The Spanish Journal of Psychology (2025), **28**, e6, 1–13
doi:10.1017/SJP.2025.1

Article

Goodness-of-fit Tests for Categorical Models of Psychological Processes: Fixing the Occasional Failures of Asymptotic Theory

Miguel A. García-Pérez and Rocío Alcalá-Quintana
Universidad Complutense, Spain

Abstract

The goodness of fit of categorical models of psychological processes is often assessed with the log-likelihood ratio statistic (G^2), but its underlying asymptotic theory is known to have limited empirical validity. We use examples from the scenario of fitting psychometric functions to psychophysical discrimination data to show that two factors are responsible for occasional discrepancies between actual and asymptotic distributions of G^2. One of them is the eventuality of very small expected counts, by which the number of degrees of freedom should be computed as $(J-1) \times I-P-K_{0.06}$, where J is the number of response categories in the task, I is the number of comparison levels, P is the number of free parameters in the fitted model, and $K_{0.06}$ is the number of cells in the implied $I \times J$ table in which expected counts do not exceed 0.06. The second factor is the administration of small numbers n_i of trials at each comparison level x_i ($1 \leq i \leq I$). These numbers should not be ridiculously small (i.e., lower than 10) but they need not be identical across comparison levels. In practice, when n_i varies across levels, it suffices that the overall number N of trials exceeds $40 \times I$ if $J = 2$ or $50 \times I$ if $J = 3$, with no n_i lower than 10. Correcting the degrees of freedom and using large n_i are easy to implement in practice. These precautions ensure the validity of goodness-of-fit tests based on G^2.

Keywords: asymptotic distribution; goodness of fit; log-likelihood ratio test; psychometric function

(Received: 12 September 2024; revised: 23 December 2024; accepted: 20 January 2025)

One of the ways in which psychology progresses is by identifying empirical relations among psychological constructs and the conditions that modulate such relations. Fulfilling this goal led to a stream of research in which data analysis focuses on testing for differences among the statistical parameters describing the univariate, bivariate, or multivariate distributions of measures of psychological constructs. Data analysis under this line of research traditionally used null hypothesis significance testing for dichotomous judgments about the presence or absence of an effect, but this approach is under scrutiny and alternatives have been proposed to pursue essentially the same goal. See, for instance, the collection of papers published in 2017 in three special sections of *Educational and Psychological Measurement* (volume 77, issues 3–5) or the analogous collection of papers published in 2019 in *The American Statistician* (volume 73, issue S1; https://www.tandfonline.com/toc/utas20/73/sup1).

An alternative stream of research whose popularity has increased in cognitive neuroscience over the past few decades consists of investigating the psychological processes that are presumably responsible for the measures of psychological constructs obtained in empirical research, which the aforementioned line of research would submit to statistical analyses. The goal here is to characterize formally the psychological processes that produce observable data, with only indirect interest in the statistical features of the data themselves. This stream of research proceeds by proposing mathematical models describing the operation of the psychological processes under study (e.g., Busemeyer & Diederich, 2010; Lewandowsky & Oberauer, 2018), an operation characterized by a set of parameters. Thus, categorical models of performance in perceptual or cognitive tasks capture the underlying operations needed to solve the applicable task and represent them in a functional description expressing how the probability of each response on the task (out of J possible responses) varies across the I levels of some variable of concern (say, I stimulus levels along a dimension of interest).[1] In some contexts, the functions that describe these changes in probability are referred to as "psychometric functions" and their mathematical expression is formally derived from the model. The parameters of a psychometric function capture the operation of the processes postulated by the model. Estimating those parameters from empirical data is the focus of this stream of research because these parameters are informative about individual differences or about the effects of experimental manipulations.

Unfortunately, parameter estimates can always be obtained, even when the categorical model does not do justice to the data.

Corresponding author: Miguel A. García-Pérez; Email: miguel@psi.ucm.es
Cite this article: García-Pérez, M. A., & Alcalá-Quintana, R. (2025). Goodness-of-fit Tests for Categorical Models of Psychological Processes: Fixing the Occasional Failures of Asymptotic Theory. *The Spanish Journal of Psychology* **28**, e6, 1–13. https://doi.org/10.1017/SJP.2025.1

[1] *Categorical models* apply to perceptual or cognitive tasks in which the number J of possible responses is finite (and generally small). When the task involves a continuous response format, *continuous models* apply instead (see, e.g., Huk, Bonnen & He, 2018; Rasanan et al., 2024). Our discussion in this study will be limited to categorical models.

This eventuality is hardly ever identifiable in the estimated values for the parameters. The reason is that parameter estimation algorithms seek values that minimize or maximize some goal function and a minimum or a maximum will always exist within the bounds of the parameter space. However, the algorithms do not incorporate any criterion to assess acceptability of the resultant estimates. Thus, a necessary step after parameter estimation is the statistical assessment of goodness of fit, which is actually the standard way of assessing the agreement between model and data and, thus, checking the trustworthiness of parameter estimates for the intended purposes of characterizing the process under study and, subsequently, drawing conclusions about individual differences or the effects of experimental manipulations.

This study discusses some little known facts concerning the assessment of model–data agreement via goodness-of-fit tests, an issue that directly bears on the validity of the resultant inferences. The plan of the study is as follows. First, and without loss of generality, the next section presents an illustrative example of a categorical model involving psychometric functions in duration discrimination tasks, where observers are asked to provide a perceptual judgment (in one of J response categories) with multiple trials at each of I stimulus magnitudes. Also in the next section, we use this sample illustration to bring up a general problem that arises when testing goodness of fit with the log-likelihood ratio statistic (G^2), namely, the eventual inaccuracy of the statistical test (i.e., the eventual disagreement between actual and nominal type I error rates). A subsequent section discusses where this problem comes from and how it can be solved. Specifically, we use simulations to show that inaccuracy has two sources: (1) a reduction of the number of degrees of freedom caused by low expected frequencies of one or more types of responses and (2) a violation of the distributional assumption caused by insufficient numbers of observations at one or more of the I stimulus magnitudes. Further simulations show that accuracy is regained by action that is in the hands of the researcher, namely, (1) subtracting one degree of freedom for each expected frequency that is lower than 0.06 and (2) ensuring that the number of observations collected at each of the I stimulus levels exceeds 10.

The context: a particular example

Consider a duration discrimination task that displays a sequence of two stimuli whose presentation durations differ. Data collection proceeds by administering a relatively large number of trials each of which pairs a fixed presentation duration (the standard) with some other duration (the comparison) from a set of I values around the standard. On each trial, observers may be asked to report a comparative judgment with a binary response format in which the only possible responses are "first longer" or "second longer." Observers may also be allowed to report that presentation duration did not seem to differ across stimuli, yielding a ternary response format in which the possible responses are "first longer," "second longer," or "equal." Each comparison duration is paired with the standard duration on a sufficiently large number of trials to obtain dependable estimates of the probability of each response at each individual pairing of standard and comparison. For simplicity and without loss of generality, we will assume that the standard duration is presented first and the comparison duration is presented second in each trial, although conventional practice randomizes the order of presentation of standard and comparison across trials.

Figure 1 shows two sets of artificial but realistic data collected with the binary task (Figure 1a) or the ternary task (Figure 1b). The standard duration was 250 ms, and the set of $I = 13$ comparison durations ranged from 100 ms to 400 ms in steps of 25 ms. The top row depicts a scenario in which the range of comparison durations covers the width of the psychometric functions, whereas the bottom row displays a scenario in which the width of the psychometric function is not fully covered by this range. In addition, 20 trials were administered at each comparison duration in the top row and 40 trials were administered at each comparison duration in the bottom row. The ordinate of each data point in each panel of Figure 1 indicates the proportion of trials in which the corresponding response was given at the comparison duration indicated by the abscissa. For conceptual clarity in the subsequent discussion of these examples, the panels in Figure 1a plot data for "first longer" and "second longer" responses, although the proportion of responses of either type is one's complement to the other.

The continuous curves in each panel of Figure 1 are psychometric functions fitted to each dataset. In the binary task, the fitted psychometric functions Ψ_1 and Ψ_2 for "first longer" and "second longer" responses are, respectively, given by

$$\Psi_1(x) = \Phi\left(\frac{\delta - \beta(x - x_s)}{\sqrt{2}}\right) \tag{1a}$$

$$\Psi_2(x) = 1 - \Psi_1(x) = \Phi\left(\frac{\beta(x - x_s) - \delta}{\sqrt{2}}\right) \tag{1b}$$

where x is the comparison duration, $x_s = 250$ ms is the standard duration, Φ is the unit-normal distribution function, and β and δ are free parameters. In the ternary task, the fitted psychometric functions Ψ_1, Ψ_2, and Ψ_E for "first longer," "second longer," and "equal" responses are, respectively, given by

$$\Psi_1(x) = \Phi\left(\frac{\delta_1 - \beta(x - x_s)}{\sqrt{2}}\right) \tag{2a}$$

$$\Psi_2(x) = \Phi\left(\frac{\beta(x - x_s) - \delta_2}{\sqrt{2}}\right) \tag{2b}$$

$$\Psi_E(x) = 1 - \Psi_1(x) - \Psi_2(x) \tag{2c}$$

where the free parameters are β, δ_1, and δ_2. Although this is immaterial here, the mathematical expressions in Eqs. 1 and 2 arise from the indecision model of García-Pérez and Alcalá-Quintana (2017) for the case in which the standard duration is presented first in each trial. The artificial data displayed in the top row of Figure 1 were generated from Eqs. 1 and 2 with $\beta = 0.035$ in both cases and $\delta = 0.3$ for binary data or $\delta_1 = -1.5$ and $\delta_2 = 0.5$ for ternary data;[2] data for the bottom row of Figure 1 were analogously simulated but with $\beta = 0.012$ instead.

Fitting psychometric functions to data implies estimating their free parameters. Maximum-likelihood estimates of the applicable parameters are indicated on the right side of each panel of Figure 1. Since the true values of the parameters are known in a simulation, one can on this occasion judge that the estimated parameters agree reasonably well with their true values, but this is a privilege never attainable in empirical research. In empirical practice, the relevant

[2]Incidentally, the facts that $\delta \neq 0$ in the binary task and $\delta_2 \neq -\delta_1$ in the ternary task imply a response bias that causes the axis of bilateral symmetry of the psychometric functions in Figure 1 to be displaced away from $x = x_s$ (for more details, see García-Pérez & Alcalá-Quintana, 2017). This feature does not have any bearing on the goodness-of-fit issues addressed in this study.

https://doi.org/10.1017/SJP.2025.1 Published online by Cambridge University Press

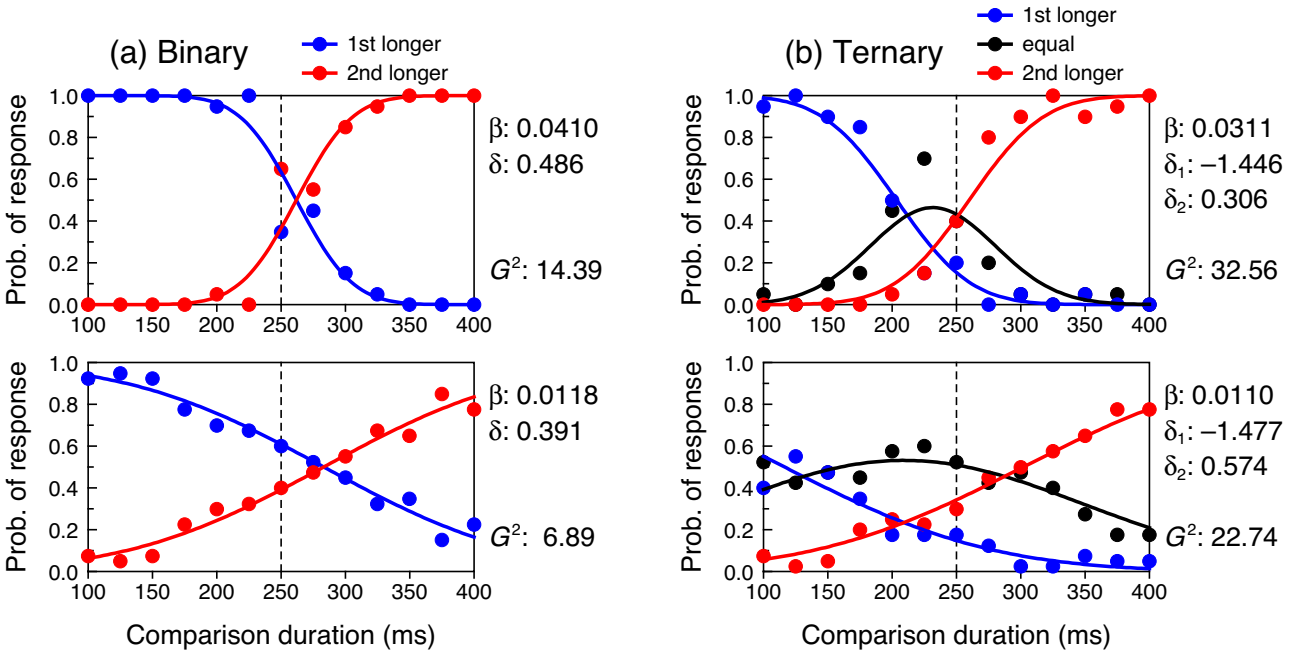

Figure 1. Artificial data (symbols) and fitted psychometric functions (curves) in binary (a) and ternary (b) discrimination tasks with a standard duration of 250 ms (vertical line in each panel) and comparison durations ranging from 100 ms to 400 ms (horizontal axis). Response types are indicated at the top of each row. Parameter estimates and the value of the G^2 statistic are indicated on the right side of each panel.

issue is not whether the estimated parameters match their unknown true values (which is impossible to elucidate) but, rather, whether the model with estimated parameter values gives a good account of the observed data. Then, an inescapable subsequent step in the analysis of empirical data involves assessing the goodness of the fit and the most common choice for this purpose consists of computing the loglikelihood ratio statistic G^2, defined as

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} O_{ij}\log\left(O_{ij}/E_{ij}\right) \quad (3)$$

where I is the number of comparison durations, J is the number of response options, O_{ij} is the observed count of responses of type j at comparison duration i, and E_{ij} is the expected count of responses of type j at comparison duration i. Other goodness-of-fit statistics are available but G^2 is more dependable than the rest (see García-Pérez, 1994; García-Pérez & Núñez-Antón, 2001, 2004a, 2004b; see also Cressie & Read, 1989), and more importantly, G^2 is congruent with the metric under which maximum-likelihood estimates are obtained (Bishop, Fienberg, & Holland, 1975, p. 126). In any case, the type of problem addressed here also affects alternative goodness-of-fit statistics in the chi-squared family.

In Eq. 3, observed counts O_{ij} are those that gave rise to the proportions plotted in Figure 1 and expected counts E_{ij} are the counts that should have been observed under the model, which render the ordinates of the fitted curves at the comparison durations. Specifically, for each comparison duration x_i ($1 \le i \le I$) at which data were collected and for each possible response type j ($1 \le j \le J$), expected counts are defined as $E_{ij} = n_i\widehat{\Psi}_j(x_i)$, where n_i is the number of trials in which the comparison duration was x_i and the hat over Ψ indicates that the model psychometric function is evaluated using estimated values for its parameters. Summands for which $O_{ij} = 0$ in Eq. 3 are not computable but they are well defined as 0 by continuity. Figure 2 displays observed and expected counts for the binary and the ternary cases plotted in the top row of Figure 1. The resultant values of G^2 are indicated on the right side of each panel in Figure 1.

Asymptotic derivation proves that G^2 has a chi-squared distribution on $(J-1)\times I-P$ degrees of freedom, where P is the number of estimated parameters. The factor $J-1$ in this expression stems from the fact that the data at each x_i have a multinomial distribution over J categories so that there are only $J-1$ degrees of freedom in each of them. The factor I stems from the facts that there is one such multinomial distribution at each x_i and that they are independent from one another. Specifically, the component of G^2 coming from the multinomial distribution at each x_i is asymptotically distributed as a chi-squared variable with $J-1$ degrees of freedom and the sum of I such independent components is asymptotically distributed as a chi-squared variable with the sum of their individual degrees of freedom (Forbes et al., 2011, p. 72). Finally, P degrees of freedom are subtracted because P model parameters constrain the parameters of the ensemble of multinomial distributions.

For the binary task in Figure 1a, where $J = 2$, $I = 13$, and $P = 2$, there are $(J-1) \times I-P = 11$ degrees of freedom and the corresponding p value for the G^2 statistic is .212 in the top panel and .808 in the bottom panel. Thus, the model is not rejected at the usual $\alpha = .05$ in any of the two cases. For the ternary task in Figure 1b, where $J = 3$, $I = 13$, and $P = 3$, there are $(J-1) \times I-P = 23$ degrees of freedom and the corresponding p value for the G^2 statistic is .089 in the top panel and .476 in the bottom panel. Again, the model is not rejected at $\alpha = .05$ in any of the two cases. One would thus be entitled to conclude that both the binary and the ternary models give a fitting account of the applicable data with the parameter values estimated in each case.

The problem: discrepancy between actual and nominal distributions of G^2

The validity of parametric statistical inference relies on the validity of the assumptions from which the asymptotic distribution of a test statistic is derived. In the case of G^2 as defined in Eq. 3, the implied assumptions are that the I multinomial distributions are independent

https://doi.org/10.1017/SJP.2025.1 Published online by Cambridge University Press

Comparison duration (ms)

(a) Binary, observed

	100	125	150	175	200	225	250	275	300	325	350	375	400
1st longer	20	20	20	20	19	20	7	9	3	1	0	0	0
2nd longer	0	0	0	0	1	0	13	11	17	19	20	20	20

(b) Binary, expected

	100	125	150	175	200	225	250	275	300	325	350	375	400
1st longer	20.00	20.00	19.99	19.88	19.27	17.15	12.69	7.03	2.69	0.67	0.11	0.01	0.00
2nd longer	0.00	0.00	0.01	0.12	0.73	2.85	7.31	12.97	17.31	19.33	19.89	19.99	20.00

(c) Ternary, observed

	100	125	150	175	200	225	250	275	300	325	350	375	400
1st longer	19	20	18	17	10	3	4	0	1	0	1	0	0
2nd longer	0	0	0	0	1	3	8	16	18	20	18	19	20
Equal	1	0	2	3	9	14	8	4	1	0	1	1	0

(d) Ternary, expected

	100	125	150	175	200	225	250	275	300	325	350	375	400
1st longer	19.77	19.16	17.61	14.70	10.62	6.37	3.07	1.16	0.34	0.08	0.01	0.00	0.00
2nd longer	0.00	0.03	0.16	0.62	1.88	4.44	8.29	12.61	16.23	18.48	19.53	19.89	19.98
Equal	0.22	0.81	2.23	4.68	7.50	9.20	8.65	6.23	3.43	1.44	0.46	0.11	0.02

Figure 2. Observed and expected counts of responses of each type at each comparison duration in the binary and ternary tasks. Observed counts are plotted in panels in the top row of Figure 1 as proportions over the 20 trials at each comparison duration. Expected counts are similarly plotted as proportions in the corresponding panels of Figure 1 in the form of the ordinate of the applicable psychometric function at each comparison duration. A red background identifies cells with expected counts of zero.

from one another, that each of them has J categories, and that P parameters are actually estimated. If these assumptions do not hold, either G^2 does not have a chi-squared distribution or the distribution does not have the computed degrees of freedom and, then, the conclusion that the model fits or does not fit the data is suspect. Apparently, all of these assumptions hold in our illustrative scenarios with the binary and ternary tasks. Specifically, data were collected at all of the I comparison durations, the applicable number J of responses were allowed under each task, and the P parameters are identifiable.[3]

Luckily, it is very easy to check out what the actual distribution of a test statistic is in the scenario on hand. All that it takes is to run a simulation in which the null hypothesis is true, generate multiple replicates of data, compute the sample value of the test statistic for each replicate, and finally plot a histogram of the sample values ideally alongside a plot of the asymptotic distribution that the test statistic should have.

A simulation study of this type was conducted with 150,000 replicates for each task. Data (i.e., observed counts) for each replicate were generated from the applicable model (i.e., Eq. 1 for the binary task and Eq. 2 for the ternary task) so that the null hypothesis is true in the simulation. Standard duration and comparison durations were the same as in Figure 1. Each replicate used a different set of parameter values that were randomly drawn. For both tasks, parameter β was drawn from a uniform distribution on either [0.030, 0.060] or [0.006, 0.012] to mimic the relative coverage of comparison durations depicted across the rows of Figure 1. Also as in Figure 1, the number n_i of trials at each comparison duration was 20 when β ranged between 0.030 and 0.060 and 40 when β ranged between 0.006 and 0.012. For the binary task, parameter δ was drawn from a uniform distribution on $[-0.25, 0.25]$. For the ternary task, parameter δ_1 was drawn from a uniform distribution on $[-1.2, -0.2]$ and parameter δ_2 was drawn from a uniform distribution on [0.2, 1.2]. Data from each replicate under each task were subjected to the same analysis illustrated in Figure 1, namely, obtaining maximum-likelihood parameter estimates and computing the value of the G^2 statistic.

Figure 3 shows the resultant distributions (histograms) of the G^2 statistic for the binary (left column) and ternary (right column) tasks alongside the asymptotic chi-squared distribution with the applicable degrees of freedom in each case (red curves). The top row depicts the scenario in which the range of comparison durations is broad relative to the width of the psychometric functions as determined by the values of parameter β across replicates (as in the top row of Figure 1); the bottom row depicts the scenario in which the range of comparison durations is narrow relative to the width of the psychometric functions as determined by the values of parameter β across replicates (as in the bottom row of Figure 1).

Clearly, G^2 does not always follow its asymptotic distribution, despite the fact that all of the implied assumptions seem to hold. At $\alpha = .05$, the rejection rate of the true null across replicates should be close to 5%, but the rejection rate computed as the percentage of replicates in which the value of G^2 exceeded the critical point under the asymptotic distribution (indicated by arrows in the panels of Figure 3) is, in the top row, a measly 0.25% in the binary task and an even lower 0.07% in the ternary task. In contrast, empirical rejection

[3] A model is not identifiable when some of its nominal parameters mix up with others under transformations so that multiple sets of parameter values produce identical model functions. In these cases, P should be the number of identifiable parameters and not the number of nominal parameters. For examples of unidentifiable models, see Crowther, Batchelder, and Hu (1995); Rammsayer and Ulrich (2001); García-Pérez and Alcalá-Quintana (2015); Heller (2017); or Qarehdaghi and Rad (2024).

https://doi.org/10.1017/SJP.2025.1 Published online by Cambridge University Press

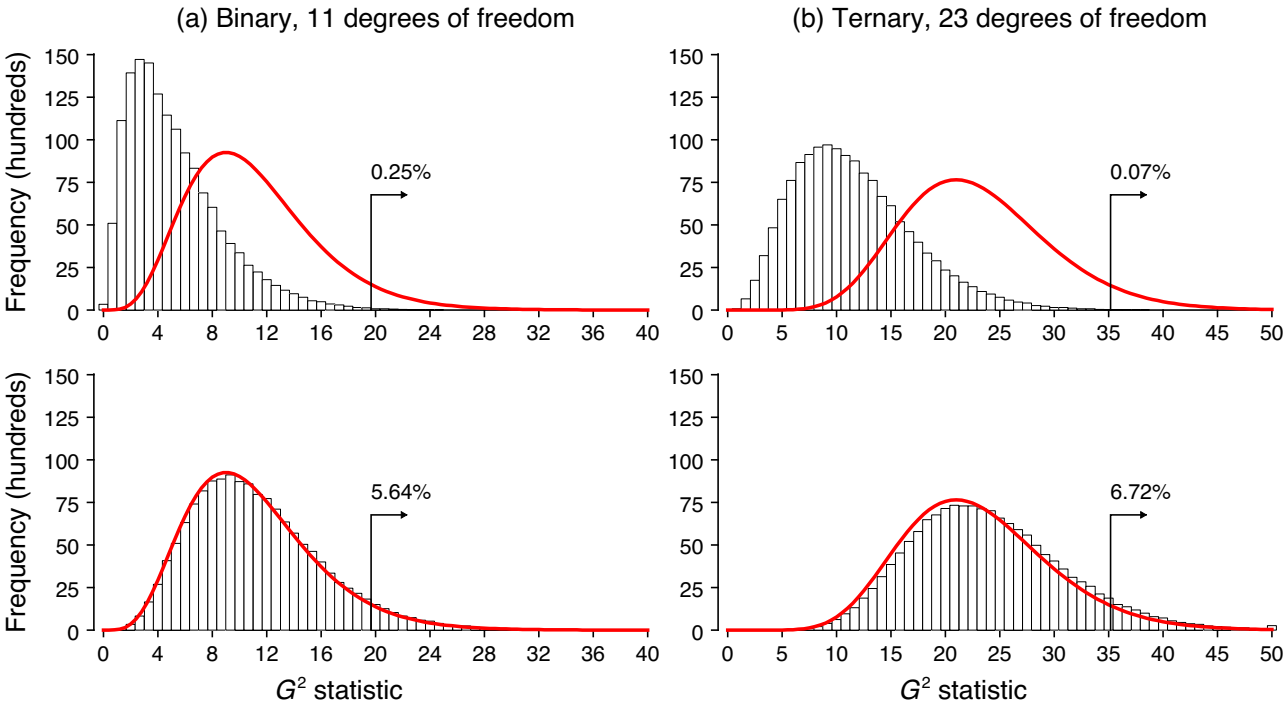

Figure 3. Actual (histograms) and asymptotic (red curves) distributions of G^2 in the binary (left panels) and ternary (right panels) tasks under the simulation conditions stated in the text. The nominal number of degrees of freedom is indicated at the top of each column. The arrow in each panel is horizontally located at the critical point for a size-.05 goodness-of-fit test, and the numeral above it is the percentage of replicates in which the value of G^2 exceeded this critical point. The top and bottom rows pertain, respectively, to simulations in which comparison durations had the coverage depicted in the top and bottom rows of Figure 1.

rates in the bottom row of Figure 3 seem to be inconsequentially different from the target 5%, although minor mismatches between actual and asymptotic distributions can be observed.

These occasional discrepancies between actual and asymptotic distributions of goodness-of-fit statistics for psychometric functions have been reported earlier (see, e.g., Wichmann & Hill, 2001).[4] Wichmann and Hill attributed it to a failure of asymptotic theory to apply in small-sample situations. Asymptotic results indicate the behavior of a test statistic as sample size goes to infinity and such results may not hold with small samples. The sample size that matters here is the number n_i of trials administered at each comparison duration (i.e., 20 in the top row of Figure 3 or 40 in the bottom row), because this determines how often each of the I multinomial distributions is sampled. A failure to reject a true null the percentage of times that α implies may naively be regarded as a blessing of the statistical test to the extent that rejections should never occur when the null is true. Yet, the downside is that this behavior is inextricably accompanied by a failure to detect instances in which the null is false. In any case, Wichmann and Hill stated that discrepancies between the asymptotic and actual distributions of goodness-of-fit statistics are unpredictable and they recommended the use of bootstrap methods to assess goodness of fit. The next section shows that discrepancies have identifiable and tractable origins that, in retrospect, are also easy to understand.

[4]Wichmann and Hill (2001) did not report results for the G^2 statistic but for the deviance D, which is a close relative to G^2 that has the same asymptotic distribution. We will continue to refer to their results despite our focus on G^2 because both goodness-of-fit measures display identical differences between actual and asymptotic distributions.

The reasons for the discrepancies

Figure 3 illustrated scenarios in which the asymptotic distribution fails to capture the actual distribution of G^2 (top row in Figure 3) and others in which no major discrepancy is observed (bottom row of Figure 3). The first part of this section discusses crucial differences between the two scenarios, the factor responsible for the misbehavior depicted in the top row of Figure 3, and corrective action to restore the agreement between actual and asymptotic distributions. The second part of this section discusses another scenario not illustrated thus far and for which the only possible corrective action is to design data collection appropriately.

Expected counts at or very near zero

Histograms in the bottom row of Figure 3 reveal that the asymptotic distribution captures the actual distribution of G^2 in the conditions under which the data were generated, although the relevant characteristics of those conditions may seem mysterious at this point. On the other hand, the conditions under which the histograms in the top row of Figure 3 were generated must have some contrasting characteristics that determine the failure of the asymptotic distribution to capture the actual distribution of G^2. We must stress at this point that the relevant factor is not the number n_i of trials per comparison duration (which was 20 in the top row of Figure 3 and 40 in the bottom row), but this will only become evident below. In any case, histograms in the top row of Figure 3 do not have a disorderly shape and, at first glance, it looks as if they also reflected chi-squared distributions in which the numbers of degrees of freedom differed from those seemingly applicable under the circumstances. In fact, it seems as if some degrees of freedom were lost somewhere along the line.

https://doi.org/10.1017/SJP.2025.1 Published online by Cambridge University Press

A close look at Figure 2 reveals that the expected counts for some cells are zero (see the cells shaded in red), which is an indication that the corresponding multinomial distribution with J response categories degenerates under the estimated parameter values: Expected counts are zero for categories that have a probability of zero, which certainly indicates that the corresponding multinomial has actually fewer than its nominal J categories. This might account for the apparent loss of degrees of freedom and suggests that the number J_i of response categories under each of the I multinomials has to be reassigned after parameter estimation. Each expected count of zero thus subtracts one degree of freedom. Then, for the particular example in the top row of Figure 1, the actual number of degrees of freedom for the binary data would be 8 (not the nominal 11) and the actual number for the ternary data would be 20 (not the nominal 23). Note that the computation of the actual number of degrees of freedom has to wait until parameters are estimated for each replicate and the resultant number of cells with expected counts of zero can be determined. Thus, the actual number of degrees of freedom would vary across replicates in each panel in the top row of Figure 3, suggesting that the histograms reflect a mixture of chi-squared distributions with different numbers of degrees of freedom. Interestingly, none of the replicates in the bottom row of Figure 3 ever resulted in any expected count of zero, which may be a sign that the nominal number of degrees of freedom is appropriate in such a case.

An obstacle to the preceding argument arises because, strictly speaking, a probability of exactly zero is impossible for any response category under the models of Eqs. 1 and 2, no matter what values the estimated parameters have. The reason is that psychometric functions have lower asymptotes at zero, implying that they can never evaluate to zero. The immediate question under the circumstances is at what small value, if any, a probability might be assimilated to zero. Furthermore, a very small probability can turn into a sizeable expected count if the number of trials is sufficiently large, which displaces the question from magnitude of probabilities to magnitude of expected counts. Since there is no straightforward answer to this question, we conducted simulations to determine the threshold expected count that warrants subtraction of degrees of freedom, as described next.

Because parameter estimation is time-consuming, the simulation assessed goodness of fit to the generating psychometric functions, as is usual in studies of sampling variability. In this situation, $P = 0$ and the nominal number of degrees of freedom is simply $(J-1) \times I$. Across simulation conditions, J was either two (binary task) or three (ternary task), I varied from 10 to 14, the I comparison durations x_i were always evenly spaced between 100 ms and 400 ms (with a standard of 250 ms), and the number n_i of trials administered at each x_i varied from 20 to 40 in steps of 10.[5] The number of replicates in each simulation condition was 300,000, and true parameter values for each replicate were drawn from a uniform distribution on [0.03, 0.06] for β (for both tasks), from a uniform distribution on [−0.25, 0.25] for δ in the binary task, and from uniform distributions on [−1.2, −0.2] and [0.2, 1.2] for δ_1 and δ_2 in the ternary task, respectively. Note that these parameter ranges are those under which the failure of asymptotics was observed in the top row of Figure 3. An additional and crucial factor in the simulations was the threshold T used to subtract one degree of freedom for each expected count that fell below T, with values of T ranging from 0.01 to 0.14 in steps of 0.01.

Thus, in each simulation condition where the nominal number of degrees of freedom was df = $(J-1) \times I$, each replicate turned up with an alternative number of degrees of freedom on application of the threshold T. Specifically, this alternative number of degrees of freedom was defined as $df^* = (J-1) \times I - K_T$, where K_T is the number of expected counts lower than T across the $I \times J$ cells in each replicate. The overall set of 300,000 replicates in each condition was subsequently split into groups for which df^* attained the same value at any given T. Within each of these groups, the analysis proceeded as illustrated in Figure 3 above, that is, (1) the actual distribution of G^2 (a histogram) within the group was compared with the probability density function of a chi-squared variable with df^* degrees of freedom and (2) the rejection rate was computed. If the loss of K_T degrees of freedom is responsible for the discrepancy between the overall distribution of G^2 and a chi-squared distribution with df degrees of freedom, then G^2 should follow a chi-squared distribution with df^* degrees of freedom within each of the groups defined according to the value of df^*. The goal of the simulation was to find out the value of T that eliminates the discrepancy, if any.

Figure 4 shows an illustrative example of this strategy by plotting results for the condition involving a ternary task (i.e., $J = 3$) with $I =$ 13 comparison durations, 20 trials per comparison, and threshold $T = 0.06$. The panel at the top left shows the overall distribution of G^2 (histogram) across the 300,000 replicates along with the probability density of a chi-squared variable with the nominal 26 degrees of freedom (red curve). This condition is equivalent to using $T =$ 0 because expected counts cannot be negative-valued and, thus, $K_0 = 0$ in the computation of df^*. These results reproduce the characteristics in the top-right panel of Figure 3, indicating that the disagreement between actual and nominal distribution of G^2 also occurs when parameters are not estimated from the data. The remaining panels in Figure 4 show analogous results within each of the groups of replicates for which application of the threshold $T =$ 0.06 resulted in the same value for df^*. It is immediately apparent that consideration of the number K_T of degrees of freedom that are lost due to small expected counts restores the agreement between the actual distribution of G^2 and its asymptotic distribution when the latter uses the actual number of degrees of freedom in each replicate (i.e., df^* instead of the nominal df). This analysis also reveals the reason for the discrepancy in the top-left panel in Figure 4: The overall distribution of G^2 is a mixture of several chi-squared distributions each with a different number of degrees of freedom. Analogous results were obtained for the binary task (i.e., $J = 2$), and we omit presentation of these results in the graphical form of Figure 4.

Our choice of the simulation condition for which to plot results in Figure 4 was not blind, particularly in what regards the value $T =$ 0.06. An overall picture of how the value of T affects rejection rates and their potential departure from the nominal α can easily be obtained. Consider summarizing the results in Figure 4 as follows. First, there is a rejection rate indicated in Figure 4 for each group of replicates and they vary from 6.65% when $df^* = 10$ to 4.23% when $df^* = 20$. Each of these rejection rates reflects a given number of

[5]Both the range of number I of stimulus levels and the range of number n_i of trials at each level that were chosen for these and other simulations in this study cover the most common decisions made by researchers estimating psychometric functions in empirical studies. This is actually the reason that similar ranges have been used in other simulation studies aimed at assessing the effects of methodological or procedural decisions in the design of protocols for the estimation of psychometric functions (see, e.g., García-Pérez, 2014; García-Pérez & Alcalá-Quintana, 2005; Lam, Dubno, & Mills, 1999; Lam, Mills, & Dubno, 1996; Leek, Hanna, & Marshall, 1992; Miller & Ulrich, 2001; and O'Regan & Humbert, 1989).

https://doi.org/10.1017/SJP.2025.1 Published online by Cambridge University Press

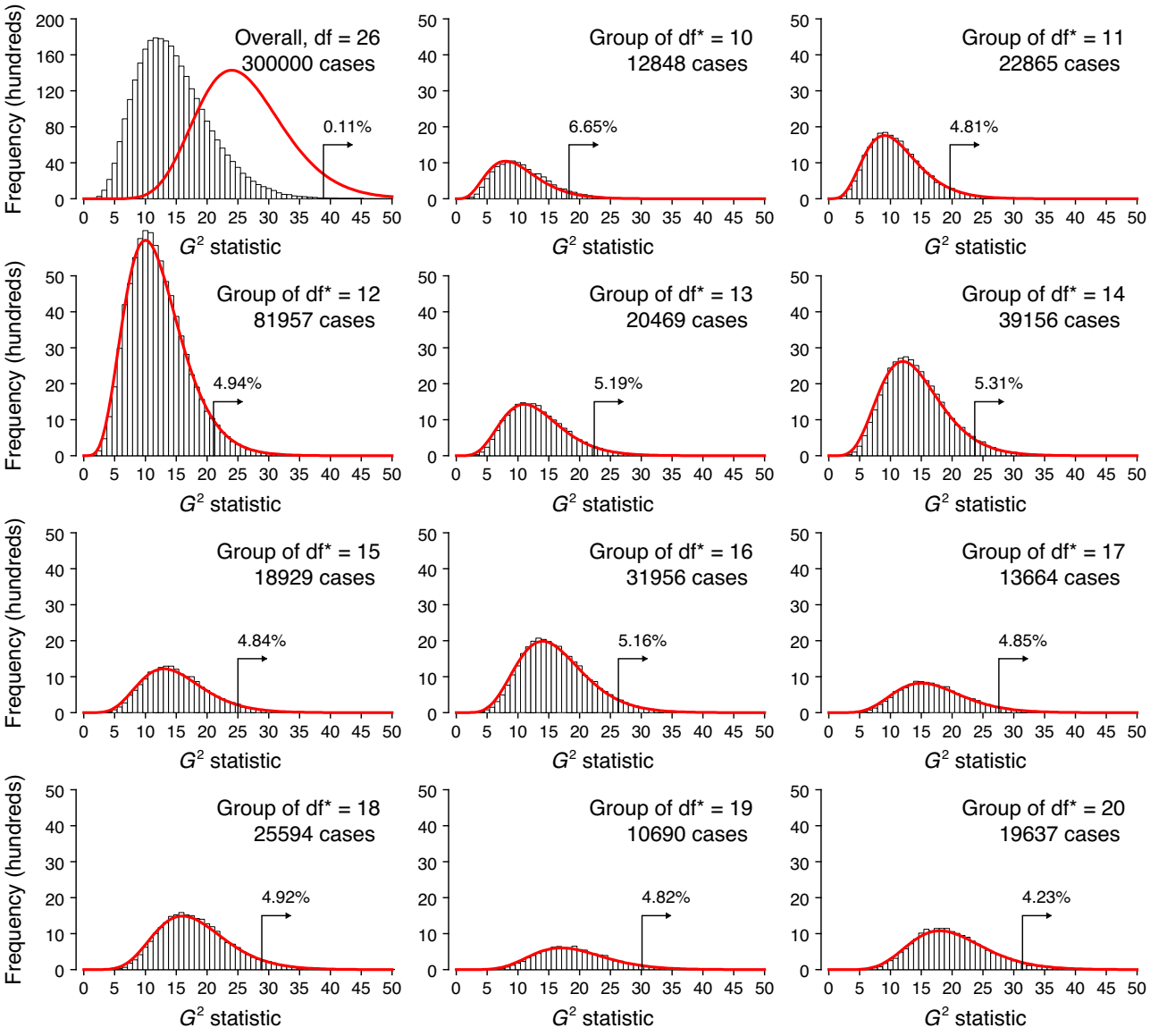

Figure 4. Actual (histograms) and asymptotic (red curves) distributions of G^2 from the simulation condition involving $J = 3$ (ternary task), $I = 13$ (13 comparison durations), and $n_i = 20$ (20 trials at each comparison duration) with no parameter estimation. The top-left panel shows the overall distribution and the asymptotic distribution with df = (3−1) × 13 = 26 degrees of freedom. The remaining panels show analogous results after determining the value of df* for each replicate by subtracting one degree of freedom for each expected count below $T = 0.06$. Each panel shows the distribution for the subset of replicates for which df* had the same value, for values between df* = 10 and df* = 20. The number of replicates and the value of df* for each group are given at the top right side of each panel. Two additional groups are omitted because the number of replicates that fell into them was too few (1999 replicates with df* = 21 and 236 replicates with df* = 22). Rejection rates and the critical point for a size-.05 test with the degrees of freedom in each panel are indicated by an arrow and the numeral above it.

rejections in each group and the overall rejection rate is the sum of the number of rejections in each group divided by the total number of replicates in the simulation. This overall rejection rate turns out to be 5.02% for the simulation condition in Figure 4, where $T = 0.06$. Similar rejection rates can be computed for the same simulation condition at other values of T, and the process can be repeated for the remaining simulation conditions. Figure 5 shows these overall rejection rates as a function of the value of T for each simulation condition. The value $T = 0.06$ stands out as that for which overall rejection rates are virtually at the nominal 5%, something that holds very precisely whether $J = 2$ or $J = 3$ (i.e., binary or ternary tasks), for all $I \in \{10, 11, 12, 13, 14\}$, and for all sample sizes defined as the number n_i of trials per comparison duration (between 20 and 40 in steps of 10). Although these combinations cover sufficiently well the breadth of sampling plans in psychophysical studies where psychometric functions are fitted to data, one may wonder whether the appropriateness of $T = 0.06$ generalizes to other combinations of values for I and J, or with other sample sizes. This is something that cannot be explored exhaustively but the convergence of results across the panels in Figure 5 seems to rule out an anecdotal coincidence. Nevertheless, we used a cross-validation approach to test the notion that the threshold $T = 0.06$ determines the actual number of degrees of freedom and, thus, reconciles G^2 with an asymptotic distribution.

The top row in Figure 3 plotted the overall distribution of G^2 when model parameters were estimated from the data in a binary task in which $P = 2$ parameters and a ternary task in which $P = 3$ parameters. In both tasks there were 13 comparison durations (i.e., $I = 13$) and 20 trials per comparison (i.e., $n_i = 20$ for all i). The

https://doi.org/10.1017/SJP.2025.1 Published online by Cambridge University Press

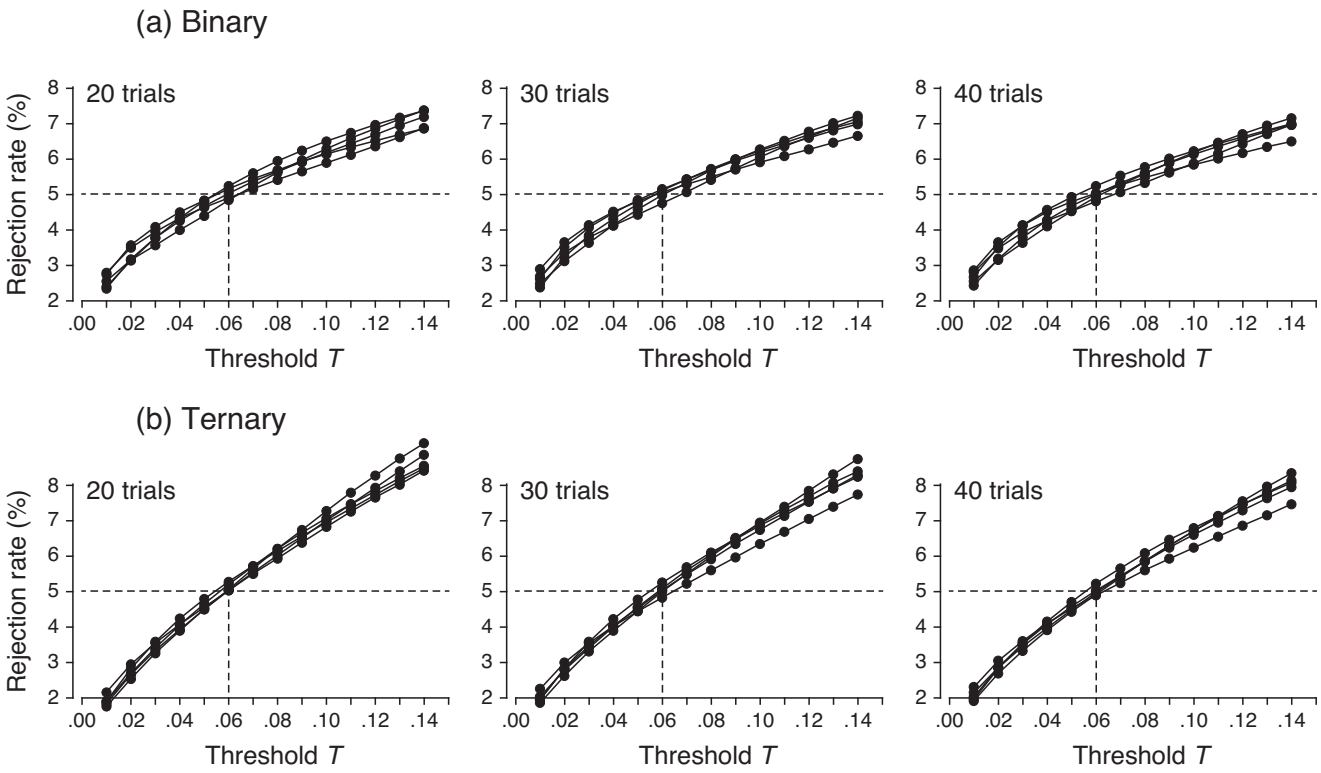

Figure 5. Overall rejection rates in the binary (a) and ternary (b) tasks with no parameter estimation as a function of the value T of the threshold used to determine the number K_T of cells with small expected counts. Each panel shows results for the number n_i of trials indicated at the top left of the panel. The five (unmarked) curves in each panel show results for each number I of comparison durations (between 10 and 14). Note that $T = 0.06$ produces rejection rates that are virtually at the target 5% level in all cases.

results indicated discrepancy with the asymptotic chi-squared distribution with nominal df = 11 (binary task) or df = 23 (ternary task). We submitted those data to the same analyses illustrated in Figures 4 and 5, and we assessed overall rejection rates when each individual replicate is tested against the critical point of a chi-squared distribution with df*. The value of df* was again computed for each replicate by subtracting (from the nominal df) one degree of freedom for each expected count that fell below T, and we also varied here the value of T from 0.01 to 0.14 in steps of 0.01. Figure 6 shows the results with the same graphical conventions of Figure 5. Again, $T = 0.06$ gives an overall rejection rate that is meaninglessly different from the nominal 5% and, then, with an allowance of two decimal places, $T = 0.06$ also comes out as the proper threshold when model parameters need to be estimated from the data.

Interestingly, we mentioned earlier that the simulations in which the distribution of G^2 was plotted in the bottom row of Figure 3 did not result in any expected count of zero for any of the replicates. More precisely, K_T was zero for all replicates at all T, so that df* = df for all replicates. Thus, the nominal number of degrees of freedom holds for each replicate and the asymptotic distribution describes adequately the actual distribution of G^2 (see the bottom row in Figure 3), except perhaps for a minor quirk that is more apparent in the ternary task and whose origin will be addressed in the next section.

Having established that the actual number of degrees of freedom for the chi-squared distribution of G^2 is df* = $(J-1) \times I - P - K_{0.06}$, a look back at Figure 2 reveals that, in the binary task with $P = 2$, $K_{0.06} = 5$, and df* = 6 the resultant $G^2 = 14.39$ (see the top panel in Figure 1a) has an associated p value of .026. Analogously, in the ternary task with $P = 3$, $K_{0.06} = 6$, and df* = 17, the resultant $G^2 = 32.56$ (see the top panel in Figure 1b) has an associated p value of .013. Thus, contrary to what the asymptotic distributions with 11 or 23 degrees of freedom render, the model is rejected for these data in both tasks at the usual $\alpha = .05$. As for the sample cases in the bottom row of Figure 1, $K_{0.06} = 0$ in both cases, and thus, the p value associated with $G^2 = 6.89$ in the binary task where $P = 2$ and df* = df = 11 is 0.808, whereas the p value associated with $G^2 = 22.74$ in the ternary task where $P = 3$ and df* = df = 23 is 0.476. Thus, the model is not rejected for these data in any task at the usual $\alpha = .05$.

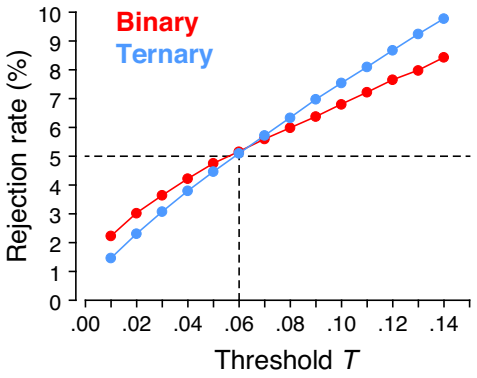

Figure 6. Overall rejection rates in the binary (red) and ternary (blue) tasks with parameter estimation as a function of the value T of the threshold used to determine the number K_T of cells with small expected counts. Data come from the simulation for which overall distributions of G^2 were plotted in the top row of Figure 3. Note that $T = 0.06$ also produces rejection rates that are closest to the nominal 5% level here.

Insufficient number n_i of trials at each x_i

In their Figure 7, Wichmann and Hill (2001) illustrated another form of disagreement in which the actual distribution of the

https://doi.org/10.1017/SJP.2025.1 Published online by Cambridge University Press

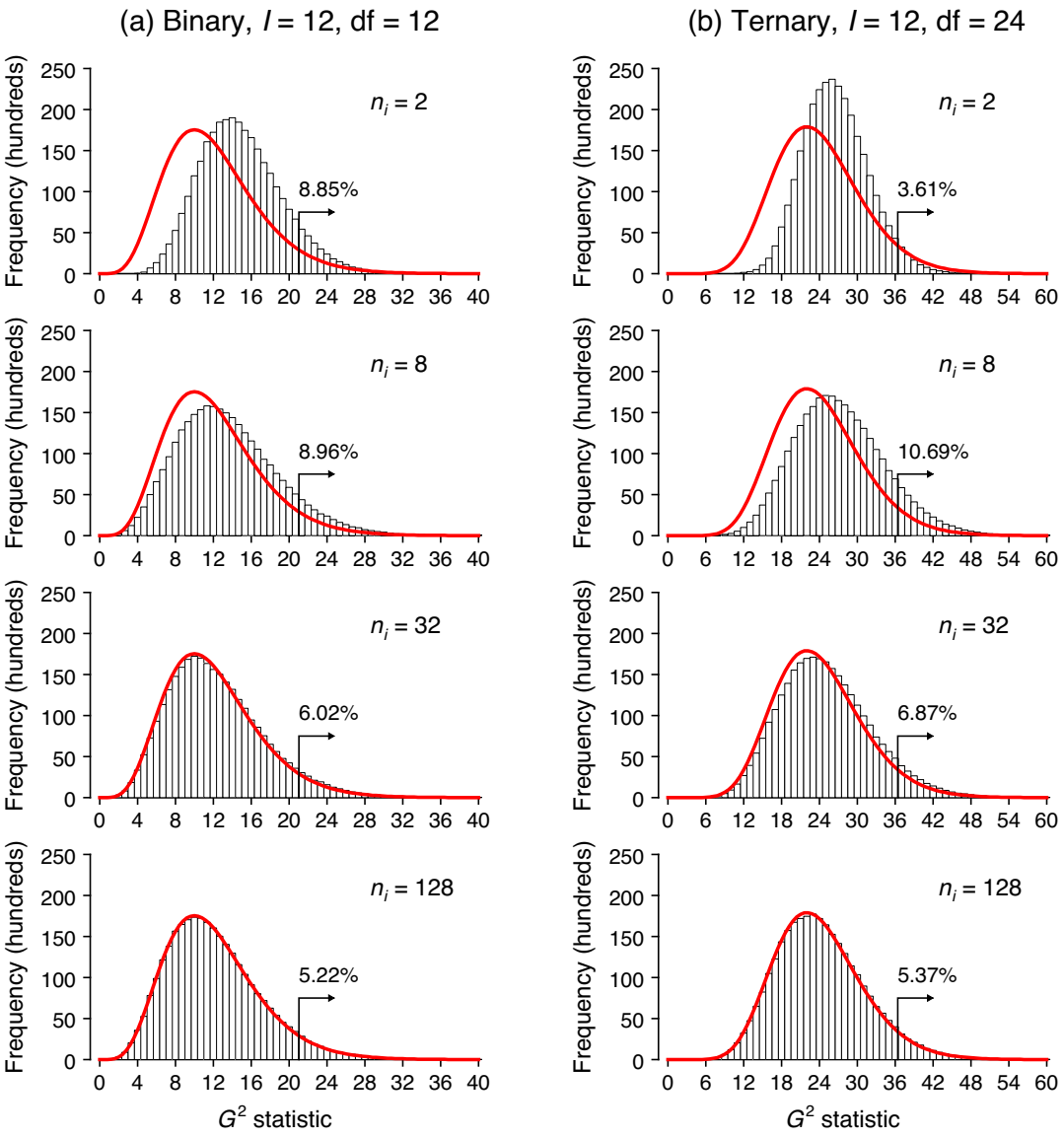

Figure 7. Actual (histograms) and asymptotic (red curves) distributions of G^2 in binary (left column) and ternary (right column) tasks from the simulation condition involving $I = 12$ (i.e., 12 comparison durations) and n_i ranging from 2 (top row) to 128 (bottom row) in multiplicative steps of 4, and with no parameter estimation. Each panel shows the distribution for the set of 300,000 replicates and the asymptotic chi-squared distribution with the applicable degrees of freedom (indicated by the value of df at the top of each column). Rejection rates and the critical point for a size-.05 test with the degrees of freedom in each panel are indicated by an arrow and the numeral above it.

deviance D is displaced to the right of its asymptotic distribution, that is, a displacement in the opposite direction to that seen in the top row of Figure 3. Their simulation conditions involved a binary task (i.e., $J = 2$) with $I = 60$ or $I = 240$ and only $n_i = 2$ trials placed at each x_i. The putative psychometric function was such that the generating probabilities of success in the binary task were evenly spaced between .52 and .85 across the I comparison levels. They conducted the simulation without estimating parameters, and in these conditions, the smaller expected counts range between $2 \times .15 = 0.30$ and $2 \times .48 = 0.96$, both of which are far above the threshold $T = 0.06$ discussed in the preceding section. Thus, the actual df* would match the nominal df and, in any case, subtracting df would instead push the actual distribution of G^2 to the left (see the top row in Figure 3). By analogy with the situation described in the preceding section, one might think that the culprit here is an increase in the number of degrees of freedom but it is hard to identify where the extra degrees of freedom might come from.

We reproduced the results of Wichmann and Hill (2001) in a simulation with G^2 that avoided parameter estimation and simply assessed goodness of fit to the generating psychometric functions, as we did to present results in Figures 4 and 5. Thus, in these simulations $P = 0$ and the nominal number of degrees of freedom was simply $(J-1) \times I$. Across simulation conditions, J was either two (binary task) or three (ternary task), I varied from 10 to 14, and the I comparison durations x_i were evenly spaced between 100 ms and 400 ms (with a standard of 250 ms). As in our preceding simulations, the number of replicates in each condition was 300,000 and true parameter values for each replicate were drawn from a uniform distribution on [0.006, 0.012] for β (for both tasks), from a uniform distribution on [−0.25, 0.25] for δ in the binary task, and from uniform distributions on [−1.2, −0.2] and [0.2, 1.2] for δ_1 and δ_2 in the ternary task, respectively. Note that the spread of values for β is that for which expected counts lower than 0.06 were not observed in our

https://doi.org/10.1017/SJP.2025.1 Published online by Cambridge University Press

previous simulations, and the reason for this choice is that we wanted to leave out of the picture any further contamination coming from small expected counts. The main independent variable in these simulations was the number n_i of trials administered at each x_i, which varied from 2 to 128 in multiplicative steps of 2.

Figure 7 shows the distributions of G^2 for the binary (left column) and ternary (right column) tasks with $I = 12$ at $n_i \in \{2, 8, 32, 128\}$ (rows). Distributions for other conditions were analogous and displayed the same pattern of increasing agreement with the asymptotic distribution as the number n_i of trials at each x_i increased. Clearly, discrepancies at small n_i are due to pushing asymptotic theory too far and the reason is also easy to identify.

Recall that the sample size that matters here is the number n_i of trials with which each of the I multinomials is sampled. Asymptotic theory holds (or fails to hold) for the components of G^2 coming from each of the I multinomials, regardless of the number I of multinomials that are aggregated into the omnibus G^2. For an illustration of the problem, consider the binary case and a single comparison level representing an individual binomial component with probability of success p so that the expected counts are $(E_1, E_2) = n \times (p, 1-p)$. With $n = 2$, the sample space consists of only three possible vectors of observed counts (O_1, O_2), namely, (0, 2), (1, 1), and (2, 0). From Eq. 3 with $I = 1$ (because we are considering only one of the I binominals), the values that G^2 can attain are $4\log\left(\frac{1}{1-p}\right)$ for $(O_1, O_2) = (0, 2)$, $2\log\left(\frac{1}{2p}\right) + 2\log\left(\frac{1}{2(1-p)}\right)$ for $(O_1, O_2) = (1, 1)$, and $4\log\left(\frac{1}{p}\right)$ for $(O_1, O_2) = (2, 0)$. Thus, this component of G^2 can only attain three values (or just two if $p = .5$) and, yet, its extremely discrete distribution is regarded as if it were a (continuous) chi-squared distribution with one degree of freedom. For $J = 3$ (i.e., the ternary case), the discreteness of the distribution of each individual component of G^2 worsens the problem when $n = 2$, because this forces one of the observed counts to be zero. Aggregating such discrete distributions over several $I > 1$ (e.g., $I = 12$ as in the top row of Figure 7) cannot fix a problem that lies within each of the component multinomials.

Obviously, as n_i increases (i.e., down the columns of Figure 7), the sample space for each of the components of G^2 increases in size and makes their (individual) asymptotic chi-squared distribution with $J-1$ degrees of freedom more attainable. This carries over to the chi-squared distribution of their aggregation into the omnibus G^2. The immediate question is, then, what is the minimum number n_i of trials that must be administered at each of the I multinomials for dependable goodness-of-fit testing. As usual, an answer of the one-size-fits-all kind does not exist. For instance, note in the third row of Figure 7 that $n_i = 32$ seems more appropriate in the binary case than it seems in the ternary case. Traces of this dependence on the value of J were also apparent in the bottom row of Figure 3, where $n_i = 40$ rendered a reasonably close asymptotic approximation to the actual distribution of G^2 in the binary task (left panel) but not so much in the ternary task (right panel). Based on the analyses reported here and other analyses of the same type, $n_i \geq 40$ (if $J = 2$) or $n_i \geq 50$ (if $J = 3$) seem sufficient, although smaller numbers of trials still provide reasonably close approximations. Nevertheless, simulations can always be run to investigate behavior under alternative scenarios, as shown next.

A scenario in which n_i is constant at all $1 \leq i \leq I$ defines what is known as the psychophysical method of constant stimuli, whose designation aims at stressing that each comparison level is paired with the standard level the same number of times. This sampling plan is inefficient because placing n_i trials at some comparison levels seems too much, whereas also placing n_i trials at other comparison levels seems too little, further resulting in poorer estimates of the parameters of the psychometric function (see, e.g., García-Pérez & Alcalá-Quintana, 2005; Watson & Fitzhugh, 1990). In the interest of efficiency, an overall number N of trials is often deployed with adaptive methods that distribute them unevenly such that n_i varies across the I comparison levels to sample the psychometric function more reasonably (for comparative examples of both types of sampling plan, see Figures 9 and 10 in García-Pérez, 2014). In the context of this study, the question arises as to how the characteristics illustrated in Figure 7 for constant n_i vary when the same overall number N of trials is adaptively and unevenly distributed such that n_i is no longer constant across the I comparison levels.

We sought the answer to this question by running a simulation identical to that for which results were reported in Figure 7 except that the N trials in each condition were now deployed adaptively for each replicate. Given that $I = 12$ and n_i varied from 2 to 128 for the simulations in Figure 7, the corresponding values for N in the current simulations would range from $12 \times 2 = 24$ trials to $12 \times 128 = 1{,}536$ trials in multiplicative steps of 2. We nevertheless excluded the uninteresting first three conditions and only ran simulations for $N \in \{192, 384, 768, 1{,}536\}$. As in the simulations of Figure 7, the standard duration was $x_s = 250$ ms and the $I = 12$ potential comparison durations were evenly spaced between $x_1 = 100$ ms and $x_{12} = 400$ ms. This spacing for comparison durations was also used as the step size for the adaptive placement rules. Adaptive placement of trials for each replicate was governed by strategies typically used in empirical practice (see, e.g., García-Pérez & Alcalá-Quintana, 2020a; García-Pérez & Peli, 2014, 2019). Specifically, 24 up–down staircases were randomly interwoven each of which deployed the same number of trials (e.g., 8 trials when $N = 192$ or 64 trials when $N = 1{,}536$). Twelve of these staircases had their starting point at x_1, and the other twelve had their starting point at x_{12}. This 24-staircase setup ensures that there will be a minimum of 12 trials placed at x_1 and a minimum of also 12 trials at x_{12}, whereas the number of trials at all intermediate comparison durations will generally be much larger. When $J = 2$ (i.e., the binary task), staircases implemented the 1-down/1-up rule such that the comparison duration for the next trial was one step higher if the response on the current trial was "first longer" or one step lower if the response was "second longer." When $J = 3$ (i.e., the ternary task), the same rules were in place for "first longer" and "second longer" responses on the current trial. In addition, upon an "equal" response in the current trial, the comparison duration for the next trial was either two steps higher or two steps lower with equiprobability. If application of these rules resulted in a comparison duration lower than x_1 or higher than x_{12} for the next trial, the boundary values x_1 or x_{12} were substituted.

Figure 8 shows the results with the same graphical conventions of Figure 7. Keep in mind that adaptive deployment of, say, $N = 384$ trials (second row in Figure 8) incurs the same cost as deploying $n_i = 32$ trials at each of $I = 12$ comparison levels (third row in Figure 7). Also, the 24-staircase strategy ensures a minimum of 12 trials at x_1 and x_{12} and generally many more trials at each of the intermediate levels. Clearly, this unevenness does not seem to cause much trouble and, then, constant numbers n_i of trials at all comparison levels are not necessary for validity of goodness-of-fit inferences as long as none of the n_i is ridiculously small (say, less than 10).

https://doi.org/10.1017/SJP.2025.1 Published online by Cambridge University Press

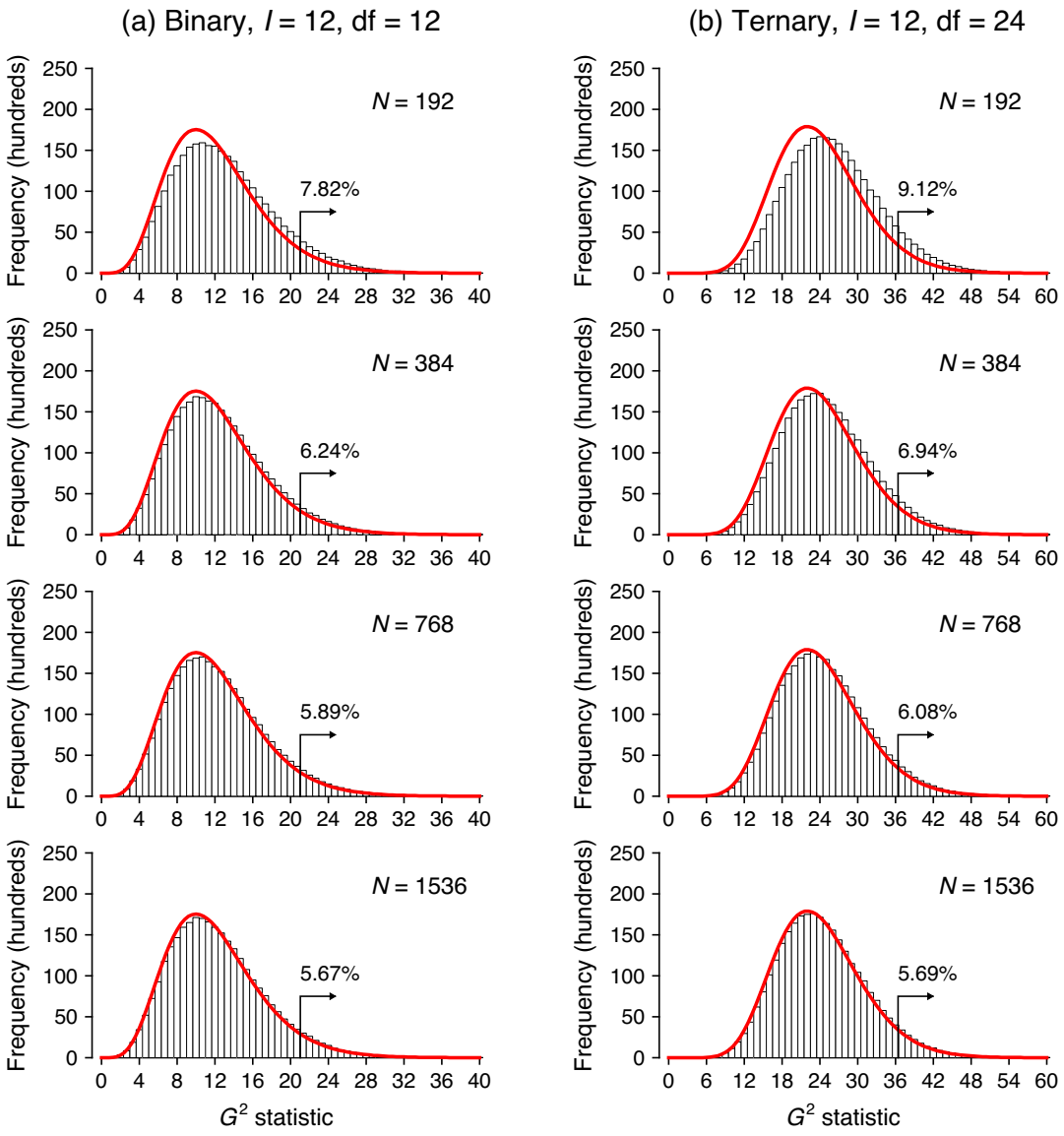

Figure 8. Actual (histograms) and asymptotic (red curves) distributions of G^2 in binary (left column) and ternary (right column) tasks with adaptive placement of N trials, ranging from 192 (top row) to 1,536 (bottom row) in multiplicative steps of 2, and with no parameter estimation. Graphical conventions as in Figure 7.

Discussion

It has often been reported that the actual distribution of G^2 in goodness-of-fit testing with categorical models differs from its asymptotic distribution, specifically in the form of the actual distribution being meaningfully displaced laterally relative to the asymptotic distribution. We corroborated the presence of these differences for testing the goodness of fit of psychometric functions, an instance of the general problem of fitting multinomial logistic-like regression models to categorical data. We have also unveiled the sources of the discrepancies, which lie in two factors. One of them is the difference between nominal and actual degrees of freedom when expected counts are zero or nearly zero, which produce leftward displacements relative to the location of the asymptotic distribution. Specifically, although the nominal number of degrees of freedom of G^2 in these applications is df = $(J-1) \times I-P$, our analyses identified that the actual number of degrees of freedom is instead df* = $(J-1) \times I-P-K_{0.06}$, where $K_{0.06}$ is the number of cells for which the expected count does not exceed 0.06. Simulations provided the evidence supporting the validity of this expression over a realistic range of values of I, J, and P. In view of the tight convergence of those results (see Figures 5 and 6), we are inclined to believe that subtraction of K_T degrees of freedom (with T = 0.06) is generally appropriate for use with G^2 in goodness-of-fit testing of categorical models although we do not have any insight on the reasons for this particular value of T and its apparent generality across the ranges of I, J, and P.

We must insist on the fact that $K_{0.06}$ can only be determined once the model has been fitted to the data on hand, because parameter estimates are needed to compute expected counts in each cell. Since the mathematical form of (and the number of parameters in) the categorical model will vary broadly across occasions, suitable custom software will still be needed by each researcher to complete the parameter estimation stage and this software cannot be replaced with a general-purpose routine (e.g., an R script) that will implement our counting-and-subtracting

https://doi.org/10.1017/SJP.2025.1 Published online by Cambridge University Press

procedure. For dependable goodness-of-fit tests via G^2, using the correct number of degrees of freedom only requires researchers to supplement their parameter estimation software with extra code that computes $K_{0.06}$ and df*.

The failure of goodness-of-fit statistics of the power-divergence family (Read and Cressie, 1988) to follow their asymptotic distribution when expected counts are small has been known for a long time. A number of studies have tried over the years to identify the conditions within which the asymptotic distribution is dependable, whereas other studies have evaluated correction terms to maximize the match between actual and asymptotic distributions. Collectively, all of these studies resulted in a relatively complex list of conditions regarding how many expected counts can be small (and how small can they be) for the asymptotic distribution to hold reasonably well (see, e.g., Delucchi, 1983). However, we are not aware of any previous attempt to identify a correction to the number of degrees of freedom in the actual distribution of G^2 that restores the validity of inferences across the board. An alternative approach consists of using bootstrap methods as advocated by Wichmann and Hill (2001), but restoring asymptotics via correcting the number of degrees of freedom is more cost-effective.

The second factor displaces the actual distribution of G^2 to the right of the asymptotic distribution, and it is caused by an insufficient number n_i of trials to sample each of the I multinomial distributions comprising computation of the G^2 statistic. This effect cannot be countered with any after-the-fact action but it can be easily circumvented at the design stage by deploying a sufficiently large numbers of trials at each x_i. It should be noted in this respect that several simulation studies have assessed parameter estimation accuracy under alternative sampling plans that incur the exact same cost, that is, scenarios in which a given total number of trials is deployed either in the form of very many trials at a small number I of stimulus levels or in the form of very few trials at a large number I of stimulus levels (see, e.g., García-Pérez & Alcalá-Quintana, 2005; Lam, Dubno, & Mills, 1999; Lam, Mills, & Dubno, 1996). At one extreme, Treutwein and Strasburger (1999) showed that using $n_i = 1$ with large I provides reasonably good parameter estimates, although they acknowledged that this choice precludes testing goodness of fit due to lack of degrees of freedom. All else being equal, sampling plans involving large values of n_i are preferable because they ensure the statistical validity of subsequent goodness-of-fit analyses. It should also be kept in mind that the minimal number n_i of trials needed for accurate parameter estimation (see, e.g., García-Pérez & Alcalá-Quintana, 2005, 2012, 2017, 2020b) may differ from the minimal number of trials needed for dependable goodness-of-fit testing and that, in addition, the overall number of trials needed for adequate model comparisons may be much larger (see, e.g., Kelber & Ulrich, 2024; Tünnermann & Scharlau, 2018, 2021). Thus, a researcher who is not just (but also) interested in testing goodness of fit must consider a number of criteria when deciding on the optimal number of trials and stimulus levels to address the research question on hand.

We must stress that computation of $(J-1) \times I-P-K_{0.06}$ as the applicable number of degrees of freedom for goodness-of-fit testing is limited to G^2 and its use with other statistics of the power-divergence family is unwarranted. We checked this out by repeating the simulations and analyses in Figures 4 and 5 for Pearson's X^2 statistic, defined as

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{4}$$

where all symbols have the same meaning as in Eq. 3. We also do not have any insight as to why the correction is useless with X^2, although the actual distribution of X^2 appears more heavily affected structurally by the presence of small expected counts and observed counts of zero. For one thing, consider a cell in which $O_{ij} = 0$ and where E_{ij} is not negligible (say, $E_{ij} > 1$). The contribution of this cell to G^2 is zero by Eq. 3, whereas its contribution to X^2 is E_{ij} by Eq. 4. Thus, cells with zero observed counts may spuriously inflate the sample value of X^2, an inflation that also comes from cells in which $O_{ij} > 0$ and E_{ij} is very small. Cells with these characteristics are relatively common in the context of fitting psychometric functions (see sample observed and expected counts in Figure 2).

We have focused on the accuracy of goodness-of-fit tests and left aside their power for reasons that are easy to describe. Estimating power curves for a statistical test requires assessing rejection rates when data are generated under an alternative hypothesis that progressively departs from the null along a quantitative dimension. The null hypothesis in goodness-of-fit tests to categorical models states that model functions with parameters estimated from the data give a good account, whereas the alternative simply states that this is not the case. Obviously, there is no quantitative dimension along which increasingly larger deviations from the null can be used to generate data. This is the reason that the power of goodness-of-fit tests is often evaluated only against fixed alternatives of interest (e.g., if Eqs. 1 or 2 should have a logistic cumulative function instead of the normal cumulative function Φ). In the context of categorical models of psychological processes, the existence of fixed alternatives of interest typically raises issues other than power, namely, model mimicry (i.e., whether alternative models are functionally equivalent; see García-Pérez & Alcalá-Quintana, 2015; Navarro, Pitt, & Myung, 2004; Starns, 2021) and model comparison and selection (i.e., which of a number of alternative models fits the data better; see Ditterich, 2015; Kelber & Ulrich, 2024; Pitt, Myung, & Zhang, 2002; Tünnermann & Scharlau, 2018, 2021).

Conclusion

When using G^2 to test the goodness of fit of psychometric functions (or, more generally, of any categorical model in a suitable scenario), statistical inferences are valid when two conditions are met. The first one is that the number of degrees of freedom is properly computed as $(J-1) \times I-P-K_{0.06}$, where J is the number of response categories in the task, I is the number of stimulus levels, P is the number of free parameters in the mathematical expression of the fitted function, and $K_{0.06}$ is the number of cells for which the expected count does not exceed 0.06 (i.e., the number of values E_{ij} in Eq. 3 that do not exceed 0.06). The second condition is that the number n_i of trials administered at comparison level x_i is not ridiculously small (i.e., not lower than 10) when n_i is the same at all comparison levels; if n_i varies across levels, the overall number N of trials must be in excess of $40 \times I$ if $J = 2$ or $50 \times I$ if $J = 3$, with no individual n_i falling below 10. Both requirements are easy to achieve in any practical situation but it should be stressed that compliance with these conditions does not solve inferential problems arising from the use of members of the power-divergence family of goodness-of-fit statistics other than G^2.

Data sharing. No data were collected for this study.

Author contribution. M. A. G.-P. and R. A.-Q involved in conceptualization of data, formal analysis, funding acquisition, methodology, and validation; M. A. G.-P. involved in software acquisition and wrote the original draft of the manuscript; and M. A. G.-P. and R. A.-Q wrote the review and edited the manuscript.

Funding statement. This work was supported by grant PID2019-110083GB-I00 from Ministerio de Ciencia e Innovación.

Conflicts of interest. None.

References

Bishop, Y. M. M., **Fienberg, S. E.**, & **Holland, P. W.** (1975). *Discrete multivariate analysis: Theory and practice*. MIT Press.

Busemeyer, J. R., & **Diederich, A.** (2010). *Cognitive modeling*. Sage.

Cressie, N., & **Read, T. R. C.** (1989). Pearson's X^2 and the loglikelihood ratio statistic G^2: A comparative review. *International Statistical Review*, **57**, 19–43. https://doi.org/10.2307/1403582

Crowther, C. S., **Batchelder, W. H.**, & **Hu, X.** (1995). A measurement-theoretic analysis of the fuzzy logic model of perception. *Psychological Review*, **102**, 396–408. https://doi.org/10.1037/0033-295X.102.2.396

Delucchi, K. L. (1983). The use and misuse of chi-square: Lewis and Burke revisited. *Psychological Bulletin*, **94**, 166–176. https://doi.org/10.1037/0033-2909.94.1.166

Ditterich, J. (2015). Distinguishing between models of perceptual decision making. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 277–290). Springer. https://doi.org/10.1007/978-1-4939-2236-9_13

Forbes, C., **Evans, M.**, **Hastings, N.**, & **Peacock, B.** (2011). *Statistical distributions* (4th ed.). Wiley.

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

García-Pérez, M. A. (1994). Parameter estimation and goodness-of-fit testing in multinomial models. *British Journal of Mathematical and Statistical Psychology*, **47**, 247–282. https://doi.org/10.1111/j.2044-8317.1994.tb01037.x

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.** (2012). 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.** (2015). Visual and auditory components in the perception of asynchronous audiovisual speech. *i-Perception*, **6**(6), 2041669515615735. https://doi.org/10.1177/2041669515615735

García-Pérez, M. A., & **Alcalá-Quintana, R.** (2017). 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., & **Alcalá-Quintana, R.** (2020a). Order effects in two-alternative forced-choice tasks invalidate adaptive threshold estimates. *Behavior Research Methods*, **52**, 2168–2187. https://doi.org/10.3758/s13428-020-01384-6

García-Pérez, M. A., & **Alcalá-Quintana, R.** (2020b). Assessing multisensory integration and estimating speed of processing with the dual-presentation timing task: Model and data. *Journal of Mathematical Psychology*, **96**, 102351. https://doi.org/10.1016/j.jmp.2020.102351

García-Pérez, M. A., & **Núñez-Antón, V.** (2001). Small-sample comparisons for power-divergence goodness-of-fit statistics for symmetric and skewed simple null hypotheses. *Journal of Applied Statistics*, **28**, 855–874. https://doi.org/10.1080/02664760120074942

García-Pérez, M. A., & **Núñez-Antón, V.** (2004a). Small-sample comparisons for goodness-of-fit statistics in one-way multinomials with composite hypotheses. *Journal of Applied Statistics*, **31**, 161–181. https://doi.org/10.1080/0266476032000148849

García-Pérez, M. A., & **Núñez-Antón, V. A.** (2004b). On the chi-square approximation to the exact distribution of goodness-of-fit statistics in multinomial models with composite hypotheses. *British Journal of Mathematical and Statistical Psychology*, **57**, 73–96. https://doi.org/10.1348/000711004849240

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.** (2019). Psychophysical tests do not identify ocular dominance consistently. *i-Perception*, **10**(2), 2041669519841397. https://doi.org/10.1177/2041669519841397

Heller, J. (2017). Identifiability in probabilistic knowledge structures. *Journal of Mathematical Psychology*, **77**, 46–57. https://doi.org/10.1016/j.jmp.2016.07.008

Huk, A., **Bonnen, K.**, & **He, B. J.** (2018). Beyond trial-based paradigms: Continuous behavior, ongoing neural activity, and natural stimuli. *Journal of Neuroscience*, **38**, 7551–7558. https://doi.org/10.1523/JNEUROSCI.1920-17.2018

Kelber, P., & **Ulrich, R.** (2024). Independent-channels models of temporal-order judgment revisited: A model comparison. *Attention, Perception, & Psychophysics*, **86**, 2187–2209. https://doi.org/10.3758/s13414-024-02915-5

Lam, C. F., **Dubno, J. R.**, & **Mills, J. H.** (1999). Determination of optimal data placement for psychometric function estimation: A computer simulation. *Journal of the Acoustical Society of America*, **106**, 1969–1976. https://doi.org/10.1121/1.427944

Lam, C. F., **Mills, J. H.**, & **Dubno, J. R.** (1996). Placement of observations for the efficient estimation of a psychometric function. *Journal of the Acoustical Society of America*, **99**, 3689–3693. https://doi.org/10.1121/1.414966

Leek, M. R., **Hanna, T. E.**, & **Marshall, L.** (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception & Psychophysics*, **51**, 247–256. https://doi.org/10.3758/BF03212251

Lewandowsky, S., & **Oberauer, K.** (2018). Computational modeling in cognition and cognitive neuroscience. In E.-J. Wagenmakers (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience, Vol. 5, Methodology* (pp. 1–35). Wiley.

Miller, J., & **Ulrich, R.** (2001). On the analysis of psychometric functions: The Spearman-Kärber method. *Perception & Psychophysics*, **63**, 1399–1420. https://doi.org/10.3758/BF03194551

Navarro, D. J., **Pitt, M. A.**, & **Myung, I. J.** (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, **49**, 47–84. https://doi.org/10.1016/j.cogpsych.2003.11.001

O'Regan, J. K., & **Humbert, R.** (1989). Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and slope estimations when small samples are used. *Perception & Psychophysics*, **46**, 434–442. https://doi.org/10.3758/BF03210858

Pitt, M. A., **Myung, I. J.**, & **Zhang, S.** (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472–491. https://doi.org/10.1037/0033-295X.109.3.472

Qarehdaghi, H., & **Rad, J. A.** (2024). EZ-CDM: Fast, simple, robust, and accurate estimation of circular diffusion model parameters. *Psychonomic Bulletin & Review*, **31**, 2058–2091. https://doi.org/10.3758/s13423-024-02483-7

Rammsayer, T., & **Ulrich, R.** (2001). Counting models of temporal discrimination. *Psychonomic Bulletin & Review*, **8**, 270–277. https://doi.org/10.3758/BF03196161

Rasanan, A. H. H., **Evans, N. J.**, **Fontanesi, L.**, **Manning, C.**, **Huang-Pollock, C.**, **Matzke, D.**, **Heathcote, A.**, **Rieskamp, J.**, **Speekenbrink, M.**, **Frank, M. J.**, **Palminteri, S.**, **Lucas, C. G.**, **Busemeyer, J. R.**, **Ratcliff, R.**, & **Rad, J. A.** (2024). Beyond discrete-choice options. *Trends in Cognitive Sciences*, **28**, 857–870. https://doi.org/10.1016/j.tics.2024.07.004

Read, T. R. C., & **Cressie, N. A. C.** (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer

Starns, J. J. (2021). High- and low-threshold models of the relationship between response time and confidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **47**, 671–684. https://doi.org/10.1037/xlm0000960

Treutwein, B., & **Strasburger, H.** (1999). Fitting the psychometric function. *Perception & Psychophysics*, **61**, 87–106. https://doi.org/10.3758/BF03211951

Tünnermann, J., & **Scharlau, I.** (2018). Stuck on a plateau? A model-based approach to fundamental issues in visual temporal-order judgments. *Vision*, **2**(3), 29. https://doi.org/10.3390/vision2030029

Tünnermann, J., & **Scharlau, I.** (2021). Big-M-small-N temporal-order judgment data. *The Quantitative Methods for Psychology*, **17**, 355–373. https://doi.org/10.20982/tqmp.17.4.p355

Watson, A. B., & **Fitzhugh, A.** (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics*, **47**, 87–91. https://doi.org/10.3758/BF03208169

Wichmann, F. A., & **Hill, N. J.** (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, **63**, 1293–1313. https://doi.org/10.3758/BF03194544