is lowest when $I = J = 2$, understandably because psychometric functions are not very different from one another in this simulation condition; yet, when $J > 2$, power increases because additional response categories allow for stronger empirical manifestation of the differences. This is also the reason why power is higher when $I = 3$ (bottom row in Fig. 6) than it is under analogous conditions when $I = 2$ (center row), given that the third

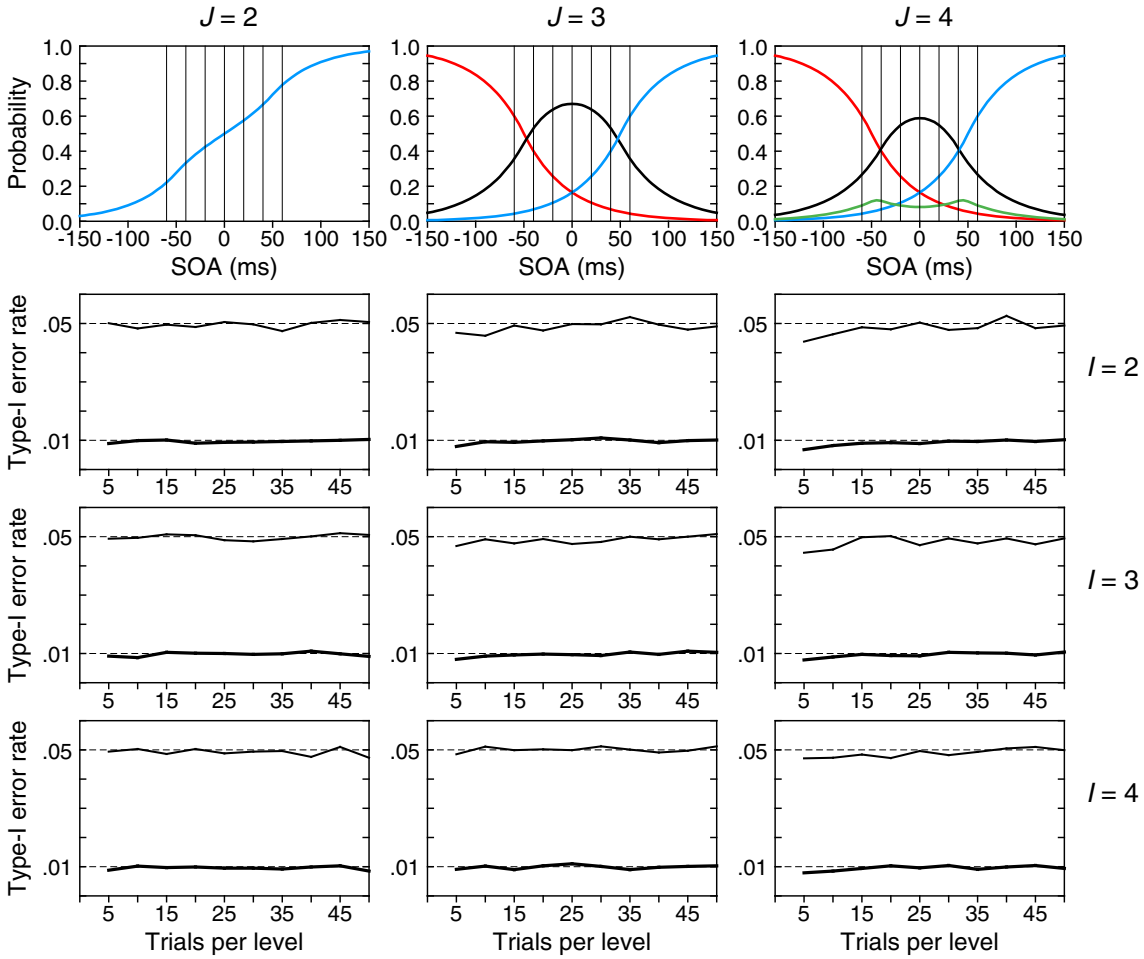



**Fig. 4** Accuracy of the generalized Mantel–Haenszel test of equality of psychometric functions at nominal $\alpha \in \{.05, .01\}$ as a function of trials per level (i.e., the size of row marginal frequencies $f_{i\bullet k}$, identical for all $i$ and $k$). The $Q_{GMH}$ statistic was computed from $K = 7$ tables sampling the psychometric functions at locations indicated by the thin vertical lines in the top row, which shows also the shape of the psychometric functions used to generate the data for the combinations of $I$ and $J$ indicated at the top of each column and at the right of each row. In comparison to Fig. 3, the only difference is that data were not generated for any of the three outer stimulus levels on each side. *SOA* stimulus-onset asynchrony

population in these simulations expands the range of differences across populations.

Perhaps surprisingly, power is higher when $J = 3$ (center column in Fig. 6) than it is under analogous conditions when $J = 4$ (right column), despite the fact that an additional response category should provide further manifestation of the differences across populations. This counterintuitive result relates to the problem discussed earlier regarding systematic differences that change sign across stimulus levels. The case $J = 3$ implies three psychometric functions only two of which are monotonic (red and blue curves in the top-center panel of Fig. 6). For these two monotonic functions, changes in location across populations produce consistent differences in sign across stimulus levels (as illustrated in Fig. 2a above), which contribute to increasing the significance of the $Q_{GMH}$ statistic. In the remaining non-monotonic psychometric function (black curves in the top-center panel of Fig. 6) location changes result in a reversal of the sign of the differences (as illustrated in Fig. 2c above), which enter the $Q_{GMH}$ statistic as misleading "evidence" of lack of differences. When $J = 4$, the fourth psychometric function is also non-monotonic (green curves in the top-right panel of Fig. 6) and the associated sign reversals contribute further "evidence" that reduces power. Our next analysis makes this problem more clearly apparent.

Figure 7 shows analogous results when psychometric functions differ in slope. For these analyses, psychometric functions differed in that $\lambda_A = \lambda_B = 1/45$ in population 1 (solid curves in the top row of Fig. 7), $\lambda_A = \lambda_B = 1/22.5$ in population 2 (dashed curves), and $\lambda_A = \lambda_B = 1/90$ in population 3 (dotted curves). Compared to results in Fig. 6, power to detect changes in slope is remarkably smaller, with an absolute lack of power when $J = 2$ (left column of Fig. 7): Rejection rates stay at the Type-I error rate as if psychometric functions did not differ across populations. It is true that the psychometric functions differ only slightly when $J = 2$ (and only at the outer stimulus levels in the conditions of Fig. 7),

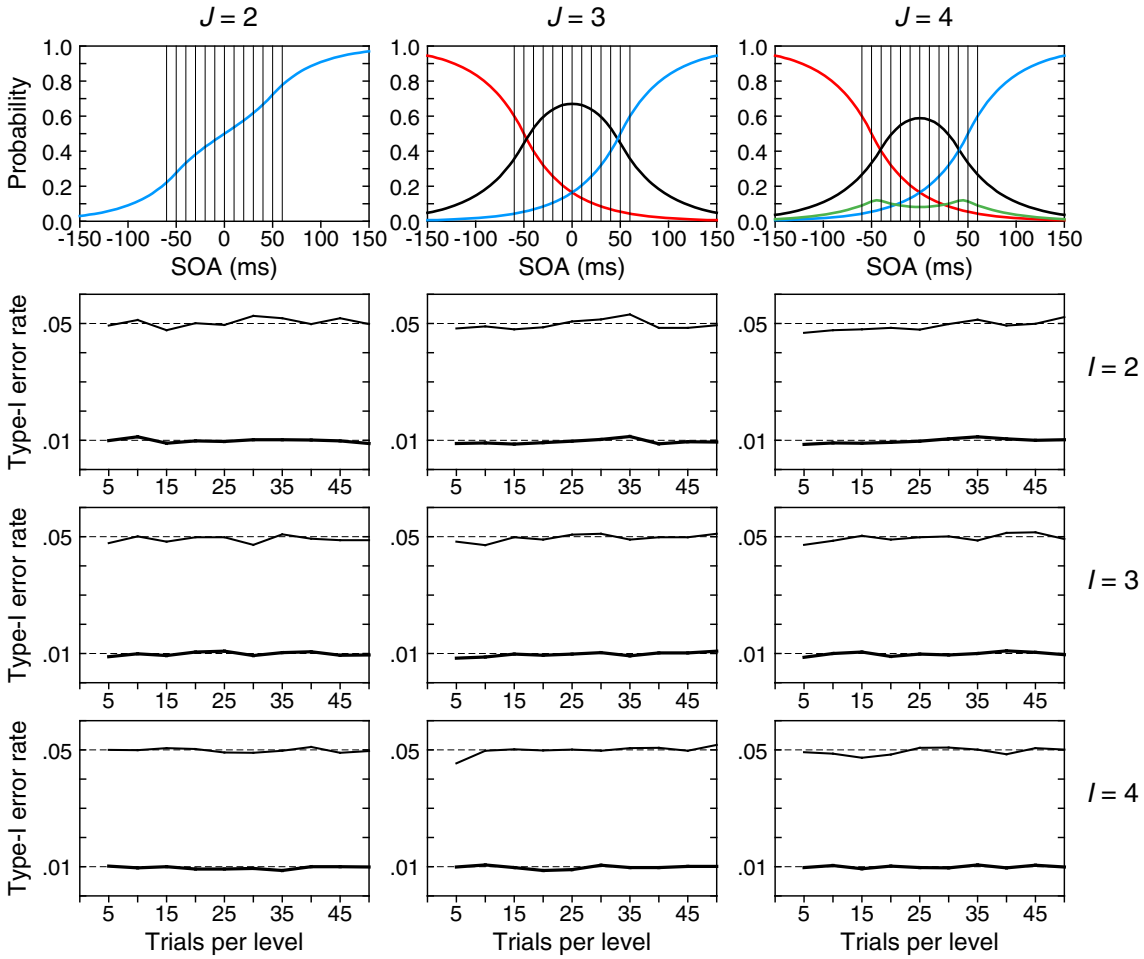



**Fig. 5** Accuracy of the generalized Mantel–Haenszel test of equality of psychometric functions at nominal $\alpha \in \{.05, .01\}$ as a function of trials per level (i.e., the size of row marginal frequencies $f_{i\bullet k}$, identical for all $i$ and $k$). The $Q_{GMH}$ statistic was computed from $K = 13$ tables sampling the psychometric functions at locations indicated by the thin vertical lines in the top row, which shows also the shape of the psychometric functions used to generate the data for the combinations of $I$ and $J$ indicated at the top of each column and at the right of each row. In comparison to Fig. 4, the only difference is that data were generated at additional stimulus levels in between each pair of consecutive levels used in Fig. 4. *SOA* stimulus-onset asynchrony

but additional simulations rendered analogous results when psychometric functions differed more (see, e.g., Fig. 10 below). The lack of power is actually caused by the fact that aggregation into **D** across $K$ averages out the opposite sign of the differences on either side of the crossing point of the psychometric functions. This is also responsible for the shallow power functions when $J > 2$ (center and right columns of Fig. 7) and the strong dependence of power on the choice of stimulus levels. Specifically, $K = 13$ stimulus levels at the locations indicated by the black circles above the top panels span the region where sign changes occur across populations, resulting in very low power particularly when $I = 2$ (black power functions). In contrast, use of $K = 7$ stimulus levels within the central region where sign changes do not occur (green circles above the top panels) substantially increases power (green power functions), and probing psychometric functions within the same region but more densely with $K = 13$ stimulus levels (magenta circles above the top panels) increases power further (magenta power functions). Although these results are easily understandable on these grounds, they raise the practical issue of how can appropriate sampling points be planned in advance without knowledge of the differences that psychometric functions may actually have across populations. One way around this unsolvable problem is to use one of the alternative tests examined later, provided they prove invulnerable to this or other threats.

Figure 8 shows results when psychometric functions differ in symmetry. Specifically, $\lambda_A = \lambda_B = 1/45$ in population 1 (solid curves in the top row of Fig. 8), $\lambda_A = 1/67.5$ and $\lambda_B = 1/30$ in population 2 (dashed curves), and $\lambda_A = 1/30$ and $\lambda_B = 1/67.5$ in population 3 (dotted curves). Note, however, that changes in symmetry also produce changes in location, as is apparent in the left panel in the top row of Fig. 8. For this reason, results are analogous to those reported in Fig. 6 above, with power being lowest when $I = J = 2$ and higher when $I = 3$

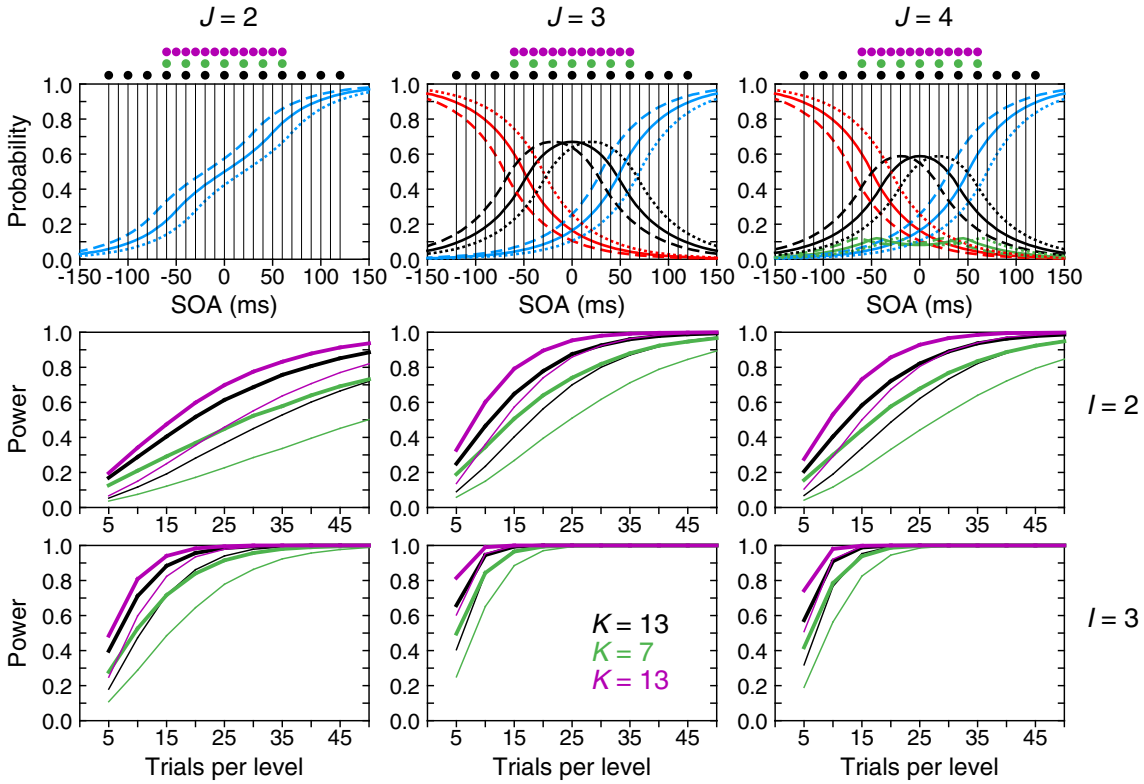



**Fig. 6** Power of the generalized Mantel–Haenszel test of equality of psychometric functions as a function of number of trials per level (i.e., the size of the row marginal frequencies $f_{i\bullet k}$, identical for all $i$ and $k$) when psychometric functions differ only by lateral translation across populations (see the top row) and data are collected with tasks that allow for $J \in \{2, 3, 4\}$ response categories (columns). In each case, data were gathered at the same sets of $K$ stimulus levels used to assess accuracy in Figs. 3–5. These sampling points are indicated by the strings of colored circles immediately above the top panels. The parameters of the psychometric functions were identical in all populations except that $\tau = 0$ in population 1 (solid curves in the top row), $\tau = 20$ in population 2 (dashed curves in the top row), and $\tau = -20$ in population 3 (dotted curves in the top row). Results for $I = 2$ (center row) involve populations 1 and 2, whereas results for $I = 3$ (bottom row) involve the three populations. Power functions are plotted at $\alpha = .05$ (thick curves) and $\alpha = .01$ (thin curves), using the color that corresponds to the sampling points used to obtain it. *SOA* stimulus-onset asynchrony

than it is when $I = 2$. Power also varies according to how different the psychometric functions are at the $K$ locations where data are collected, and power is also higher for $J = 3$ than it is for $J = 4$.

Results presented thus far describe how power increases with number of trials per stimulus level for arbitrary differences across populations. Results presented next describe how power increases with effect size (loosely defined as parametric differences between psychometric functions) at the fixed sample size $f_{i\bullet k} = 20$ for all $i$ and $k$.

Given the mathematical form and parameters of our psychometric functions (Eqs. 13–16), parameter $\tau$ determines their location so that psychometric functions that only differ in this parameter have the exact same shapes and differ only by lateral translation. A measure of effect size that captures such variation is the difference $\Delta\tau$ between $\tau$ parameter values, quantified by the largest paired difference among the $I$ populations involved. In the present simulations as well as in the coming simulations that assess power, effect size will be kept constant across variations in the number $I$ of populations by making populations 1 and 2 differ the most in the value of the parameter under study. (In contrast, effect size by this measure varied with $I$ in the preceding simulations because the applicable parameter of population 3 increased the largest difference.) The top row of Fig. 9 shows psychometric functions that differ in location for $J \in \{2, 3, 4\}$. Solid curves have the same parameter values used earlier and depict psychometric functions for population 1 in this analysis. Dashed curves differ only in that $\tau$ has a value such that $\Delta\tau = 60$, the largest effect size for which power is reported in the center and bottom rows of Fig. 9; intermediate effect sizes produce proportionally smaller lateral translations. These (families of) dashed curves are the psychometric functions used for population 2 in this analysis; when $I = 3$, parameter $\tau$ for population 3 had a value midway between those for populations 1 and 2.

The center and bottom rows of Fig. 9 show power as a function of $\Delta\tau$ for each combination of $I$ and $J$ when $K \in \{7, 13\}$ at the sampling points used earlier. Clearly, power increases as a sigmoidal function of $\Delta\tau$. It is also apparent that power is slightly higher when $I = 2$ (center row) than it is when $I = 3$ (bottom row) under otherwise identical conditions.

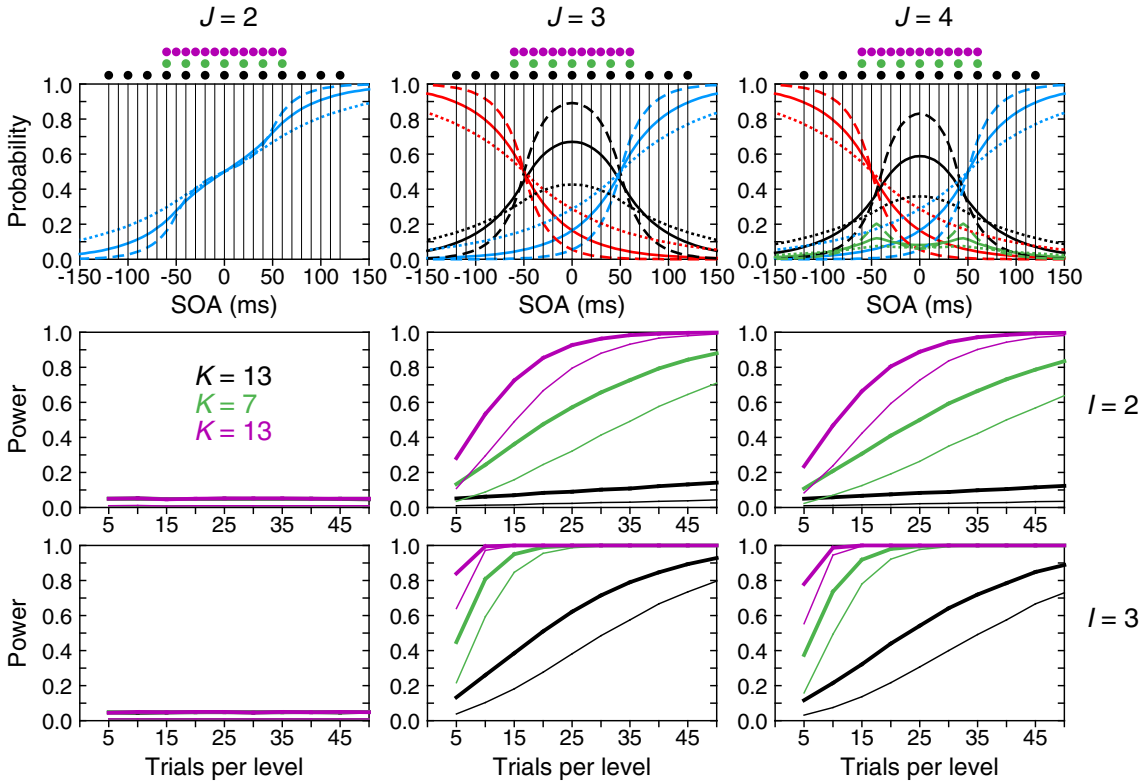



**Fig. 7** Power of the generalized Mantel–Haenszel test of equality of psychometric functions as a function of number of trials per level (i.e., the size of the row marginal frequencies $f_{i\bullet k}$, identical for all $i$ and $k$) when psychometric functions differ only in slope across populations (see the top row). Layout and graphic conventions as in Fig. 6. The parameters of the psychometric functions were identical in all populations except that $\lambda_A = \lambda_B = 1/45$ in population 1 (solid curves in the top row), $\lambda_A = \lambda_B = 1/22.5$ in population 2 (dashed curves in the top row), and $\lambda_A = \lambda_B = 1/90$ in population 3 (dotted curves in the top row). Results for $I = 2$ (center row) involve populations 1 and 2, whereas results for $I = 3$ (bottom row) involve the three populations. *SOA* stimulus-onset asynchrony

This is understandable on consideration that $\Delta\tau$ only indicates the largest parametric difference among populations so that the case $I = 2$ involves larger overall differences than the case $I = 3$ (where a third set of psychometric functions is placed midway between the two other sets). On the other hand, power also varies with the number and location of the $K$ stimulus levels at which data are collected, again understandably because differences in the ordinates of the psychometric functions vary across stimulus levels.

Note that these results relate to those displayed in Fig. 6, where power was reported as a function of sample size when $\Delta\tau = 20$ (for $I = 2$) and $\Delta\tau = 40$ (for $I = 3$). The ordinates of the power functions in Fig. 9 at these effect sizes match those in the corresponding panels of Fig. 6 at $f_{i\bullet k} = 20$, with only minor differences due to the different relative locations of the $K$ sampling points with respect to the psychometric functions.

It is also interesting to compare the power functions in Fig. 9 for $I = J = 2$ with power functions reported by Logvinenko et al. (2012) for their test in analogous conditions (see their Figs. 6 and 7). Power functions for the generalized Mantel–Haenszel test asymptote to unity as effect size increases (left panel in the center row of Fig. 9), whereas comparable power functions in Figs. 6 and 7 of Logvinenko et al. asymptote instead below unity.

Figure 10 shows analogous results for psychometric functions that differ in slope as shown in the top row for the largest difference used in our analyses. These differences are determined by parameters $\lambda_A$ and $\lambda_B$, which were kept equal in value to preserve symmetry. Effect size is now defined via $\Delta\mu$, the difference between $1/\lambda_A$ (or $1/\lambda_B$) in populations 1 and 2, rendering the differences shown in the top row of Fig. 10 when $\Delta\mu = 90$; the third population used when $I = 3$ had a slope midway between the other two. In line with results presented in Fig. 7, power is null when $J = 2$ (left column in Fig. 10) due to the changing sign of the differences between psychometric functions on either side of the point at which they cross. This feature of the psychometric functions also have consequences when $J > 2$ (center and right columns in Fig. 10), although the diverse power functions in those cases reflect an interaction with the location of the $K$ stimulus levels at which the psychometric functions are probed. Reasonably large power is obtained with the magenta sampling points, which probe only the central region where sign differences do not occur. As seen in the top-center panel of Fig. 10 (for $J = 3$), within that range of

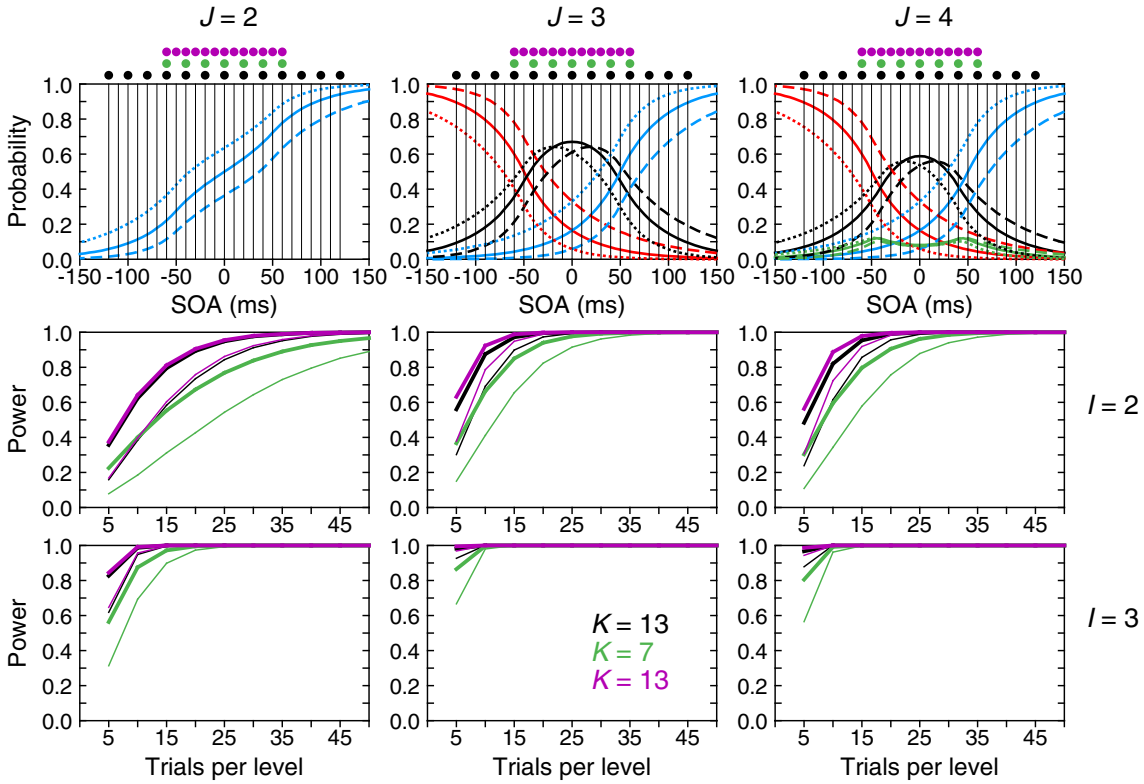



**Fig. 8** Power of the generalized Mantel–Haenszel test of equality of psychometric functions as a function of number of trials per level (i.e., the size of the row marginal frequencies $f_{i\bullet k}$, identical for all $i$ and $k$) when psychometric functions differ only in symmetry across populations (see the top row). Layout and graphic conventions as in Fig. 6. The parameters of the psychometric functions were identical in all populations except that $\lambda_A = \lambda_B = 1/45$ in population 1 (solid curves in the top row), $\lambda_A = 1/67.5$ and $\lambda_B = 1/30$ in population 2 (dashed curves in the top row), and $\lambda_A = 1/30$ and $\lambda_B = 1/67.5$ in population 3 (dotted curves in the top row). Results for $I = 2$ (center row) involve populations 1 and 2, whereas results for $I = 3$ (bottom row) involve the three populations. *SOA* stimulus-onset asynchrony

stimulus levels the solid red curve is (almost) always below the dashed red curve, the solid black curve is always above the dashed black curve, and the solid blue curve is (almost) always below the dashed blue curve. Power is lower when the same region is probed with fewer stimulus levels (green sampling points and green power functions), and power is substantially reduced when stimulus levels extend into the region where differences between psychometric functions change sign (black sampling points and black power functions). It is also clear that power is higher when $I = 2$ than it is when $I = 3$ under identical conditions, again because the psychometric function for the third population in these simulations was midway between the other two.

Results in Fig. 10 also relate to those displayed in Fig. 7, where power was reported as a function of sample size when $\Delta\mu = 22.5$ (for $I = 2$) and $\Delta\mu = 45$ (for $I = 3$), although slope in the present analyses varies only in one direction (increasingly shallower), whereas it varied in both directions (shallower and steeper) in Fig. 7. Thus, the ordinates of the power functions in Fig. 10 at comparable effect sizes are only close to those in the corresponding panels of Fig. 7 at $f_{i\bullet k} = 20$, also due to differences in the relative location of the $K$ sampling points.

To summarize, the generalized Mantel–Haenszel test for equality of psychometric functions is accurate but powerless to detect differences that result in psychometric functions that cross and change the sign of their differences across the selected stimulus levels.

## Empirical Type-I error rate and power of the generalized Berry–Mielke test

Accuracy analyses of the generalized Berry–Mielke test revealed the exact same patterns reported in Figs. 3, 4, and 5 for the generalized Mantel–Haenszel test and graphic presentation of these results is omitted. Neither of these tests is thus superior in terms of accuracy, but they differed non-uniformly as to power, as discussed next.

Figure 11 shows power functions for the generalized Berry–Mielke test in the conditions of Fig. 9, namely, as a function differences in location of the psychometric functions when $f_{i\bullet k} = 20$ for all $i$ and $k$. As surmised, the generalized Berry–Mielke test does not parallel the generalized Mantel–Haenszel test in these cases, but the drop in power is relatively small. A comparison of power functions in the left panel in the center row of Fig. 11 (for $I = J = 2$) with power functions in

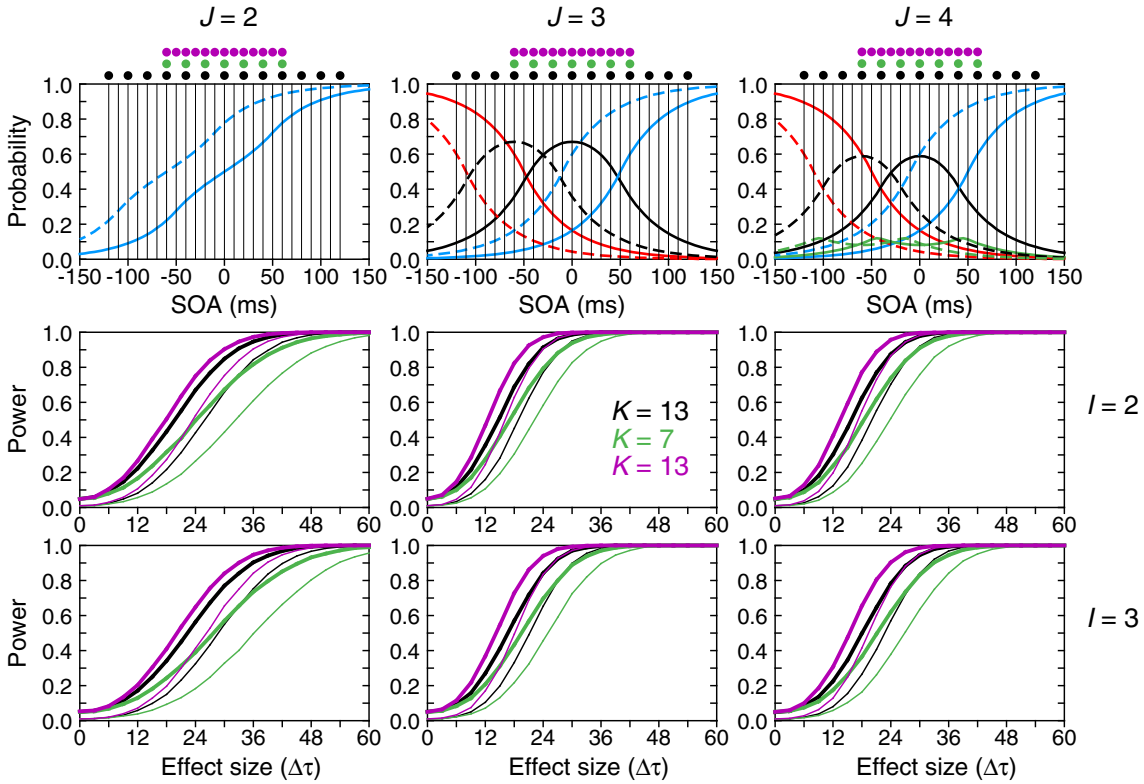



**Fig. 9** Power of the generalized Mantel–Haenszel test of equality of psychometric functions as a function of effect size, defined with respect to the lateral displacement caused by parameter τ. The number of trials per level was $f_{i\bullet k}$ = 20, identical for all $i$ and $k$. Layout and graphic conventions as in Fig. 6. The parameters of the psychometric functions were identical in all populations except that τ = 0 in population 1 (solid curves in the top row), τ = Δτ in population 2 (with values given by the abscissa in the panels in the center and bottom rows; dashed curves at the top have Δτ = 60), and τ = Δτ/2 in population 3. Results for $I$ = 2 (center row) involve populations 1 and 2, whereas results for $I$ = 3 (bottom row) involve the three populations. *SOA* stimulus-onset asynchrony

Figs. 6 and 7 of Logvinenko et al. (2012) reveals that the generalized Berry–Mielke test also outperforms the test developed by Logvinenko et al.

When psychometric functions differ across populations in slope, the generalized Berry-Mielke test substantially outperforms the generalized Mantel–Haenszel test, particularly when stimulus levels span a region where the sign of the differences between psychometric functions reverses (see Fig. 12, compared to Fig. 10). The low power that can still be noted in the left column of Fig. 12 (for $I$ = 2) is due to the fact that psychometric functions did not differ much between populations, as can be seen in the top-left panel of Fig. 12 for the largest difference involved in theses simulations. In such conditions, the psychometric functions only differ meaningfully at the outer extremes, which are only probed with the sampling points indicated by the black circles above the top panels.

In sum, the accuracy of the generalized Berry–Mielke test is identical to that of the generalized Mantel–Haenszel test and its power is only slightly lower when psychometric functions do not cross. However, when psychometric functions cross, the generalized Berry–Mielke test can detect differences that go undetected by the generalized Mantel–Haenszel test.

## Empirical Type-I error rate and power of the split Mantel–Haenszel test

Compared to the preceding tests, the accuracy of the split Mantel–Haenszel was negligibly inferior only when row marginal frequencies were small ($f_{i\bullet k} \leq 10$) with more than two response categories ($J > 2$) and only two populations ($I$ = 2). Otherwise, accuracy was identical to that of the two other tests. Graphical presentation of these results is omitted.

Figure 13 shows power functions in the same conditions in which power functions were reported for the two other tests in Figs. 9 and 11, namely, when psychometric functions differ only in location. Remarkably, the power of the split Mantel–Haenszel test in this case is only minimally lower than that of the generalized Mantel–Haenszel test. Splitting is unnecessary with the non-crossing psychometric functions for $J$ = 2 (top panel in the left column of Fig. 13) and these results show that splitting does not meaningfully affect power in this case. When $J$ = 3 or $J$ = 4 (center and right columns of Fig. 13), the monotonic psychometric functions for the first and third categories (red and blue curves in the top panels) do not demand splitting; however, the psychometric functions for the second category (black curves) and for the fourth

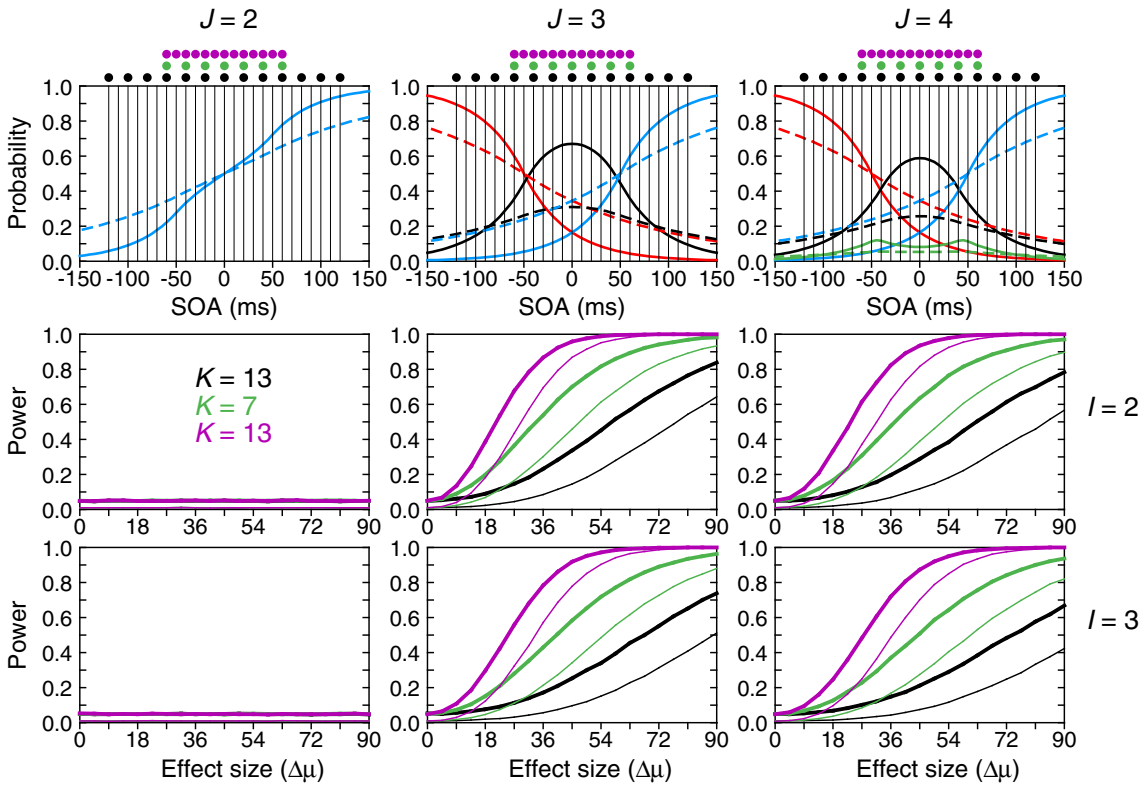



**Fig. 10** Power of the generalized Mantel–Haenszel test of equality of psychometric functions as a function of effect size, defined with respect to the slope determined by parameters $\lambda_A$ and $\lambda_B$, which were kept equal to preserve symmetry. The number of trials per level was $f_{i\bullet k} = 20$, identical for all $i$ and $k$. Layout and graphic conventions as in Fig. 6. The parameters of the psychometric functions were identical in all populations except that $1/\lambda_A = 1/\lambda_B = \mu = 45$ in population 1 (solid curves in the top row), $1/\lambda_A = 1/\lambda_B = \mu + \Delta\mu$ in population 2 (with values given by the abscissa in the panels in the center and bottom rows; dashed curves at the top have $\Delta\mu = 90$), and $1/\lambda_A = 1/\lambda_B = \mu + \Delta\mu/2$ in population 3. Results for $I = 2$ (center row) involve populations 1 and 2, whereas results for $I = 3$ (bottom row) involve the three populations. *SOA* stimulus-onset asynchrony

category (green curves, when $J = 4$) are non-monotonic and their differences are better assessed with a split at the stimulus level where these functions cross. Given the way in which differences were produced for these simulations, the optimal split is not at the central stimulus level but rather at an off-center location that varies for the actual data in each replicate. We decided against tailoring the split to the data in each replicate so that the results in Fig. 13 can show that slightly suboptimal splits do not compromise power. And, again, a comparison of power functions in the left panel in the center row of Fig. 13 (for $I = J = 2$) with power functions in Figs. 6 and 7 of Logvinenko et al. (2012) reveals that the split Mantel–Haenszel test outperforms the test devised by Logvinenko et al.

Figure 14 shows power functions for the split Mantel–Haenszel test in the same conditions in which power functions were reported for the two other tests in Figs. 10 and 12, namely, when psychometric functions differ in slope. When $J = 2$, splitting improves power relative to the generalized Mantel–Haenszel test (compare the left columns in Figs. 10 and 14) and to the same extent achieved with the generalized Berry–Mielke test (compare the left columns in Figs. 12 and 14). Recall that the apparently poor power observed in Figs. 12 and 14 when $J = 2$ is due to psychometric functions that do not differ much across populations, especially when they are probed at the green or magenta sampling points. When $J = 3$ or $J = 4$, split computation also increases power relative to the generalized Mantel–Haenszel test when the set of stimulus levels spans a region where psychometric functions cross (black sampling points and black power functions). With stimulus levels confined to a region where the psychometric functions do not cross (green or magenta sampling points and power functions), power functions for the split Mantel–Haenszel test are virtually identical to those of the generalized Mantel–Haenszel test.

## Comparison with CATANOVA

Variants of ANOVA have been proposed to deal with categorical data, which might thus circumvent the problems that ANOVAs bring when the data are counts or proportions. For instance, Light and Margolin (1971; see also Margolin & Light, 1974; Gitlow, 1976) proposed a one-way analysis of

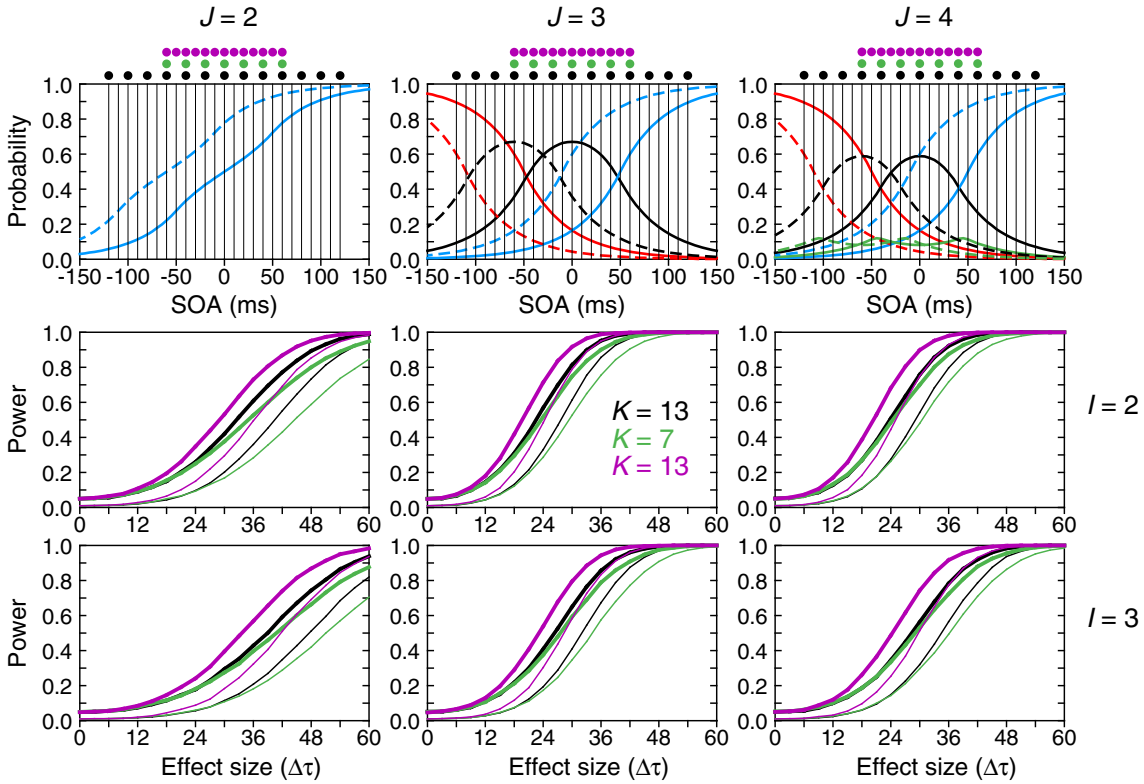



**Fig. 11** Power of the generalized Berry–Mielke test when psychometric functions differ in lateral displacement across populations. Graphic layout and simulation conditions are identical to those for which the power of the generalized Mantel–Haenszel test were reported in Fig. 9. *SOA* stimulus-onset asynchrony

variance for categorical data (which they referred to as CATANOVA) for $I \times J$ contingency tables that implies $K = 1$ and is unsuitable for our applications. Two-way extensions for use with $I \times J \times K$ tables were developed by Anderson and Landis (1980), Onukogu (1985a, b), and Singh (1996), although Gabriel (1963) had developed an analogous method much earlier. Here we will assess the performance of CATANOVA using Singh's method, which ensures an orthogonal decomposition of the sums of squares (the reader is referred to Singh's paper for details about computation of the test statistics).

Two-way CATANOVA tests for effects of factor A (in our case, population), effects of factor B (in our case, stimulus level), and effects of interaction. Remarkably, no study seems to have documented the accuracy and power of two-way CATANOVA, which makes our comparison most needed. We first checked (and confirmed) that Type-I error rates for all effects are adequate when the true distribution of responses across categories varies neither across populations (i.e., factor A) nor across stimulus levels (i.e., factor B), which implies the unrealistic case of flat psychometric functions. However, CATANOVA performed rather poorly under the conditions of interest in our context, namely, when effects of factor B exist (i.e., when the distribution of responses across the $J$ categories varies across stimulus levels).

We ran simulations to investigate the statistical properties of two-way CATANOVA for $J \in \{2, 3, 4\}$ and $I \in \{2, 3, 4\}$ at $\alpha \in \{.05, .01\}$. For reasons that will become clear immediately, results are reported only for a subset of the conditions used earlier and only for the broad set of $K = 13$ sampling points. Because real psychometric functions are not flat, strong effects of stimulus level (factor B) exist that make the corresponding test highly significant, but effects of population (factor A) or interaction effects do not exist under the conditions in which data are generated for accuracy studies. Then, rejection rates for the corresponding hypotheses should stay at the nominal level. Figure 15 shows the accuracy of two-way CATANOVA in the format of Fig. 3. Lines of different color indicate the rejection rate for the test of main effects of factor A (population), interaction effects, and either of them under a Bonferroni correction. Quite clearly, empirical rejection rates are far below their nominal levels. When a test is as inaccurate as this, its eventual power is irrelevant and meaningless but we checked that it is nominally inferior to that of our three other tests (results not shown). Hence, two-way CATANOVA is not a viable option to test equality of psychometric functions. The reason for its unruly performance is perhaps that the CATANOVA test statistics were derived under the assumption that null interaction also implies the absence of main effects (see Eq. 2.4 in Onukogu, 1985a),

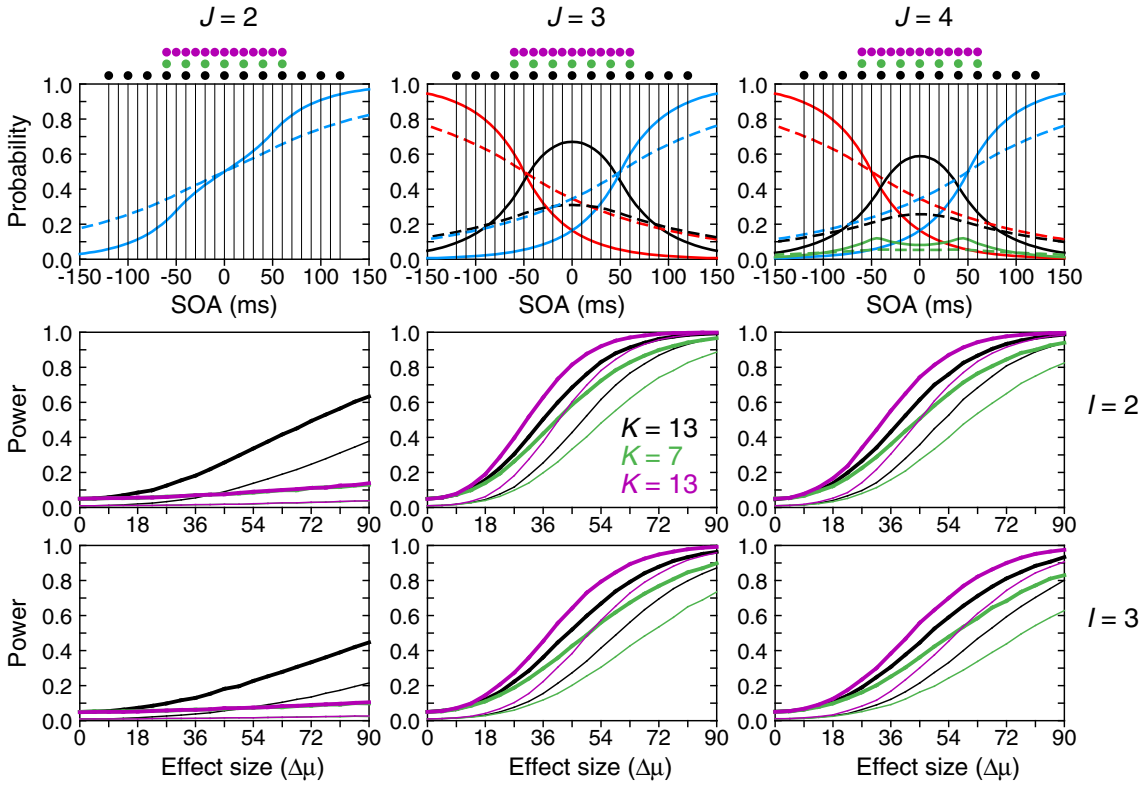



**Fig. 12** Power of the generalized Berry–Mielke test when psychometric functions differ in slope across populations. Graphic layout and simulation conditions are identical to those for which the power of the generalized Mantel–Haenszel test were reported in Fig. 10. *SOA* stimulus-onset asynchrony

an assumption that does not hold for real psychometric functions with which the distribution of responses across the $J$ categories differs across the $K$ stimulus levels. Derivation of a variant of CATANOVA where main and interaction effects are independent of one another is beyond the scope of this paper and unnecessary for our purposes.

## Illustrative applications

This section discusses potential applications of these tests and illustrates them using empirical data from published studies. A comparison with a parametric approach is also given at the end of this section using the artificial data presented in Fig. 1 and using also some of the empirical data presented next.

An obvious application for tests of equality of psychometric functions is in the assessment of whether empirical data collected across several sessions involving the same condition can reasonably be aggregated before further analyses are conducted. This was the motivation for the test developed by Logvinenko et al. (2012) and it is a frequent concern in empirical studies that require the collection of large amounts of data across several sessions, where the observers' sensory state can vary across sessions (for empirical evidence to this effect see, e.g., García-Pérez, 2010; Leek, Hanna, & Marshall, 1991; von Dincklage, Olbrich, Baars, & Rehberg, 2013). A formal test is surely more dependable than judging by eye whether the shape described by data from different sessions look alike (e.g., Hutsell & Jacobs, 2013; Oliveira & Machado, 2008; Yang, Meijer, Buitenweg, & van Gils, 2016). In this type of application, each observer's data on each experimental condition are analyzed separately with $I$ standing for the number of sessions at which data had been collected.

A second application in which each observer's data are also analyzed separately is in the assessment of the effects of experimental manipulations in within-subjects designs where the same observers provide data under all conditions. In these cases, $I$ stands for the number of conditions under study. An interesting form of this analysis arises in the assessment of order or position effects, by which psychometric functions vary with the order or position in which two stimuli are displayed in dual-presentation tasks (Alcalá-Quintana & García-Pérez, 2011; Bausenhart, Dyjas, & Ulrich, 2015; Dyjas & Ulrich, 2014; García-Pérez, 2014b; García-Pérez & Alcalá-Quintana, 2011a, 2011b; García-Pérez & Peli, 2011, 2014, 2015; Self, Mookhoek, Tjalma, & Roelfsema, 2015; Ulrich & Vorberg, 2009; von Castell, Hecht, & Oberfeld, 2017). In this case, $I$ stands for the two orders or positions of presentation of stimuli. However, even in studies that use dual-presentation tasks for other substantive purposes, order or position effects contaminate parameter estimates

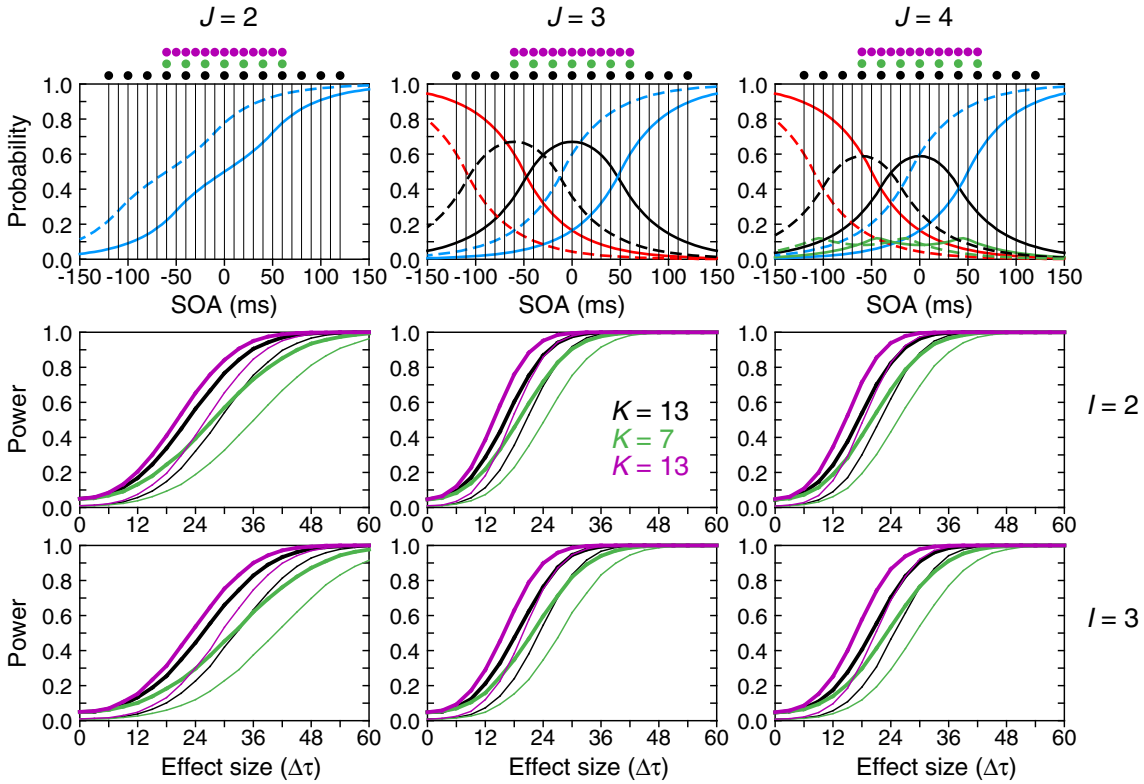



**Fig. 13** Power of the split Mantel–Haenszel test when psychometric functions differ in lateral displacement across populations. Graphic layout and simulation conditions are identical to those for which the power of the generalized Mantel–Haenszel test were reported in Fig. 9 and the power of the generalized Berry–Mielke test were reported in Fig. 11. *SOA* stimulus-onset asynchrony

considerably (see Ulrich & Vorberg, 2009). An assessment of their presence is good practice, as is the use of methods to eliminate this contamination.

An also conceivable application is in the context of between-subjects studies, where differences between groups are under scrutiny (e.g., patients vs. normal controls, or musicians vs. non-musicians), but also when different experimental conditions are for some reason administered to different (but presumably equivalent) groups. In these cases, *I* stands for the number of groups to be compared, but it is obvious that these analyses must be carried out with data aggregated across all observers in each group because it is certainly irrelevant whether some observer's psychometric function differs from that of someone else in other group. Such analyses may be preceded by separate analyses of homogeneity within each group, in which *I* would stand for the number of observers in the corresponding group.

The applications just described are illustrated next using data from a study by Lee and Noppeney (2014), in which judgments of audiovisual synchrony were collected at $K = 13$ SOAs ranging from −360 ms to 360 ms in steps of 60 ms. A binary synchrony judgment (SJ2) task was used in which observers simply report whether audio and video signals seemed subjectively simultaneous or asynchronous, rendering psychometric functions of the type shown in Fig. 2c above. The study involved a mixed 2 × 3 × 2 factorial design with group membership (musicians and non-musicians) as a between-subjects factor and with type of stimulus (speech, sinewave speech, and music) and stimulus duration (short and long) as within-subjects factors. Overall, 32 trials were administered at each SOA to each of 21 musicians and 20 non-musicians under each within-subjects condition. Type of stimulus and SOA were randomly interwoven in each block of 312 trials (3 types of stimuli × 13 SOAs × 8 repetitions) and different blocks involved short or long stimuli. Two blocks for each stimulus duration were administered on each of two days. Occasional observers missed a block or performed an additional block, which slightly altered the total number of trials per SOA per condition across observers.

The study used a parametric approach by fitting psychometric functions separately to each observer's data in each condition and subsequently conducting ANOVAs to assess mean differences in the outcome measure, which was the width of the *temporal integration window* (TIW) defined as the normalized area under the fitted psychometric function within the interval [−360, 360]. Our examples will focus instead on nonparametric comparisons of psychometric functions within and across selected conditions. These examples only aim at illustrating the various applications of the tests and they should not be misconstrued as a suggestion that such analyses address the main goals of Lee and Noppeney (2014), that this is how they should have analyzed their data

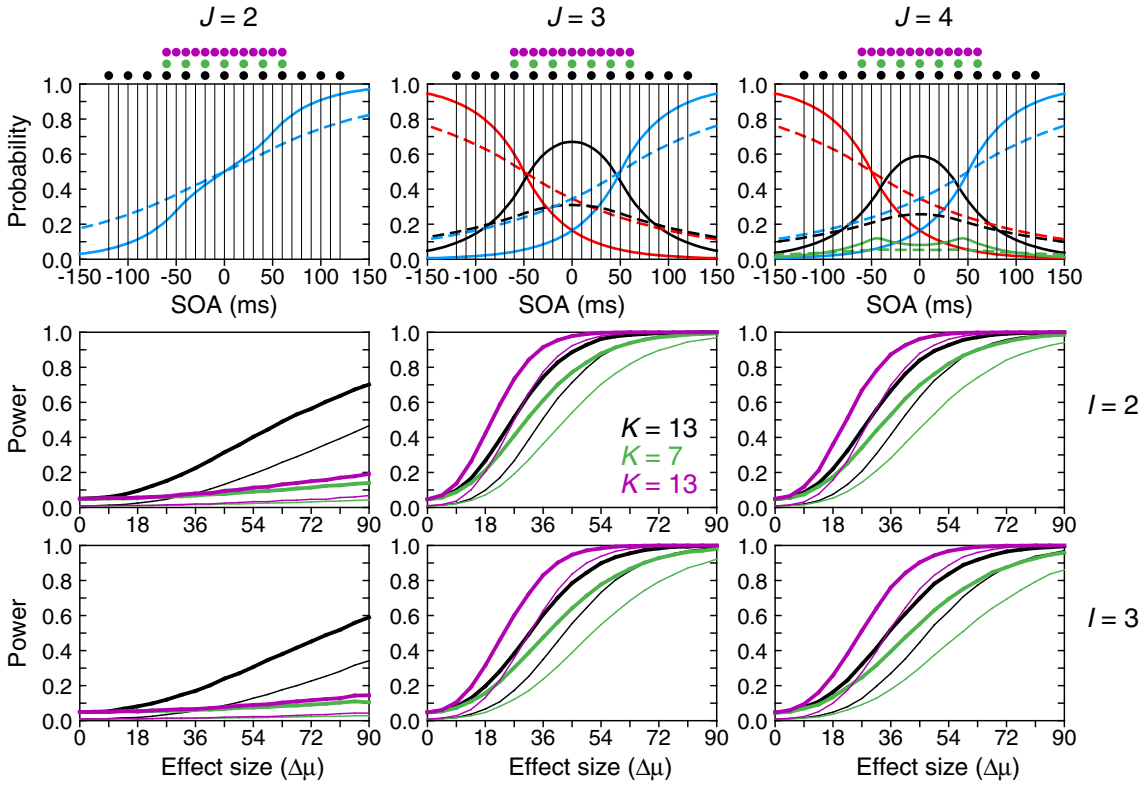



**Fig. 14** Power of the split Mantel–Haenszel test when psychometric functions differ in slope across populations. Graphic layout and simulation conditions are identical to those for which the power of the generalized Mantel–Haenszel test were reported in Fig. 10 and the power of the generalized Berry–Mielke test were reported in Fig. 12. *SOA* stimulus-onset asynchrony

to serve those goals, or that all these nonparametric tests should be conducted in all empirical studies.

Data from other studies involving dual-presentation tasks will subsequently illustrate the use of tests of equality of psychometric functions to assess order or position effects. The characteristics of those studies are described later.

### Stability of performance across days

Our first inquiry regards the stability of each observer's performance across the two days in which data were collected. One would expect that the data collected on each day come from the same underlying psychometric function, which calls for tests of equality of psychometric functions across the two days for each participant in each condition. We will restrict our analysis to observers who did not miss any block of trials, resulting in a grand total of 231 tests (instead of the 41 × 3 × 2 = 246 tests that complete data would have allowed). In each of these tests, $I = 2$ (the two days), $J = 2$ (the two response categories in the SJ2 task), $K = 13$ (the 13 SOAs), and $f_{i\bullet k} = 16$ for all $i$ and $k$. Even if the true psychometric function for each participant in each condition were the same on both days, at $\alpha = .05$ one still has to expect 5 % rejections (i.e., about 12 out of 231). The numbers of rejections were instead 88 (38.1 %), 65 (28.1 %), and 93 (40.3 %) by the generalized Mantel–Haenszel test, the generalized Berry–Mielke test, and the split Mantel–Haenszel test (with a 6–7 split), respectively. These different numbers of rejections across tests are well in line with the differences in power documented earlier in this paper.

One cannot obviously conclude that each rejection reveals non-equality of psychometric functions (because $100\alpha\%$ rejections are expected in case of equality) and that each non-rejection reveals equality (because power is not unity and, then, a certain percentage of non-rejections will always occur in case of non-equality). The moderately large percentage of rejections just mentioned nonetheless indicates that stability of performance across days is untenable (García-Pérez, 2012, 2017), which leaves the question open as to whether or not aggregating data from both days for all observers is still sensible for further analyses. This is not a question that statistical tests can provide an answer for. Figure 16 shows sample cases of rejection (top row) and non-rejection (bottom row) by the $Q_{GMH}$ statistic. Even when equality is rejected, the paths of the data on each day may justify aggregation for subsequent analyses.

### Homogeneity of groups

Our second inquiry regards the homogeneity of each group in each within-subjects condition, using data aggregated across the two days for each observer regardless of missing or extra sessions. This analysis requires 12 tests: six within-subjects

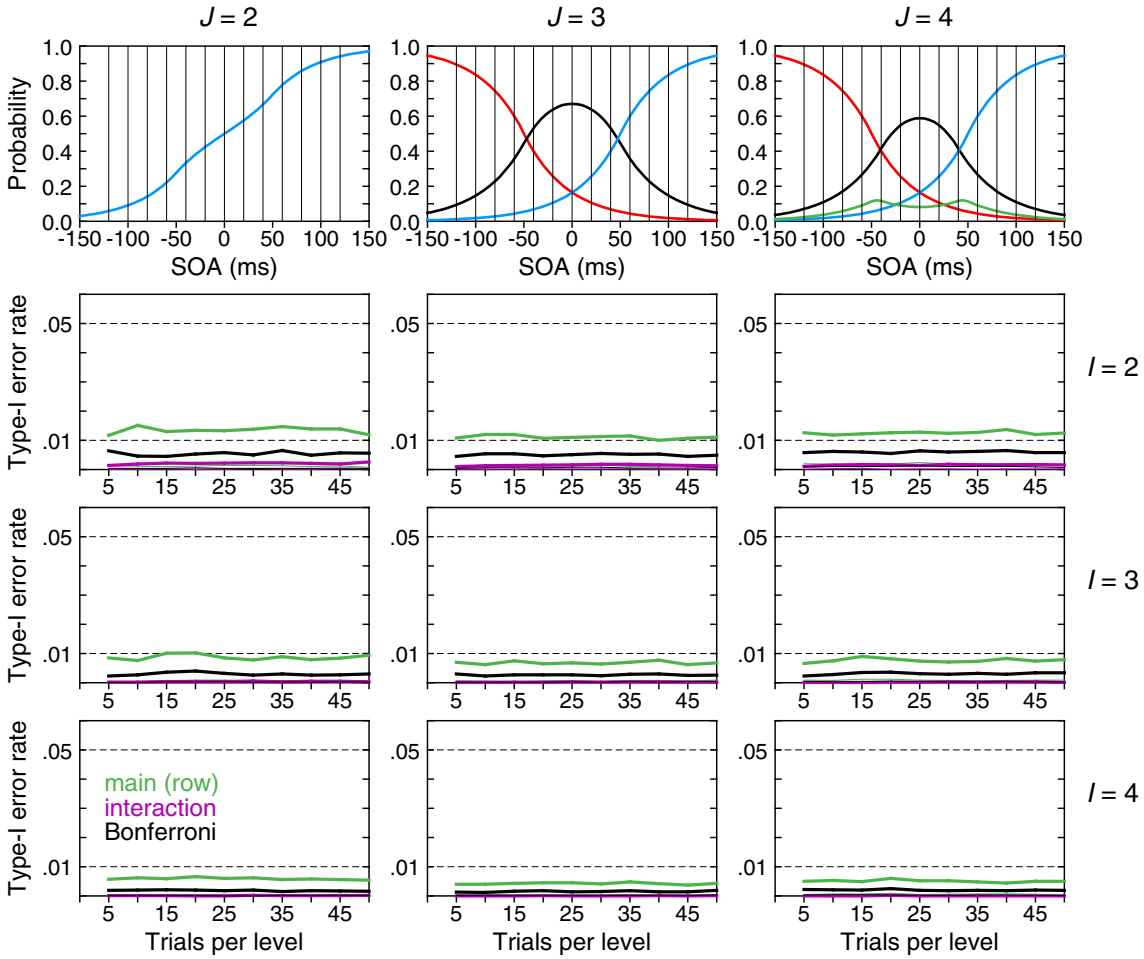



**Fig. 15** Accuracy of CATANOVA under the exact same conditions illustrated in Fig. 3 for the generalized Mantel–Haenszel test. The curves in the three bottom rows reflect the empirical Type-I error rate of the test for main effects of the row dimension (population; green curves), the effects of interaction (magenta curves), and a combined test for either type of effect with a Bonferroni correction (black curves). Note that the visible thick curves that meander near or below an ordinate of .01 are actually those pertaining to $\alpha = .05$; the curves pertaining to $\alpha = .01$ are sitting virtually at an ordinate of 0. With these non-flat psychometric functions, main effects of the column dimension (stimulus level) exist which were invariably significant and are not reported in the plots. *SOA* stimulus-onset asynchrony

conditions in each of two groups. In each of the six tests for the group of musicians, $I = 21$ (the observers), $J = 2$ (the two response categories in the SJ2 task), $K = 13$ (the 13 SOAs), and $f_{i\bullet k} = 32$ for all $i$ and $k$ in most cases (i.e., the number of trials per SOA aggregated across the two days), although this number was instead 24 for observers who missed a block for some stimulus duration and 40 for observers who completed an extra block with some stimulus duration; in the group of non-musicians the only difference is that $I = 20$ instead. There is obviously no reason to expect all musicians (or non-musicians, for that matter) to display analogous performance, but large and disparate individual differences among the members of each group surely taint subsequent group comparisons (Estes, 1956; Estes & Maddox, 2005). At $\alpha = .05$, all 12 tests resulted in rejections by the generalized Mantel–Haenszel test, the generalized Berry–Mielke test, and the split Mantel–Haenszel test (also with a 6–7 split). Figure 17 shows the empirical psychometric functions of the 21 musicians in each of the six within-subjects conditions, revealing large individual differences that naturally substantiate the significant tests.

### Differences between speech and sinewave speech

The third illustration involves a comparison of performance for speech and sinewave speech stimuli of the same duration, which is again carried out individually for each observer. This implies 41 tests (one per observer), each with $I = 2$ (the two types of speech at, e.g., short duration), $J = 2$, $K = 13$, and $f_{i\bullet k} =$ 24, 32, or 40 for all $i$ and $k$ according to the amount of data collected from each observer. Although, in principle, one might surmise that the two types of speech stimuli should produce similar performance by each observer, at $\alpha = .05$ the numbers of rejections were 31 (75.6 %), 28 (68.3 %), and 33 (80.5 %) by the generalized Mantel–Haenszel test, the generalized Berry–Mielke test, and the split Mantel–Haenszel test (also with a 6–7 split), respectively. An analogous set of 41

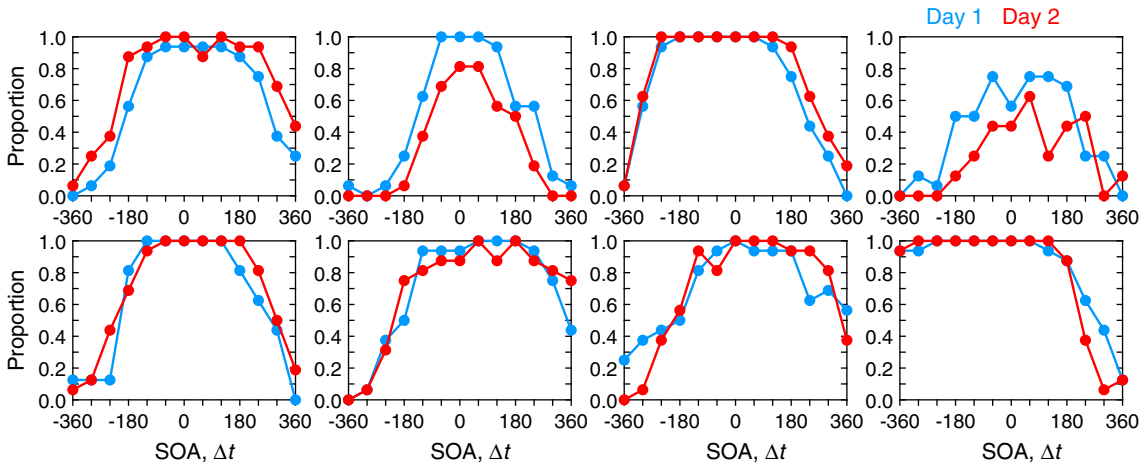



**Fig. 16** Sample psychometric functions for eight observers (panels) in the study of Lee and Noppeney (2014), with data from day 1 (blue curves) and day 2 (red curves). Equality of psychometric functions was rejected by the $Q_{GMH}$ test for cases in the top row and was not rejected for cases in the bottom row. All data come from the group of musicians with sinewave speech stimuli of long duration. *SOA* stimulus-onset asynchrony

tests for long speech and long sinewave speech stimuli rendered 30 (73.2 %), 25 (61.0 %), and 29 (70.7 %) rejections, respectively by the generalized Mantel–Haenszel, generalized Berry–Mielke, and split Mantel–Haenszel tests. A graphic illustration of these results is omitted.

A percentage of rejections well beyond the 5 % rate expected by chance thus document meaningful within-subject differences in perception of audiovisual synchrony for speech and sinewave speech at both stimulus durations. This comparison was not relevant to Lee and Noppeney's (2014) goals and, thus, they did not report tests of equality of means conducted on their outcome measure. However, from their Fig. 2, it seems unlikely that the mean TIW will differ significantly between speech and sinewave speech at any of the two durations. This judgment is admittedly speculative, but a discrepancy between the outcomes of parametric and nonparametric tests will be demonstrated in full later.

## Differences between groups

Finally, checking for differences between groups in each within-subjects condition requires six tests (one per condition) with $I = 2$ (the two groups, using data aggregated across all observers in each group), $J = 2$, $K = 13$, and very large $f_{i \bullet k}$

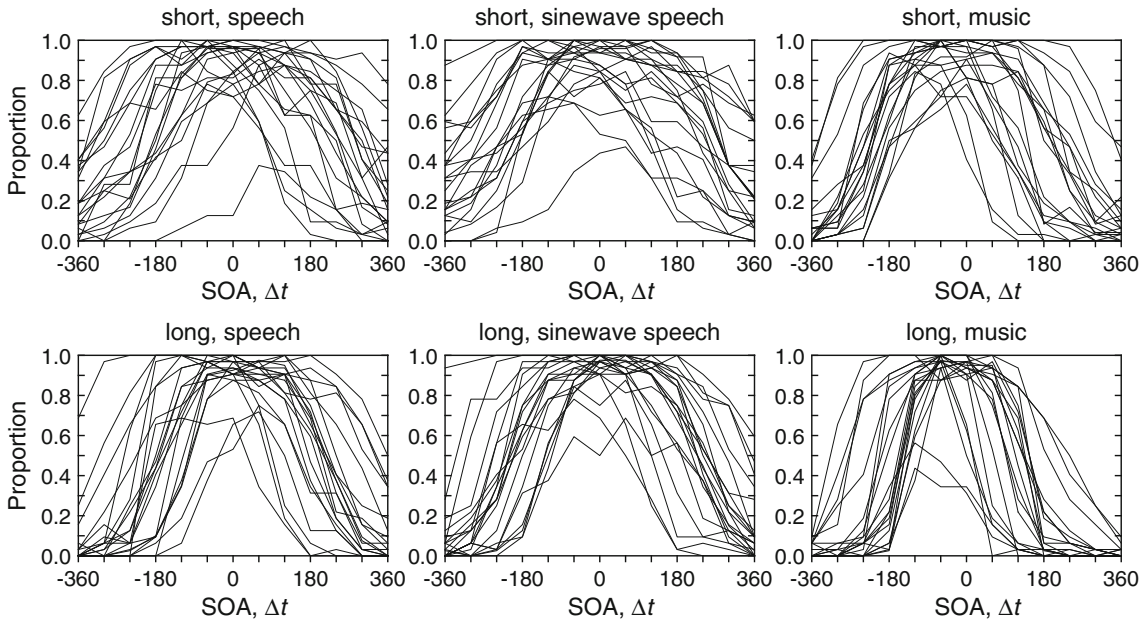



**Fig. 17** Empirical psychometric functions of the 21 musicians (undifferentiated curves in each panel) who served in each and all of the experimental conditions (panels) in the study of Lee and Noppeney (2014). *SOA* stimulus-onset asynchrony

(coming from the overall number of trials administered across all the members of each group) that also differ across $i$ due to the different sizes of each group. All the tests turned out significant by all three statistics at $\alpha = .05$, as might be expected given the systematic differences displayed graphically in Fig. 18.

This type of group comparison, which requires aggregation of data from all the observers in each group, could arguably be performed via conventional ANOVA with group membership as a between-subjects factor, stimulus level as a repeated-measures factor, experimental condition as another within-subjects factor (in case an ANOVA is not conducted separately for each condition), and with the proportion of "synchronous" responses given by each observer as the dependent variable. However, the applicability of such an approach is limited to conditions in which $J = 2$.

### Order or position effects in dual-presentation tasks

Our first illustration of this type assesses position effects in data from the fourth experiment reported by Self et al. (2015), a visual contrast discrimination study that involved two within-subjects conditions (referred to as "figure" and "ground"; for details, see Self et al.). Test and standard stimuli were presented simultaneously on each trial, one above and one below a fixation point, with presentation locations randomly interwoven across trials (so that $I = 2$; upper or lower location of the test stimulus). A ternary task was used (i.e., $J = 3$) in which observers had to indicate on each trial whether the upper stimulus had higher contrast, the lower stimulus had higher contrast, or both seemed to have the same contrast. Data were collected at $K = 11$ levels of the test stimulus, with $f_{i \bullet k} = 40$ at each $i$ and $k$ in each within-subjects condition for each of seven observers. This implies 14 tests of equality.

Figure 19 shows the empirical psychometric functions for each observer in each condition. Each panel shows the two sets of psychometric functions to be compared. At $\alpha = .05$, the numbers of rejections were 13 (92.9 %), 11 (78.6 %), and 12 (85.7 %) by the generalized Mantel–Haenszel test, the generalized Berry–Mielke test, and the split Mantel–Haenszel test (with a 5–6 split), respectively, and note that the 5–6 split seems optimal by eye in all cases. Again, massive rejections reveal the presence of meaningful position effects and, thus, advice against aggregating data across presentation position and demand fitting instead psychometric functions separately for each presentation position, as Self et al. (2015) actually did. The only case in which equality across presentation positions was not rejected by any of the three statistics is that of observer #6 in the "ground" condition (see the corresponding panel in Fig. 19). Contentious cases where equality was only rejected by one or two statistics are those of observers #3 and #4 in the "figure condition." In all other cases, empirical differences across presentation orders appear sufficiently systematic to substantiate rejection of equality by all three tests. Note, however, that the direction of the differences (direction of lateral displacement of one set of psychometric functions relative to the other) varies across observers, an issue to which we will come back later.

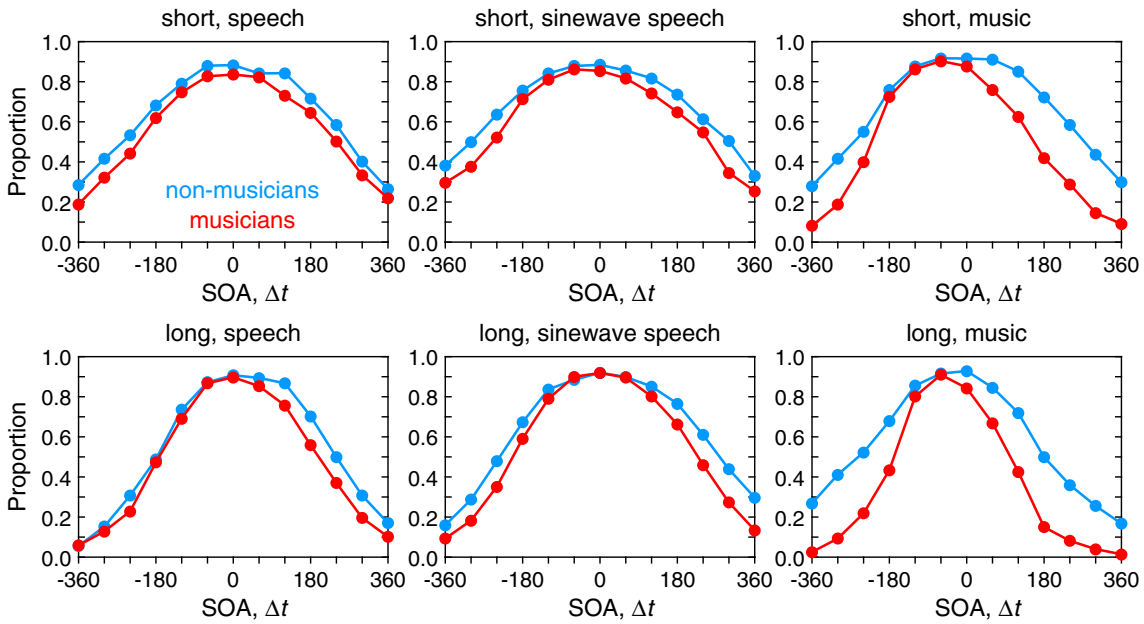



**Fig. 18** Empirical psychometric functions for aggregated data from all non-musicians (blue curves) and all musicians (red curves) in each experimental condition (panels) in the study of Lee and Noppeney (2014). Discrepancies between the paths described by these data and the paths of analogous curves plotted in Fig. 1 of Lee and Noppeney occur because each proportion plotted here is computed after aggregating raw data, whereas each proportion plotted there is the average of the individual proportions computed for each observer. *SOA* stimulus-onset asynchrony

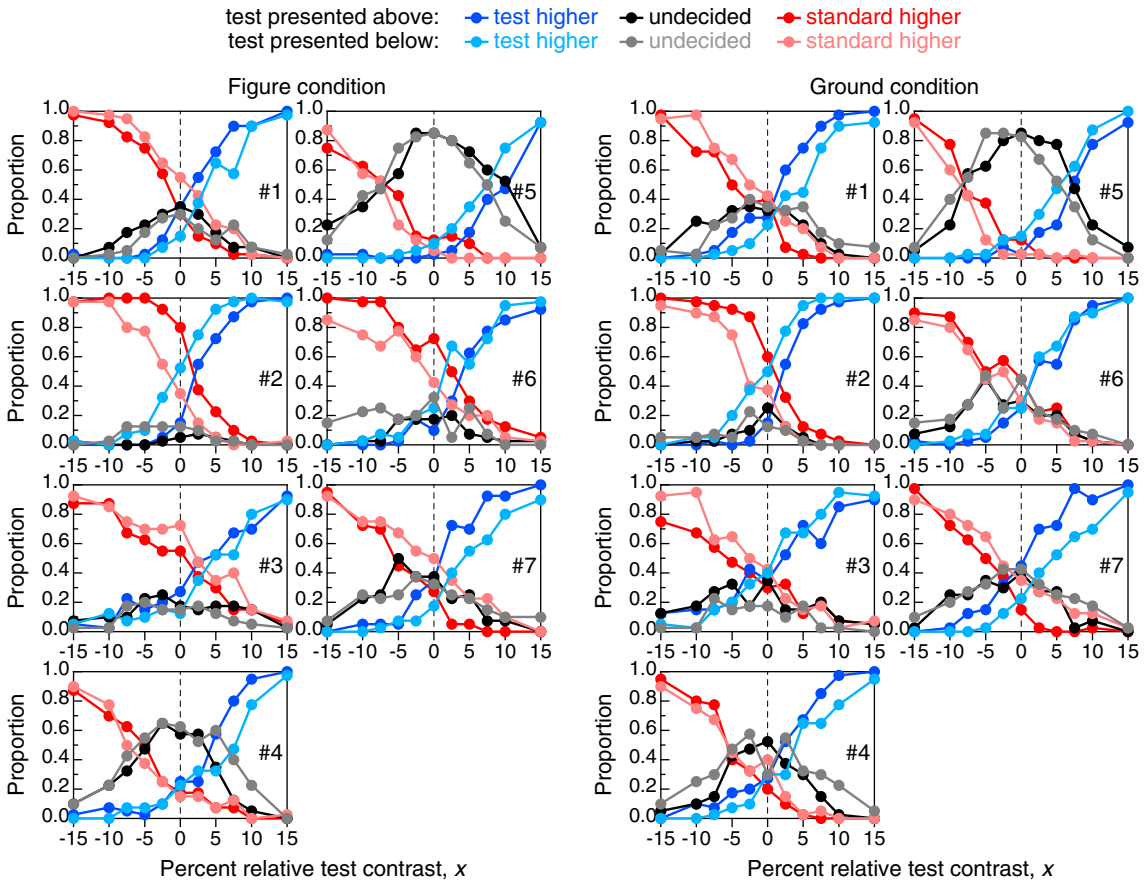



**Fig. 19** Empirical psychometric functions for visual contrast discrimination for each of seven observers (panels) in two conditions (figure and ground; left and right parts), from experiment 4 of Self et al. (2015). Responses in each of the three categories (test higher, standard higher, and undecided) were separated according to the position (above or below) at which the test stimulus was presented in each trial (see the legend at the top). The vertical dashed line in each panel indicates the contrast of the standard stimulus

Our second illustration assesses order effects and uses data collected with an adaptive method, which implies that row marginal frequencies $f_{i\bullet k}$ vary across $i$ and $k$ with the potential for $f_{i\bullet k} = 0$ at one or more $i$ in one or more $k$. Data come also from a study on visual contrast discrimination that used a sequential presentation of test and standard stimuli in each trial (Alcalá-Quintana & García-Pérez, 2011; their Experiment 2), thus resulting in a potential for order effects. Five observers served in three within-subjects conditions, thus requiring 15 tests of equality of psychometric functions across presentation orders (so that, again, $I = 2$). Data were collected with a ternary task (i.e., $J = 3$) identical to that of Self et al. (2015). In each of the three conditions (sensory desensitization during the first interval of each dual-presentation trial, sensory desensitization during the second interval, and no desensitization), 768 trials under each presentation order were randomly interwoven within and across consecutive blocks. The stimulus level on each trial was selected adaptively and separately for each presentation order, which implies that the overall number of stimulus levels as well as their values were not required in advance to be the same for both presentation orders. The adaptive method selected stimulus levels from a fixed lattice and, then, the eventual occurrence of tables with $N_k = 1$ still left a sufficiently large number of tables with large $N_k$. Some stimulus levels were indeed used with only one of the presentation orders (i.e., $f_{i\bullet k} = 0$ for some $i$ at some $k$) and, for stimulus levels that were used with both presentation orders, the number of trials varied across presentation orders (i.e., $f_{i\bullet k}$ varies across $i$ at all $k$).

Figure 20 shows the empirical psychometric functions for each observer in each condition. Each panel shows again the two sets of psychometric functions to be compared. Data at stimulus levels that were used in only one of the $I$ populations are uninformative for lack of data from the other population to compare with. Hence, for data collected with adaptive methods, the test of equality of psychometric functions would use responses from only the set of $K$ stimulus levels used in both populations (or, more generally, in more than one population when $I > 2$). Thus, for the case in the top-left panel of Fig. 20, only stimulus levels from $x = -1.25$ to $x = -0.8$ would be used, yielding $K = 9$. This is effectively what the computations involved in the tests end up doing when provided instead with data from the 18 (common and unique) stimulus levels (from $x = -1.4$ to $x = -0.55$ in our sample case), with $f_{ijk}$

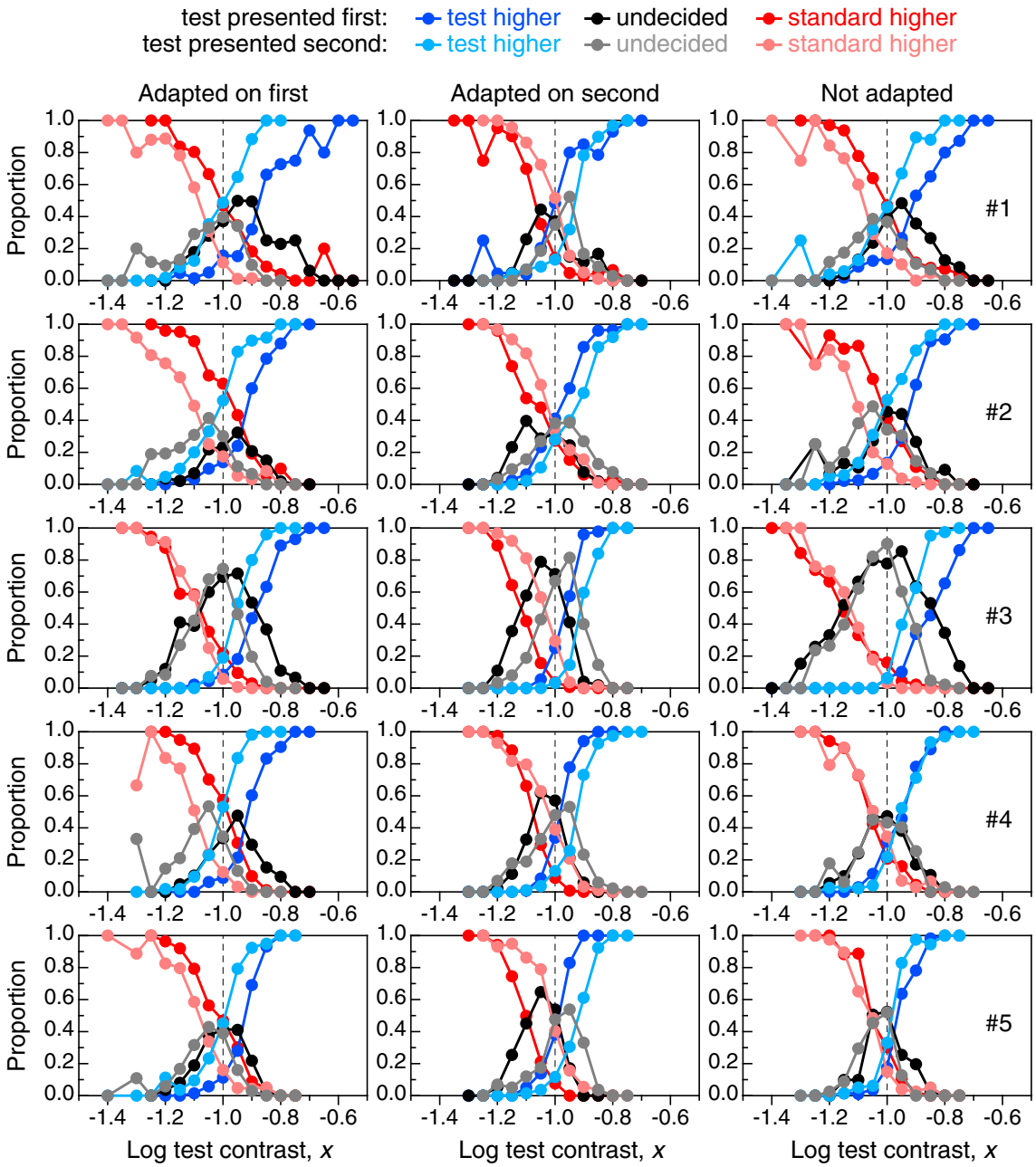



**Fig. 20** Empirical psychometric functions for visual contrast discrimination for each of five observers (rows) in three adaptation conditions (columns), from Experiment 2 of Alcalá-Quintana and García-Pérez (2011). Responses in each of the three categories (test higher, standard higher, and undecided) were separated according to the order of presentation of test and standard stimuli in each trial (test first or test second; see the legend at the top). The vertical dashed line in each panel indicates the contrast of the standard stimulus

= 0 at all $j$ for the stimulus levels $k$ at which no data had actually been collected in population $i$. Equality of psychometric functions across presentation orders was rejected at $\alpha = .05$ in all 15 cases by the generalized Berry–Mielke test; the generalized Mantel–Haenszel test and the split Mantel–Haenszel test (with an even split) rejected equality in all cases except for observer #4 in the "not adapted" condition, and note again that an even split also seems adequate by eye with these data.

## Comparison with a parametric approach

We will first revert to the data in Fig. 1 for a comparison of the outcomes of nonparametric and parametric approaches. The artificial situation remains unspecified besides the fact that data were generated for a purported SJ3 task with the model of perception of temporal order discussed earlier and using the same parameter values $(\lambda_A, \lambda_B, \tau, \delta_1, \delta_4) = (1/45, 1/45, 0, -50, 50)$ for both populations. Populations 1 and 2 in this example

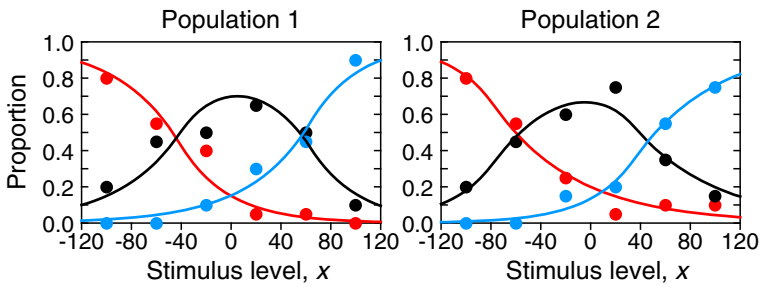
might well reflect the two sessions in which data for some condition were collected, representing the empirically realistic case in which the need to interweave several conditions in a session limits the amount of data that can be collected per condition (as in Lee and Noppeney, 2014, among many other studies). In a situation like this, a researcher will have as many two-population cases of this type as there are observers and conditions in the study and, hence, multiple situations for tests of equality of psychometric functions.

The outcomes of our nonparametric tests for the singled-out artificial data in Fig. 1 were reported earlier: None of the three tests rejected equality of psychometric functions. This outcome happens to match the reality that generated the data although this type of assessment is impossible in empirical studies. The main point, however, is that nonparametric approaches allow for formal tests to assess equality of psychometric functions in each individual case. As seen next, this is simply impossible with parametric approaches.

Parametric approaches fit suitable psychometric functions to the data and compare either their estimated parameters or performance measures derived from them. We used the software in Alcalá-Quintana and García-Pérez (2013) to obtain parameter estimates and performance measures for the model psychometric functions in Eq. 15 in each of the two populations, with the results shown in Fig. 21. The fit was good in both cases by the log-likelihood ratio statistic ($p = .402$ and $p = .592$ in populations 1 and 2, respectively). Beyond this point, assessing equality of psychometric functions involves subjective judgments of similarity of the fitted functions or subjective judgments of differences between parameter estimates ($\lambda_A$, $\lambda_B$, $\tau$, $\delta_1$, and $\delta_4$) or performance measures (PSS or SR, defined in the caption to Fig. 21) across populations. The reason that only informal judgments are possible is that each psychometric function has been reduced to a single set of distinct quantities ($\lambda_A$, $\lambda_B$, $\tau$, $\delta_1$, $\delta_4$, PSS, and SR), which precludes statistical tests of equality. Comparing the numerical values by eye is hampered by the fact that they are rather inaccurate estimates of the corresponding true values because of the scarce data used to obtain them, an unavoidable inconvenience in the empirical context of this discussion. Standard errors of estimation are not available either but they are known to be relatively large with magnitudes that vary across parameters as a function of number of trials and the location of the $K$ sampling points relative to the unknown shape and location of the true psychometric functions (see, e.g., Fig. 15 in Alcalá-Quintana & García-Pérez, 2013; see also García-Pérez, 2014a). Then, even an informal assessment of equality based on standard errors of estimation is unfeasible, besides the fact that such type of comparison may be inconclusive due to different outcomes for different parameters.

Because formal tests of equality of psychometric functions are unfeasible, the conventional approach takes a detour by testing for *equality of means* (of parameter values or performance measures) at a group level. In the situation under discussion, this may involve paired-samples $t$ tests or repeated-measures ANOVAs to assess mean differences in the estimates of, say, the SR across sessions, with the means computed over observers. In other words, conventional parametric approaches assess group-level differences in the mean of isolated quantitative aspects of psychometric functions. Interesting as these analyses may be *in a second stage*, they differ from testing equality of psychometric functions per se. Differences between means may not be significant when psychometric functions actually differ on an observer-by-observer basis: Equality of means is a necessary consequence of equality of psychometric functions but it is not a sufficient condition for it. An illustration in the context of order or position effects using empirical data will reveal why this parametric strategy is inadequate.

We mentioned that order or position effects vary in direction across observers. Figure 19 revealed that data points for one of the presentation positions systematically lie above data points for the other in some observers (e.g., observer #2 in both conditions), whereas it is the other way around for other observers (e.g., observer #7 also in both conditions). The same holds for order effects in Fig. 20.




True and estimated parameters

| | True | Pop. 1 | Pop. 2 |
|---|---|---|---|
| $1/\lambda_A$ | 45.00 | 49.19 | 36.43 |
| $1/\lambda_B$ | 45.00 | 38.28 | 65.78 |
| $\tau$ | 0.00 | -11.55 | 20.69 |
| $\delta_1$ | -50.00 | -52.69 | -56.22 |
| $\delta_4$ | 50.00 | 52.69 | 56.22 |
| PSS | 0.00 | 4.97 | -4.55 |
| SR | 94.67 | 100.70 | 104.41 |

**Fig. 21** Psychometric functions in Eq. 15 fitted to the data in Fig. 1. The table on the right lists the true parameters with which data were generated (identical for both populations) and the estimated parameters for each population. Parameters $\delta_1$ and $\delta_4$ were estimated under the constraint $\delta_1 = -\delta_4$, which holds for their true counterparts. PSS and SR at the bottom of the lists are performance measures extracted from the fitted functions. The PSS (point of subjective simultaneity) is the stimulus level at which the psychometric function for S responses (black curve) peaks; the SR (width of the synchrony range) is the range of stimulus levels at which S responses are more prevalent than AF or BF responses

These effects make data look as if psychometric functions were laterally shifted in opposite directions across presentation orders or positions. Evidence of individual differences in the direction and magnitude of this shift is overwhelming (e.g., Bausenhart, Dyjas, & Ulrich, 2015; Dyjas & Ulrich, 2014; see also the many other studies discussed in García-Pérez & Alcalá-Quintana, 2011a, 2017b). A parametric approach first measures the signed magnitude of the shift for each observer via estimated parameters and then tests the average shift against zero. With a mixture of observers showing meaningful shifts in each direction, the average shift may not be significantly different from zero and will be mistaken for a "proof" that position or order effects are not present in the data.

To substantiate this claim numerically, we used the software in García-Pérez and Alcalá-Quintana (2017b) to fit suitable psychometric functions to data from each observer and condition in Fig. 19. The fit was good as measured by the log-likelihood ratio statistic, which did not reject the fitted model in any case at $\alpha = .05$. The lateral shift across presentation positions was then measured for each observer and condition using the estimated parameters, which rendered values ranging from −3.52 to 3.37 (in the units of the horizontal axis) with an average of 0.138 and a standard deviation of 2.264. The effect size is negligible (Cohen's $d_z = 0.06$) and a one-sample $t$ test did not reject a zero mean ($t_{13} = 0.219$, $p = .829$). Obviously, this result only says that the *average magnitude of position effects* is reasonably null, which is not to say that position effects are not present in the data and should not be taken care of. Only nonparametric tests can assess within-subject position effects adequately and we showed earlier that their use revealed significant effects in almost all cases: Equality of psychometric functions was rejected massively on an observer-by-observer basis.

## Practical recommendations

The software included as Electronic Supplementary Material conducts the three nonparametric tests presented here (with the user-selected split for the $S$-$Q_{GMH}$ statistic) taking care of all their subtleties (e.g., removal of tables for which $\gamma_{T_k} < 0.5$ in case of the $G$ statistic, proper consideration of the degrees of freedom of the $S$-$Q_{GMH}$ statistic, etc.) and returning information about these aspects (see examples in the user's manual that accompanies the software). Then, the only practical concern for a user is the criterion by which a test should be selected and, eventually, the choice of a split for the $S$-$Q_{GMH}$ statistic. These issues are discussed next.

Test accuracy is not a consideration in the choice of a test because all of them maintain the nominal Type-I error rate. However, power differs non-uniformly across tests depending on the number $J$ of response categories and, when $J = 2$, depending also on whether empirical data display evidence that the psychometric functions for the $I$ populations cross.

In principle, with binary tasks (i.e., $J = 2$) that render monotonic psychometric functions, the generalized Mantel–Haenszel test is the optimal choice with non-crossing data (as in Fig. 2a); with crossing data (as in Fig. 2b), the optimal choices are instead the generalized Berry–Mielke test and the split Mantel–Haenszel test with a split at the crossing point. The same criterion applies to binary tasks that render instead non-monotonic psychometric functions, where the most common case is that the functions cross as illustrated in Fig. 2c and, then, the optimal choices are again the generalized Berry–Mielke test and the split Mantel–Haenszel test with a split at the crossing point. However, it may so happen that such non-monotonic psychometric functions do not cross (see Figs. 16 and 18). In such cases, the optimal choice is again the generalized Mantel–Haenszel test.

With non-binary tasks (where $J > 2$) crossings will almost always be observed in empirical data whenever psychometric functions actually differ across populations (see some empirical examples in Figs. 19 and 20). The generalized Mantel–Haenszel test should thus be avoided and the optimal choices are again the generalized Berry–Mielke test and the split Mantel–Haenszel test. Because of the multiple psychometric functions and crossings when $J > 2$, the optimal split for the latter test is at the stimulus level where the ensemble of psychometric functions has a vertical axis of approximate bilateral symmetry. It can easily be seen in the panels of Figs. 19 and 20 that the vertical dashed line used for an even split in those examples is optimal or near-optimal by this criterion. However, the choice of split is not a serious concern because our results reveal that slightly suboptimal splits do not reduce power meaningfully.

We have also shown that power varies with the number and location of the stimulus levels at which data are collected. This factor does not have any bearing on the choice of a test for the data on hand, but it raises the question as to whether criteria exist to pre-select stimulus levels that maximize power. Ideal criteria obviously exist: Select levels at which psychometric functions differ the most. However, without knowledge of whether and where such differences occur, this ideal is impossible to realize but some action can nevertheless be taken. Consider Fig. 16 and remember that data were collected for all observers at a fixed set of pre-selected stimulus levels. In retrospect, these levels were inadequate for the observer in the bottom-right panel: The left-most increasing range of the psychometric functions is entirely missing, the central range where they do not differ is over-represented, and only some evidence of differences shows at the right-most decreasing range. Use of adaptive methods would have rendered more informative data and a more powerful test because these methods choose stimulus levels wherever necessary to cover

the entire range of the psychometric function (for an illustration addressing the exact same problem, see Fig. 10 in García-Pérez, 2014a).

Finally, the number of trials placed at each stimulus level also affects the power of the tests, although this has again no bearing on the choice of a test for the data on hand. More data always implies more power but it also places more burden on the observers and lengthens the duration of experimental sessions, which are the only practical limitations to collecting the large amounts of data that would ensure large power.

## Conclusion

This paper has defined and compared several nonparametric tests for equality of psychometric functions. Three of them (the generalized Mantel–Haenszel test, the generalized Berry–Mielke test, and the split Mantel–Haenszel test) have adequate (and indistinguishable) accuracy but they differ slightly as to power. In contrast, the CATANOVA test also assessed in this paper is highly inaccurate. Of the three advisable tests, the generalized Mantel–Haenszel test is the most powerful when the psychometric functions to be compared do not cross but it is unable to detect differences materializing in psychometric functions that cross. In contrast, the generalized Berry–Mielke test and the split Mantel–Haenszel test overcome the problems that cause the failure of the generalized Mantel–Haenszel test.

These three tests of equality of psychometric functions also outperform the test proposed by Logvinenko et al. (2012) in the limited scenario where the latter is applicable, namely, $I = J = 2$ and monotonic psychometric functions that differ only by lateral translation. The three tests presented here are more general, as they can be used in other scenarios (i.e., $I \geq 2, J \geq 2$, and monotonic or non-monotonic psychometric functions that differ in any respect). The tests can be used when row marginal frequencies $f_{i\bullet k}$ are identical for all $i$ and $k$ (i.e., with data collected with the method of constant stimuli) or when data are collected instead with adaptive methods that render row marginal frequencies $f_{i\bullet k}$ that vary across $i$ and $k$, including the potential for $f_{i\bullet k} = 0$ at one or more $i$ in one or more $k$. Adaptive collection of data is generally advisable because of its ability to probe psychometric functions where needed and, hence, to increase the power of these tests.

Although the tests were presented in the context of assessing the equality of psychometric functions, they are applicable whenever homogeneity of distributions across strata needs to be assessed. For instance, the detection of differential item functioning (DIF; Finch, 2016) in educational or psychological measurement requires a comparison of item responses (in $J$ categories) among $I$ groups of individuals, with respondents in each group classified into $K$ strata according to their total score. The generalized Mantel–Haenszel test is one of the most common approaches to detect DIF but this approach is known to be insensitive to nonuniform crossing DIF, a pattern of differences whose sign varies among groups across the $K$ strata in a form thoroughly analogous to that illustrated in Fig. 2b (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Uttaro & Millsap, 1994). Our results show that the generalized Berry–Mielke test and the split Mantel–Haenszel test are suitable nonparametric tests to detect nonuniform crossing DIF.

To facilitate applications, fully documented software in MATLAB and R is available as Supplementary Material to test equality of psychometric functions (or DIF detection or, more generally, homogeneity of distributions in any context) with the statistic of choice ($Q_{GMH}$, $G$, or $S$-$Q_{GMH}$) and with the user's choice of split for the $S$-$Q_{GMH}$ statistic.

**Acknowledgements** This research was supported by grants PSI2015-67162-P (Ministerio de Economía y Competitividad), MTM2013-40941-P and MTM2016-74931-P (Ministerio de Economía y Competitividad and FEDER), UFI11/03 (Universidad del País Vasco UPV/EHU), and IT-642-13 (Departamento de Educación del Gobierno Vasco - UPV/EHU Econometrics Research Group). We thank Hweeling Lee and Matthew Self for sharing their data for our illustrations.

## References

Agresti, A. (1983). Testing marginal homogeneity for ordinal categorical variables. *Biometrics*, *39*, 505–510. https://doi.org/10.2307/2531022

Agresti, A. (2002). Categorical data analysis (second edition). New York: Wiley.

Alcalá-Quintana, R., & García-Pérez, M. A. (2011). A model for the time-order error in contrast discrimination. *Quarterly Journal of Experimental Psychology*, *64*, 1221–1248. https://doi.org/10.1080/17470218.2010.540018

Alcalá-Quintana, R., & García-Pérez, M. A. (2013). Fitting model-based psychometric functions to simultaneity and temporal-order judgment data: MATLAB and R routines. *Behavior Research Methods*, *45*, 972–998. https://doi.org/10.3758/s13428-013-0325-2

Anderson, R. J., & Landis, J. R. (1980). Catanova for multidimensional contingency tables: Nominal-scale response. *Communications in Statistics – Theory and Methods*, *9*, 1191–1206. https://doi.org/10.1080/03610928008827952

Bausenhart, K. M., Dyjas, O., & Ulrich, R. (2015). Effects of stimulus order on discrimination sensitivity for short and long durations. *Attention, Perception, & Psychophysics*, *77*, 1033–1043. https://doi.org/10.3758/s13414-015-0875-8

Berry, K. J., & Mielke, P. W., Jr. (1988). Monte Carlo comparisons of the asymptotic chi-square and likelihood-ratio tests with the nonasymptotic chi-square test for sparse $r \times c$ tables. *Psychological Bulletin*, *103*, 256–264. https://doi.org/10.1037/0033-2909.103.2.256

Birch, M. W. (1965). The detection of partial association, II: The general case. *Journal of the Royal Statistical Society, Series B*, *27*, 111–124. Retrieved from http://www.jstor.org/stable/2984488

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*, 364–366. https://doi.org/10.2307/2284382

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

Capa, R.L., Duval, C. Z., Blaison, D., & Giersch, A. (2014). Patients with schizophrenia selectively impaired in temporal order judgments. *Schizophrenia Research*, *156*, 51–55. https://doi.org/10.1016/j.schres.2014.04.001

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, *46*, 440–464. Retrieved from http://www.jstor.org/stable/2345686

Donohue, S. E., Woldorff, M. G., & Mitroff, S. R. (2010). Video game players show more precise multisensory temporal processing abilities. *Attention, Perception, & Psychophysics*, *72*, 1120–1129. https://doi.org/10.3758/APP.72.4.1120

Droit-Volet, S., Bigand, E., Ramos, D., & Oliveira Bueno, J. L. (2010). Time flies with music whatever its emotional valence. *Acta Psychologica*, *135*, 226–232. https://doi.org/10.1016/j.actpsy.2010.07.003

Dyjas, O., & Ulrich, R. (2014). Effects of stimulus order on discrimination processes in comparative and equality judgements: Data and models. *Quarterly Journal of Experimental Psychology*, *67*, 1121–1150. https://doi.org/10.1080/17470218.2013.847968

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140. https://doi.org/10.1037/h0045156

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*, 403–408. https://doi.org/10.3758/BF03193784

Eubank, R. L. (1997). Testing goodness of fit with multinomial data. *Journal of the American Statistical Association*, *92*, 1084–1093. https://doi.org/10.1080/01621459.1997.10474064

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

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

Finch, W. H. (2016). Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. *Applied Measurement in Education*, *29*, 30–45. https://doi.org/10.1080/08957347.2015.1102916

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). Statistical distributions (fourth edition). New York: Wiley.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*, 2–18. https://doi.org/10.1037/a0024338

Gable, P. A., & Poole, B. D. (2012). Time flies when you're having approach-motivated fun: Effects of motivational intensity on time perception. *Psychological Science*, *23*, 879–886. https://doi.org/10.1177/0956797611435817

Gabriel, K. R. (1963). Analysis of variance of proportions with unequal frequencies. *Journal of the American Statistical Association*, *58*, 1133–1157. https://doi.org/10.1080/01621459.1963.10480694

García-Pérez, M. A. (2010). Denoising forced-choice detection data. *British Journal of Mathematical and Statistical Psychology*, *63*, 75–100. https://doi.org/10.1348/000711009X424057

García-Pérez, M. A. (2012). Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in Psychology*, *3*:325. https://doi.org/10.3389/fpsyg.2012.00325

García-Pérez, M. A. (2014a). Adaptive psychophysical methods for non-monotonic psychometric functions. *Attention, Perception, & Psychophysics*, *76*, 621–641. https://doi.org/10.3758/s13414-013-0574-2

García-Pérez, M. A. (2014b). Does time ever fly or slow down? The difficult interpretation of psychophysical data on time perception. *Frontiers in Human Neuroscience*, *8*:415. https://doi.org/10.3389/fnhum.2014.00415

García-Pérez, M. A. (2017). Thou shalt not bear false witness against null hypothesis significance testing. *Educational and Psychological Measurement*, *77*, 631–662. https://doi.org/10.1177/0013164416668232

García-Pérez, M. A., & Alcalá-Quintana, R. (2005). Sampling plans for fitting the psychometric function. *Spanish Journal of Psychology*, *8*, 256–289. https://doi.org/10.1017/S113874160000514X

García-Pérez, M. A., & Alcalá-Quintana, R. (2011a). Interval bias in 2AFC detection tasks: Sorting out the artifacts. *Attention, Perception, & Psychophysics*, *73*, 2332–2352. https://doi.org/10.3758/s13414-011-0167-x

García-Pérez, M. A., & Alcalá-Quintana, R. (2011b). Improving the estimation of psychometric functions in 2AFC discrimination tasks. *Frontiers in Psychology*, *2*:96. https://doi.org/10.3389/fpsyg.2011.00096

García-Pérez, M. A., & Alcalá-Quintana, R. (2012a). On the discrepant results in synchrony judgment and temporal-order judgment tasks: A quantitative model. *Psychonomic Bulletin & Review*, *19*, 820–846. https://doi.org/10.3758/s13423-012-0278-y

García-Pérez, M. A., & Alcalá-Quintana, R. (2012b). Response errors explain the failure of independent-channels models of perception of temporal order. *Frontiers in Psychology*, *3*:94. https://doi.org/10.3389/fpsyg.2012.00094

García-Pérez, M. A., & Alcalá-Quintana, R. (2015a). Converging evidence that common timing processes underlie temporal-order and simultaneity judgments: A model-based analysis. *Attention, Perception, & Psychophysics*, *77*, 1750–1766. https://doi.org/10.3758/s13414-015-0869-6

García-Pérez, M. A., & Alcalá-Quintana, R. (2015b). The left visual field attentional advantage: No evidence of different speeds of processing across visual hemifields. *Consciousness and Cognition*, *37*, 16–26. https://doi.org/10.1016/j.concog.2015.08.004

García-Pérez, M. A., & Alcalá-Quintana, R. (2015c). Visual and auditory components in the perception of asynchronous audiovisual speech. *i-Perception*, *6*(6), 1–20. https://doi.org/10.1177/2041669515615735

García-Pérez, M. A., & Alcalá-Quintana, R. (2017a). Perceived temporal order and simultaneity: Beyond psychometric functions. In: A. Vatakis, F. Balci, A. Correa, & M. Di Luca (Eds.), Timing and time perception: Procedures, measures, and applications. Boston: Brill.

García-Pérez, M. A., & Alcalá-Quintana, R. (2017b). The indecision model of psychophysical performance in dual-presentation tasks: Parameter estimation and comparative analysis of response formats. *Frontiers in Psychology*, *8*:1142. https://doi.org/10.3389/fpsyg.2017.01142

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. https://doi.org/10.1080/03610910802538351

García-Pérez, M. A., & Peli, E. (2011). Visual contrast processing is largely unaltered during saccades. *Frontiers in Psychology*, *2*:247. https://doi.org/10.3389/fpsyg.2011.00247

García-Pérez, M. A., & Peli, E. (2014). The bisection point across variants of the task. *Attention, Perception, & Psychophysics*, *76*, 1671–1697. https://doi.org/10.3758/s13414-014-0672-9

García-Pérez, M. A., & Peli, E. (2015). Aniseikonia tests: The role of viewing mode, response bias, and size–color illusions. *Translational*

*Vision Science & Technology*, *4*(3):9. https://doi.org/10.1167/tvst.4.3.9

Gil, S., Rousset, S., & Droit-Volet, S. (2009). How liked and disliked foods affect time perception. *Emotion*, *9*, 457–463. https://doi.org/10.1037/a0015751

Gitlow, H. S. (1976). CATANOVA: A program for computing analysis of variance for categorical data. *Journal of Marketing Research*, *13*, 408–409. Retrieved from http://www.jstor.org/stable/3151028

Hutsell, B. A., & Jacobs, E. A. (2013). Attention and psychophysics in the development of stimulus control. *Journal of the Experimental Analysis of Behavior*, *100*, 282–300. https://doi.org/10.1002/jeab.54

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*:863. https://doi.org/10.3389/fpsyg.2013.00863

Landis, J. R., Heyman, E. R., & Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. *International Statistical Review*, *46*, 237–254. https://doi.org/10.2307/1402373

Lee, H., & Noppeney, U. (2014). Music expertise shapes audiovisual temporal integration windows for speech, sinewave speech, and music. *Frontiers in Psychology*, *5*:868. https://doi.org/10.3389/fpsyg.2014.00868

Leek, M. R., Hanna, T. E., & Marshall, L. (1991). An interleaved tracking procedure to monitor unstable psychometric functions. *Journal of the Acoustical Society of America*, *90*, 1385–1397. https://doi.org/10.1121/1.401930

Lewis, T., Saunders, I. W., & Westcott, M. (1984). The moments of the Pearson chi-squared statistic and the minimum expected value in two-way tables. *Biometrika*, *71*, 515–522. https://doi.org/10.1093/biomet/71.3.515

Light, R. J., & Margolin, B. H. (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association*, *66*, 534–544. https://doi.org/10.1080/01621459.1971.10482297

Logvinenko, A. D., Tyurin, Y. N., & Sawey, M. (2012). A test for psychometric function shift. *Behavior Research Methods*, *44*, 503–515. https://doi.org/10.3758/s13428-011-0155-z

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748. https://doi.org/10.1093/jnci/22.4.719

Margolin, B. H., & Light, R. J. (1974). An analysis of variance for categorical data, II: Small sample comparisons with chi square and other competitors. *Journal of the American Statistical Association*, *69*, 755–764. https://doi.org/10.1080/01621459.1974.10480201

Mielke, P. W., Jr., & Berry, P. W. (1985). Non-asymptotic inferences based on the chi-square statistic for *r* by *c* contingency tables. *Journal of Statistical Planning and Inference*, *12*, 41–45. https://doi.org/10.1016/0378-3758(85)90051-5

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*, 257–274. https://doi.org/10.1177/014662169602000306

Nisen, J. A., & Schwertman, N. C. (2008). A simple method of computing the sample size for chi-square test for the equality of multinomial distributions. *Computational Statistics and Data Analysis*, *52*, 4903–4908. https://doi.org/10.1016/j.csda.2008.04.007

Oliveira, L., & Machado, A. (2008). The effect of sample duration and cue on a double temporal discrimination. *Learning and Motivation*, *39*, 71–94. https://doi.org/10.1016/j.lmot.2007.06.001

Onukogu, I. B. (1985a). An analysis of variance of nominal data. *Biometrical Journal*, *27*, 375–383. https://doi.org/10.1002/bimj.4710270404

Onukogu, I. B. (1985b). Reasoning by analogy from ANOVA to CATANOVA. *Biometrical Journal*, *27*, 839–849. https://doi.org/10.1002/bimj.4710270802

Pardo, M. C. (1998). Improving the accuracy of goodness-of-fit tests based on Rao's divergence with small sample size. *Computational Statistics and Data Analysis*, *28*, 339–351. https://doi.org/10.1016/S0167-9473(98)90131-1

Read, T. R. C. (1984). Small-sample comparisons for the power divergence goodness-of-fit statistics. *Journal of the American Statistical Association*, *79*, 929–935. https://doi.org/10.1080/01621459.1984.10477113

Rogers, H. J., & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel–Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*, 105–116. https://doi.org/10.1177/014662169301700201

Self, M. W., Mookhoek, A., Tjalma, N., & Roelfsema, P. R. (2015). Contextual effects on perceived contrast: Figure–ground assignment and orientation contrast. *Journal of Vision*, *15*(2)2. https://doi.org/10.1167/15.2.2

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, *13*, 238–241. Retrieved from http://www.jstor.org/stable/2984065

Singh, B. (1996). On CATANOVA method for analysis of two-way classified nominal data. *Sankhyā: The Indian Journal of Statistics, Series B*, *58*, 379–388. Retrieved from http://www.jstor.org/stable/25052964

Somes, G. W. (1986). The generalized Mantel–Haenszel statistic. *American Statistician*, *40*, 106–108. https://doi.org/10.1080/00031305.1986.10475369

Tipples, J. (2010). Time flies when we read taboo words. *Psychonomic Bulletin & Review*, *17*, 563–568. https://doi.org/10.3758/PBR.17.4.563

Ulrich, R., & Vorberg, D. (2009). Estimating the difference limen in 2AFC tasks: Pitfalls and improved estimators. *Attention, Perception, & Psychophysics*, *71*, 1219–1227. https://doi.org/10.3758/APP.71.6.1219

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel–Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement*, *18*, 15–25. https://doi.org/10.1177/014662169401800102

von Castell, C., Hecht, H., & Oberfeld, D. (2017). Measuring perceived ceiling height in a visual comparison task. *Quarterly Journal of Experimental Psychology*, *70*, 516–532. https://doi.org/10.1080/17470218.2015.1136658

von Dincklage, F., Olbrich, H., Baars, J. H., & Rehberg, B. (2013). Habituation of the nociceptive flexion reflex is dependent on interstimulus interval and stimulus intensity. *Journal of Clinical Neuroscience*, *20*, 848–850. https://doi.org/10.1016/j.jocn.2012.07.013

Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*, *118*, 75–83. https://doi.org/10.1016/j.cognition.2010.10.002

Wilbiks, J. M. P., & Dyson, B. J. (2013). Effects of temporal asynchrony and stimulus magnitude on competitive audio–visual binding. *Attention, Perception, & Psychophysics*, *75*, 1883–1891. https://doi.org/10.3758/s13414-013-0527-9

Yang, H., Meijer, H. G., Buitenweg, J. R., & van Gils, S. A. (2016). Estimation and identifiability of model parameters in human nociceptive processing using yes-no detection responses to electrocutaneous stimulation. *Frontiers in Psychology*, *7*:1884. https://doi.org/10.3389/fpsyg.2016.01884