

Classic-Probability Accounts of Mirrored (Quantum-Like) Order Effects in Human
Judgments

David Kellen

Syracuse University

Henrik Singmann

University of Zurich

William H. Batchelder

University of California, Irvine

Author Note

David Kellen, Department of Psychology, Syracuse University. Henrik Singmann, Faculty of Psychology, University of Zürich. William H. Batchelder, Department of Cognitive Sciences, University of California, Irvine.

The first two authors contributed equally to this manuscript.

We thank Zheng Joyce Wang for providing the raw data and Jerome Busemeyer for his detailed comments on a previous version. David Kellen and Henrik Singmann received support from the Swiss National Science Foundation Grant 100014_165591. William Batchelder received support from the US National Science Foundation Grant #1534471.

Correspondence concerning this article should be addressed to David Kellen, Department of Psychology, Syracuse University, 430 Huntington Hall, Syracuse, NY 13244. Electronic mail may be sent to davekellen@gmail.com.

Abstract

Using a large data corpus, Wang, Solloway, Shiffrin, and Busemeyer (2014) showed that order effects in the responses given to pairs of related agree/disagree questions presented in succession follow a specific pattern termed QQ-equality. The fact that QQ-equality corresponds to a parameter-free prediction of a proposed quantum-probability model, together with the failure of several alternative classic-probability accounts, led Wang et al. to conclude that it constitutes strong evidence for the quantum nature of human judgments and to issue a challenge for the development of suitable classic-probability accounts. We respond to Wang et al.'s challenge by discussing a class of repeat-choice models that is able to yield the QQ-equality as a parameter-free prediction (or a very-likely prediction a priori) and provide an overall account of the data that is comparable to the quantum model. The success of this class of models establishes a plausible benchmark against which quantum accounts of order effects – like the ones observed in this data corpus – can be compared. Finally, we argue that the assumption of respondent homogeneity implied in Wang et al.'s use of aggregated data is extremely problematic for some of the alternative models discussed here (but not necessarily for the quantum account), leading to spurious rejections at non-negligible rates. We also discuss how a move away from aggregated data could help resolve some theoretical challenges that the quantum account of QQ-equality currently faces.

Keywords: human judgments, order effects, quantum probability

In recent years, there has been a surge of research developing models for psychological data based on quantum probability models (for introductions and reviews, see Busemeyer & Bruza, 2012; Pothos & Busemeyer, 2013). A major motivation for developing models based on quantum probability theory is that they provide an elegant characterization of sometimes-problematic phenomena in human judgments, such as order effects or violations of classic logic and probability theory. For example, a quantum account can seamlessly describe the so-called “Linda effect” (Tversky & Kahneman, 1983), in which participants tend to perceive the conjunction “Linda is active in the feminist movement and a bank teller” as more likely than the simple statement “Linda is a bank teller” (Busemeyer, Pothos, Franco, & Trueblood, 2011; Busemeyer, Wang, Pothos, & Trueblood, 2015). Quantum accounts achieve this by representing probabilities as arising from projections into a Hilbert space, where for example, some basic operations are not commutative. Due to its many departures from classic probability theory, the recent use of quantum theory in cognitive modeling has become a polarizing but exciting topic among researchers (see the numerous comments to Pothos & Busemeyer, 2013). Recently, Wang, Solloway, Shiffrin, and Busemeyer (2014) tested a critical prediction of their quantum-probability model that holds independently of any particular model parametrization. This prediction, which was considered by Busemeyer and colleagues to be “the strongest to date” (Busemeyer et al., 2015, p. 241), concerns the direction and magnitude of question-order effects in polls and questionnaires (Moore, 2002). Consider response matrices \mathbf{P} and \mathbf{Q} in the upper-left and lower-left side of Table 1, both summarizing “yes” and “no” responses to two questions A and B:

Question A: “Do you generally think Bill Clinton is honest and trustworthy?”

Question B: “Do you generally think Al Gore is honest and trustworthy?”

In one question-order condition, question A was asked first, immediately followed by question B (condition \mathbf{P}), whereas in a second condition the question order was reversed (condition \mathbf{Q}). The responses given to these questions in each of the two

conditions are given by the two 2×2 matrices \mathbf{P} and \mathbf{Q} in Table 1. In *both* matrices, the first and second rows report the cases in which respondents responded “yes” and “no” to question A (Clinton), respectively. The first and second columns show the cases in which respondents responded “yes” and “no” to question B (Gore), respectively. For example, $\mathbf{P}_{1,2}$ indicates the proportion of cases in which respondents responded “yes” to question A, followed by response “no” to question B. Alternatively, $\mathbf{Q}_{1,2}$ indicates the proportion of times respondents answered “no” to question B, followed by a “yes” to question A.

The data reported in \mathbf{P} and \mathbf{Q} in Table 1 show that the marginal response proportions differ between the two conditions: Clinton is more often considered to be honest when question B (Gore) is asked first rather than second (58.8% versus 53.5%). In contrast, Gore is less often considered to be honest when question A (Clinton) is asked first (66.6% versus 76.1%). The order effects found in these data are made clear when looking at the difference between \mathbf{P} and \mathbf{Q} , reported in the upper-right side of Table 1. For example, the proportion of respondents who responded “yes” to both questions diminishes when questioned about Clinton first. According to Moore (2002), this kind of order effects are typically characterized as a byproduct of shifting comparison standards: When questioned first about Bill Clinton’s honesty, respondents compare him to some standard based on their background knowledge and memories. But when Clinton’s honesty is questioned after Gore’s, then the proximity of the two questions leads respondents to incorporate their previous assessment of Gore in their comparison standard, potentially affecting their judgment. An analogous scenario is expected to occur when the question about Gore is preceded by the question about Clinton. The verbal account given by Moore is close to the way order effects occur in a quantum probability account, where the response probabilities for the second question are based on an updated belief state that is determined by the response given to the first question.

One intriguing aspect in the differences found between \mathbf{P} and \mathbf{Q} is their mirror-like pattern: The difference found between the cases in which people responded “yes” to

both questions (-0.0726) is pretty much symmetrical to the difference found when both responses were “no” (0.0756). A similar pattern is found in the order effects of the other two response patterns (0.0192 versus -0.0224). Overall, this mirror-like pattern suggests that order effects amount to a shift in probability across the response matrices’ diagonals (see the $\mathbf{P} - \mathbf{Q}$ matrix in the lower-right part of Table 1). Wang et al. (2014) argued that according to their quantum-probability model, the sums of the diagonals of $\mathbf{P}-\mathbf{Q}$, which will be designated as QQ-values (Quantum Question values), *are always expected to be zero*.¹ This result, which is designated as *QQ-equality*, emerges from the fact that the quantum model is constrained to produce shifts in probability mass across the two diagonals of the $\mathbf{P}-\mathbf{Q}$ matrix.² According to the quantum-probability account, when individuals do not have a clear opinion on both politicians, their measurements are *complementary*: these measures cannot be obtained simultaneously and the order of the measurements (questions) matter in the sense that the first provides a contextualized state upon which the second will be framed. Wang and Busemeyer (2015) argued that complementarity will depend on the individuals’ experience with the subject being questioned and is also expected to be a function of factors such as respondents’ age, level of cognitive development, among others. Complementarity is then mostly expected when respondents are presented with uncommon or familiar pairs of questions for which answers have to be produced on the fly.

Using a corpus of 72 datasets, 70 coming from representative US surveys (most containing more than 1000 adult respondents) and two coming from their own experimental studies, Wang et al. (2014) tested QQ-equality directly. The test relied on a constrained parametrization of the two multinomial distributions associated with matrices \mathbf{P} and \mathbf{Q} , a parametrization that is represented here by the tree model illustrated in Figure 1. According to this model, which we will refer to as the *QQ-test model*, probability mass can only shift between cells in the same diagonal, namely between response pairs “yes”-“yes” and “no”-“no” (e.g., between $\mathbf{P}_{1,1}$ and $\mathbf{P}_{2,2}$), and

¹Wang et al. (2014) showed that under more relaxed assumptions, their quantum model can at most predict minute deviations from zero (see their Supplemental Information).

²A formal proof of this result given by Wang et al. (2014).

between pairs “yes”-“no” and “no”-“yes” (e.g., between $\mathbf{P}_{1,2}$ and $\mathbf{P}_{2,1}$). Parameter θ_1 determines how much probability mass is attributed to each diagonal of the matrices. The shifts within the diagonals are captured by parameters θ_2 , θ_3 , θ_4 , and θ_5 . Irrespective of the values taken by the model parameters, QQ-equality is bound to hold. Besides QQ-equality, this model imposes no other constraints on the multinomial distributions.

Goodness-of-fit tests showed that the QQ-test model was rejected in approximately 5% of the datasets, in line with the nominal rejection rates expected under the null hypothesis. Figure 2 depicts the observed distribution of QQ-values (Left Panel), which is concentrated around zero, demonstrating an impressive consistency of the data with QQ-equality. This result is reflected in the summed misfits (quantified by the G^2 statistic), which did not turn out to be statistically significant (summed $G^2(72) = 75.91$, $p = .35$).³ The right panel of Figure 2 further shows that there is no obvious relationship between the magnitude of the QQ-values and the overall size of the order effects. The goodness-of-fit tests conducted by Wang et al. (2014) assume that responses are *independent and identically distributed* (i.i.d.) within each tree in Figure 1.⁴ Realistically, the i.i.d. assumption should not be expected to hold here given that the data come from heterogeneous respondents that belong to groups associated with quite distinct data-generating probability distributions (e.g., Conservatives, Liberals). Fortunately, it can be easily shown that QQ-equality is bound to hold under heterogeneity. This fact, together with QQ-test model’s lack of further constraints other than QQ-equality dismiss any major concerns with Wang et al.’s testing of the latter using aggregated data.

In addition to demonstrating the presence of QQ-equality, Wang et al. (2014) checked whether the data as a whole could be successfully described by some plausible cognitive models based on classic probability theory (see their Supplemental

³Data and scripts using the MPTinR package (Singmann & Kellen, 2013) are made available at <https://osf.io/n9q2m/>

⁴Under i.i.d., the goodness of fit of the QQ-test model (here quantified by the G^2 statistic) is asymptotically distributed as a χ^2 distribution with one degree of freedom (the data in each study provide a total of six degrees of freedom)

Information). It is important to note that the QQ-test model and these alternative cognitive models operate at distinct levels, given that the former does not postulate any specific processes, whereas the latter do.⁵ In this sense they should not be seen as direct competitors. However, the relative success of alternative models has the potential of informing us of the suitability of the quantum-probability account. For instance, previous work demonstrated the inability of classic Bayesian and Markov models to account for this kind of order effects, suggesting that the adoption of a quantum-probability account may be necessary (e.g., Busemeyer, Wang, & Lambert-Mogiliansky, 2009). In the present case, Wang et al. showed that two classic-probability models, a *repeat-choice model* and an *anchoring-adjustment model*, were rejected by the data as they failed to accurately characterize the observed order effects while assuming QQ-equality. Although Wang et al. admitted that the development of better classic-probability models is both possible and desirable, they also argued that these models are likely to be *overly flexible* and predict QQ-equality *only within limited ranges of parameter values*. Because the quantum account (instantiated by the QQ-test model illustrated in Figure 1) is bound to predict QQ-equality irrespective of parameter values, the data will always be more supportive of it. Overall, Wang et al. demonstrate that the observation of QQ-equality in human judgments introduces strong constraints on theories concerning order effects.

The goal of the remainder of the present manuscript is to provide a first response to Wang et al.'s (2014) call for the development of classic-probability models that predict QQ-equality. We will focus on the class of repeat-choices models, in which the model originally discussed by Wang et al. can be included. As will be shown below, some repeat-choice models are constrained to predict QQ-equality, whereas others predict QQ-equality with high likelihood. Also, some of these models are able to fit the data corpus as well as Wang et al.'s quantum account. Altogether, these theoretical and empirical results indicate that the class of repeat-choice models provides a suitable

⁵This assertion deserves some clarification: Quantum theories of cognition do in fact postulate specific processes (e.g., they postulate initial belief states that undergo a series of transformations or updates). However, the QQ-test model used by Wang et al. (2014) does not explicitly characterize any of these processes, only the QQ-equality prediction that emerges from them.

alternative to the quantum account. Finally, we challenge the implicit assumption of respondent homogeneity in Wang et al.’s tests and raise an important issue that so far appears to have been overlooked in this line of research, namely the impact of respondent heterogeneity in the testing and rejection of some of the candidate models. Specifically, some of the simpler candidate models considered can be spuriously rejected due to distortions introduced by the aggregation of heterogeneous respondents. The problem is that although QQ-equality is always preserved in aggregated data, *other properties are not*. This situation is extremely problematic when attempting to test models that impose constraints above and beyond QQ-equality in a fair manner.

Varieties of Repeat-Choice Models

In order to facilitate the introduction and discussion of the class of repeat-choice models, we will begin by introducing a general model, which we will refer to as \mathcal{M}_0 . All repeat-choice models discussed hereon are special cases of \mathcal{M}_0 . Figure 3 provides an illustration of how \mathcal{M}_0 accounts for the cells of matrices \mathbf{P} and \mathbf{Q} . Figure 4 describes a non-exhaustive hierarchy of submodels that can be derived from \mathcal{M}_0 . As shown in Figure 3, the model assumes a set of four so-called preference states $S_{i,j}$ that characterize the probabilities associated to a (sampled) respondent having a specific prior belief or opinion regarding the two questions posed (where subscript 1 codes “yes” and 2 codes “no”).⁶ These preference states are assumed to be unaffected by the order of the questions.

When individuals provide a first response based on state $S_{i,j}$ there is a probability $a_{\mathbf{O},i,j}$ that the subsequent response will be a function on the latter one. This possibility simply reflects the fact that the response to the second question occurs with the knowledge that a previous question was just asked. For example, if a respondent notices that the two questions are related (e.g., she would keep in mind that Clinton and Gore were part of the same US administration), she might tap into the same information structures when producing her second response. If the second response happens to be a

⁶Note that we are using the term “preference” rather loosely. Moreover, note that because $\sum_{i=1}^2 \sum_{j=1}^2 S_{i,j} = 1$, a completely unconstrained characterization of these probabilities can be achieved via three non-redundant $S_{i,j}$ parameters.

function of the first, then with probability $r_{\mathbf{O},i,j}$ the second response is exactly the same as the previous one (i.e., there is an *assimilation* effect), and with probability $1 - r_{\mathbf{O},i,j}$ the opposite response is produced (i.e., there is an *contrast* effect). With probability $1 - a_{\mathbf{O},i,j}$ the second response is entirely determined by $S_{i,j}$, which means that no order effect is expected. These probabilities a and r are dependent on the question order ($\mathbf{O} = \mathbf{P}$ or $\mathbf{O} = \mathbf{Q}$) as well as on the respondent's preference state.

As it stands, \mathcal{M}_0 is not a very interesting nor useful model given that it can account for any data that could be observed in \mathbf{P} and \mathbf{Q} (e.g., all types of order effects as well as their absence). To make matters worse, the number of free parameters it postulates (nineteen) is much larger than the six degrees of freedom provided by the data (i.e., the model is oversaturated; see Bamber & Van Santen, 2000). However, some special cases of \mathcal{M}_0 described in Figure 4 do impose testable constraints, including QQ-equality. The special cases of \mathcal{M}_0 described in Figure 4 and discussed below follow two distinct assumptions: The subset of \mathcal{M}_1 models assumes that the processes underlying response dependencies are independent of question order, whereas the subset of \mathcal{M}_2 models assumes that such processes depend on the order in which questions are posed.

The simplest member of the class of repeat-choice model discussed here, model $\mathcal{M}_{1ar,S}/\mathcal{M}_{2ar,S}$, only assumes four parameters. This restricted model assumes that the preference states are *conditionally independent* such that the four possible preference states $S_{i,j}$ are defined by independent response probabilities for each of the questions, with $S_{1,1} = S_1.S_{.1}$, $S_{1,2} = S_1.(1 - S_{.1})$, $S_{2,1} = (1 - S_1).S_{.1}$, and $S_{2,2} = (1 - S_1).(1 - S_{.1})$. This model also assumes that the probability of the second response being dependent on the first, as well as the probability of the previous response being repeated, is the same across different preference states and question orders (i.e., the model assumes single a and r parameters). Although this simple model always predicts QQ-equality, it grossly fails to characterize the data corpus. But as discussed in detail below, such failure is ultimately non-informative with respect to the adequacy of the model as it is expected to fail when fitting aggregated data coming from heterogeneous respondents.

Submodel Assuming Question-Order Independence

First, let us consider a submodel \mathcal{M}_1 in which the a and r parameters are independent of the order in which the questions are posed (i.e., $a_{\mathbf{P},i,j} = a_{\mathbf{Q},i,j}$ and $r_{\mathbf{P},i,j} = r_{\mathbf{Q},i,j}$, so we drop the question-order subscript \mathbf{O}). This implies that the probability of the second response being dependent on the first is independent of the order the questions. In a way, one could argue that this restriction reflects the notion that some individuals (in particular those without any strong or clear opinions about both questions) might simply produce a second response that repeats the previous one or generates the opposite response, without really expressing their preference state. However, one does not necessarily have to commit to this specific interpretation, as these parameter restrictions could simply reflect the notion that respondents are aware of their preferences for both questions but sometimes decide to express a second response as a function of their first, in a way that is independent of the question order. Submodel \mathcal{M}_1 yields the following equations for \mathbf{P} and \mathbf{Q} :

$$\begin{aligned}
\mathbf{P}_{1,1} &= S_{1,1}a_{1,1}r_{1,1} + S_{1,1}(1 - a_{1,1}) + S_{1,2}a_{1,2}r_{1,2}, \\
\mathbf{P}_{1,2} &= S_{1,1}a_{1,1}(1 - r_{1,1}) + S_{1,2}a_{1,2}(1 - r_{1,2}) + S_{1,2}(1 - a_{1,2}), \\
\mathbf{P}_{2,1} &= S_{2,1}a_{2,1}(1 - r_{2,1}) + S_{2,1}(1 - a_{2,1}) + S_{2,2}a_{2,2}(1 - r_{2,2}), \\
\mathbf{P}_{2,2} &= S_{2,1}a_{2,1}r_{2,1} + S_{2,2}a_{2,2}r_{2,2} + S_{2,2}(1 - a_{2,2}), \\
\mathbf{Q}_{1,1} &= S_{1,1}a_{1,1}r_{1,1} + S_{1,1}(1 - a_{1,1}) + S_{2,1}a_{2,1}r_{2,1}, \\
\mathbf{Q}_{1,2} &= S_{1,2}a_{1,2}(1 - r_{1,2}) + S_{1,2}(1 - a_{1,2}) + S_{2,2}a_{2,2}(1 - r_{2,2}), \\
\mathbf{Q}_{2,1} &= S_{1,1}a_{1,1}(1 - r_{1,1}) + S_{2,1}a_{2,1}(1 - r_{2,1}) + S_{2,1}(1 - a_{2,1}), \\
\mathbf{Q}_{2,2} &= S_{1,2}a_{1,2}r_{1,2} + S_{2,2}a_{2,2}r_{2,2} + S_{2,2}(1 - a_{2,2}).
\end{aligned}$$

Although \mathcal{M}_1 has a total of eleven parameters, still outnumbering by far the six degrees of freedom provided by the data, it still imposes testable constraints.

Specifically, \mathcal{M}_1 predicts QQ-equality, *irrespective of the values taken by its parameters*:

$$\begin{aligned}
\mathbf{P}_{1,1} - \mathbf{Q}_{1,1} + \mathbf{P}_{2,2} - \mathbf{Q}_{2,2} &= S_{1,2}a_{1,2}r_{1,2} - S_{2,1}a_{2,1}r_{2,1} + S_{2,1}a_{2,1}r_{2,1} - S_{1,2}a_{1,2}r_{1,2} \\
&= \mathbf{P}_{1,2} - \mathbf{Q}_{1,2} + \mathbf{P}_{2,1} - \mathbf{Q}_{2,1} \\
&= S_{1,1}a_{1,1}(1 - r_{1,1}) - S_{2,2}a_{2,2}(1 - r_{2,2}) \\
&+ S_{2,2}a_{2,2}(1 - r_{2,2}) - S_{1,1}a_{1,1}(1 - r_{1,1}) = 0
\end{aligned}$$

When fitted to the 72 datasets analyzed by Wang et al. (2014), \mathcal{M}_1 produced the exact same G^2 values as the QQ-test model (summed $G^2 = 75.91$). Also, the p -values for \mathcal{M}_1 were obtained with a semi-parametric bootstrap procedure (van de Schoot, Hoijsink, & Deković, 2010) and, and aside from the variability coming from the bootstrap sampling, were pretty much identical to the p -values of the QQ-test model (see Figure 5, Left Panel). Furthermore, the rate of significant misfits (i.e., $p < .05$) is at the nominal level of $\alpha = .05$ (4 out of 72 data sets are significant which corresponds to .056). These results simply reflect the fact that that the QQ-test model and \mathcal{M}_1 occupy the same prediction space, in which probability mass can only shift across the matrices' diagonals (without the imposition of any further constraints).

Submodel Assuming That Question Order Matters

In the previous subsection, we showed that imposing the constraint that the a and r parameters were independent of question order (i.e., $a_{\mathbf{P},i,j} = a_{\mathbf{Q},i,j}$ and $r_{\mathbf{P},i,j} = r_{\mathbf{Q},i,j}$) led to the prediction of QQ-equality with a model that is equivalent to the QQ-test model. However, this imposed constraint can be seen as not quite representative of the way many individuals would base their second response on the first one: Consider a respondent that finds Clinton to be dishonest but believes that Gore is honest (i.e., $S_{2,1}$). According to \mathcal{M}_1 , in the case of question order \mathbf{P} (i.e., Clinton - Gore), when confronted with the second (Gore) question, the respondent will notice the similarity between the two questions and repeat his previous “no” response to Clinton with

probability $a_{2,1}r_{2,1}$. In the case of **Q** (i.e., Gore - Clinton), the *same* respondent would respond “yes” to both questions *with the same probability*. Intuitively, one would expect that these probabilities would differ as a function of the match/mismatch between the two preferences, as well as a function of the specific preference manifested in the first response. Because of this discrepancy between our intuition and the behavior of submodel \mathcal{M}_1 , we will introduce an alternative submodel of \mathcal{M}_0 .

Let us then consider another submodel, which we will refer to as \mathcal{M}_2 , that incorporates the above-described desiderata. \mathcal{M}_2 assumes that the response probabilities associated to **P** and **Q** correspond to:

$$\begin{aligned}
\mathbf{P}_{1,1} &= S_{1,1}a_{1,1}r_{1,1} + S_{1,1}(1 - a_{1,1}) + S_{1,2}a_{1,2}r_{1,2}, \\
\mathbf{P}_{1,2} &= S_{1,1}a_{1,1}(1 - r_{1,1}) + S_{1,2}a_{1,2}(1 - r_{1,2}) + S_{1,2}(1 - a_{1,2}), \\
\mathbf{P}_{2,1} &= S_{2,1}a_{2,1}(1 - r_{2,1}) + S_{2,1}(1 - a_{2,1}) + S_{2,2}a_{2,2}(1 - r_{2,2}), \\
\mathbf{P}_{2,2} &= S_{2,1}a_{2,1}r_{2,1} + S_{2,2}a_{2,2}r_{2,2} + S_{2,2}(1 - a_{2,2}), \\
\mathbf{Q}_{1,1} &= S_{1,1}a_{1,1}r_{1,1} + S_{1,1}(1 - a_{1,1}) + S_{2,1}a_{1,2}r_{1,2}, \\
\mathbf{Q}_{1,2} &= S_{1,2}a_{2,1}(1 - r_{2,1}) + S_{1,2}(1 - a_{2,1}) + S_{2,2}a_{2,2}(1 - r_{2,2}), \\
\mathbf{Q}_{2,1} &= S_{1,1}a_{1,1}(1 - r_{1,1}) + S_{2,1}a_{1,2}(1 - r_{1,2}) + S_{2,1}(1 - a_{1,2}), \\
\mathbf{Q}_{2,2} &= S_{1,2}a_{2,1}r_{2,1} + S_{2,2}a_{2,2}r_{2,2} + S_{2,2}(1 - a_{2,2}).
\end{aligned}$$

Like \mathcal{M}_1 , this model has eleven free parameters. According to \mathcal{M}_2 , in cases where the preferences in the two questions disagree ($S_{1,2}$ and $S_{2,1}$), the postulated a and r parameters depend on the question order and the first question asked. Specifically, \mathcal{M}_2 restricts \mathcal{M}_0 such that $a_{\mathbf{P},i,j} = a_{\mathbf{Q},j,i}$ and $r_{\mathbf{P},i,j} = r_{\mathbf{Q},j,i}$. For instance, in the case of condition **P**, a respondent in state $S_{2,1}$ would respond “no” to the Gore question with probability $a_{2,1}r_{2,1}$, whereas in condition **Q**, she would respond “yes” to both questions with probability $a_{1,2}r_{1,2}$. When the preferences in the two questions agree ($S_{1,1}$ and $S_{2,2}$), \mathcal{M}_2 's predictions are equivalent to \mathcal{M}_1 in the sense that the same a and r parameters apply to both **P** and **Q**. Based on these predictions, it is easy to see that

\mathcal{M}_2 does not necessarily predict QQ-equality. For example:

$$\mathbf{P}_{1,1} - \mathbf{Q}_{1,1} + \mathbf{P}_{2,2} - \mathbf{Q}_{2,2} = (S_{1,2} - S_{2,1})a_{1,2}r_{1,2} + (S_{2,1} - S_{1,2})a_{2,1}r_{2,1}.$$

Given the robust observation of QQ-equality, the lack of such a constraint can be perceived as a shortcoming or even as a disqualifying feature. However, when comparing models based on the constraints they impose on data, it is important to keep in mind the distinction between *likelihood-function-based* and *parameter-based* constraints. So-called likelihood-function-based constraints are part of the model's functional form and do not depend on parameter values. In contrast, parameter-based constraints focus on the plausibility and/or admissibility of certain parameter values. This distinction is often made when comparing Bayesian and non-Bayesian methods of hypothesis testing and model selection (for a discussion, see Lee, 2016).

In the present case, the possibility of deviations from QQ-equality raises questions regarding its likelihood and expected magnitude. Such questions can be answered via an evaluation of \mathcal{M}_2 's *prior predictive distribution* (e.g., Lee, 2016). This distribution is a Bayesian construct that gives the distribution of data that is expected under a model a priori. We obtained a prior predictive distribution for QQ-values by computing them over a large number of parameter values that were sampled from non-informative prior distributions. These non-informative priors simply reflect our present ignorance regarding the parameters. As can be seen in the left panel of Figure 6, the prior-predictive distribution of QQ-values obtained with non-informative priors is highly peaked at 0, with 95% of values between -0.22 and 0.22. This result indicates that QQ-equality is extremely likely. It is important to note that the spread of the prior predictive distribution is a function of the upper bound of the a parameters. Because the priors used were non-informative, a parameters could take on any value between 0 and 1. However, if one incorporates the plausible notion that the probability of the second response being dependent on the first one is rather low (i.e., a is between 0 and .50), then the resulting (zero-centered) prior predictive distribution has 95% of its values between -0.11 and 0.11 (see the right panel of Figure 6). The latter range of

values is somewhat close to the actual data, whose minimal and maximal values were -0.09 and 0.09.

When fitted to the 72 datasets analyzed by Wang et al. (2014), \mathcal{M}_2 produced a summed G^2 of 79.04, which is marginally worse than the summed misfits of \mathcal{M}_1 and the QQ-test model ($\Delta G^2 = 3.13$). Despite this marginal difference at the level of summed misfits, we found considerable differences when comparing them at the level of single datasets. Specifically, the majority of \mathcal{M}_2 's misfits, with a summed G^2 of 50.99, can be attributed to ten (14%) datasets for which the model was rejected ($p < .05$). Out of the remaining datasets, thirty-two (44%) \mathcal{M}_2 were perfectly accounted for ($G^2 = 0$; compared to 5 or 7% for \mathcal{M}_1), and the remaining thirty (42%) led to small, non-significant misfits.⁷ The distribution of p -values for \mathcal{M}_2 is provided in the center panel of Figure 5, and looks distinct from the roughly uniform distribution that is found for the QQ-test model or \mathcal{M}_1 (see the left panel of Figure 5) and what is typically expected in the fits from model that corresponds to the data-generating processes. However, this distribution of p -values does not necessarily indicate that the model is unsuitable. For example, this kind of p -value distributions are found in models whose prediction space corresponds to a convex polytope (for a review, see Davis-Stober, 2009).⁸ The right panel of Figure 5 directly compares the misfits of \mathcal{M}_2 and \mathcal{M}_1 (and the QQ-test model) and shows that for many data sets for which \mathcal{M}_2 provides a perfect account \mathcal{M}_1 shows some misfit while the same does not seem to hold the other way round. Overall, nothing besides the slightly larger rejection rate (14%) speaks against the suitability of this model.

⁷We entertained the possibility that that the data sets for which \mathcal{M}_2 provided significant misfit were outliers in some sense, but we did not find any evidence for this idea. The ten data sets with significant misfit were (in order of decreasing misfit): 1 Pew study, the abortion question pair, and 8 more Pew studies (66 of the 72 data sets were Pew studies on various topics). Furthermore, G^2 values obtained with \mathcal{M}_2 were not related to any characteristic of the data set with the exception of the magnitude of the order effect w , $r(70) = .31$, $p = .008$. Such a correlation was also observed for \mathcal{M}_1 and the QQ-test model, $r(70) = .42$, $p = .0002$.

⁸For example, consider the hypothesis $\mathcal{H}_0 : \theta \leq .50$ regarding the rate parameter of a binomial distribution. The sampling distribution of the G^2 statistic of this hypothesis follows a mixture of χ^2 distributions, specifically $\frac{1}{2}\chi_{df=0}^2 + \frac{1}{2}\chi_{df=1}^2$. The distribution of p -values coming from this mixture is not uniform, as it places half of the probability mass on $p = 1$ and the other half on p -values between .50 and 0.

Testing Additional Parameter Constraints

Given that \mathcal{M}_1 and \mathcal{M}_2 have eleven free parameters to describe the six independent data points provided by the two question orders, it is interesting to check whether further restricted versions can account for the data and whether they yield identifiable parameters. The restricted models considered here are described in Figure 4 and their relative goodness-of-fit performance (quantified via the G^2 statistic) is reported in Table 2. As previously mentioned, these tests imply that the data come from homogeneous respondents, which is very unlikely to hold here. The presence of respondent heterogeneity can be problematic for some of the models included here, an issue that we will discuss in greater detail in the following section. But in the present analyses we simply follow Wang et al. (2014)'s approach and assume that relying on aggregate data is appropriate for the purposes of model comparison.

The goodness-of-fit results reported in Table 2 show that most constraints lead to relatively small increases in misfit relative to the QQ-test model, with the exception of model $\mathcal{M}_{2yn,a}$ ($\Delta G^2 = 57.84$), models enforcing all a and r parameters to take on the same value ($\Delta G^2 = 260.82$), and models imposing conditional independence on the preference states $S_{i,j}$ ($\Delta G^2 = 1102.30$ and $\Delta G^2 = 1360.79$). Among the simpler models, the best-performing one in terms of goodness of fit was $\mathcal{M}_{2yn,r}$.⁹ Given that this model yields identifiable parameters, it is interesting to see what kind of characterization of the data it provides. Figure 7 (right panel) shows the distribution of parameters for $\mathcal{M}_{2yn,r}$: First, the estimates of the $S_{i,j}$ states indicate that $S_{1,1}$ and $S_{2,2}$ are more frequent than the incongruent states $S_{1,2}$ and $S_{2,1}$. Moreover, both a parameters seem to have very similar distributions, with medians around .20 and .30, respectively. These small values are in line with the notion that only a limited portion of second responses is based on the first response (reflecting the small magnitude of the order effects). With regards to the r parameter, both its mean and median are around .50 suggesting that

⁹We directly compared the performances of the QQ-test model and submodel $\mathcal{M}_{2yn,r}$ using the Fisher Information Approximation (FIA; Kellen, Klauer, & Bröder, 2013). FIA is a model-selection statistic that penalizes models according to the flexibility that results from their functional form. Although $\mathcal{M}_{2yn,r}$ has one free parameter more than the QQ-test model, it was considered to be the most parsimonious of the two, outperforming the QQ-test model in 92% of the data sets.

across all data sets assimilation or contrast effects are about equally likely to appear. The latter result is not at all implausible given the large variety of questions considered in the data corpus.

The Problem of Aggregating Heterogeneous Data

When researchers aggregate data across items and/or respondents, they are (almost invariably) illegitimately assuming that these observations are all independent and identically distributed. The risks associated with aggregating heterogeneous sources of data have become well-known since Estes (1956), and continuous efforts have been made in order to develop methods that detect (e.g., Smith & Batchelder, 2008) and accommodate heterogeneity (e.g., Klauer, 2006; Lee & Wagenmakers, 2013). However, they are still often overlooked, which sometimes leads to controversial findings (e.g., Davis-Stober, Park, Brown & Regenwetter, 2016; Wulff & van den Bos, in press). In the present case, one could raise the concern that the presence of heterogeneity among respondents, which is virtually certain according to previous work on the modeling of survey data, is likely to affect the modeling of QQ-equality. Given that Wang et al.'s (2014) data corpus consists almost entirely of aggregate data obtained from previously-published surveys across a wide range of topics, we are unable to implement direct statistical tests on respondent heterogeneity.

The fact that the models were fitted to aggregate data, with each individual respondent only contributing with a single data point (one observation in one of the cells of matrix \mathbf{P} or \mathbf{Q}), raises important concerns regarding the interpretation of the parameter estimates. Specifically, it seems unlikely that the probabilities estimated in the QQ-test model or the different repeat-choice models accurately capture the stochastic nature of individuals' responses but rather the relative proportions of different types of individuals. Assuming otherwise, that inter-individual variability is equivalent to the intra-individual variation, amounts to an assumption known as *ergodicity*. The problems this assumption can generate have been well documented in the literature (e.g., Molenaar, 2004).

One important way in which individuals differ is that their response patterns tend to be consistent with one among several qualitatively-distinct subgroups, which in turn implies that matrices \mathbf{P} and \mathbf{Q} emerge from *mixtures of probability distributions*. For example, it is expected that political views (e.g., whether one is a supporter of the American Republican or Democratic party) lead to distinct response patterns for the Clinton and Gore questions. Indeed, it is well established in social-survey research that questions involving politics as well as many other matters of opinion have sub-populations with distinct response patterns. When data are aggregated, the differences between these sub-populations can produce spurious violations of the conditional independence of responses (e.g., Duch, Palmer, & Anderson, 2000).

Concerning QQ-equality per se, it seems extremely unlikely that across several datasets on distinct topics, such a result could emerge from mixtures of subgroups for which QQ-equality does not hold. Moreover, QQ-equality *cannot* be spuriously violated due to aggregation of heterogeneous data, if QQ-equality holds for all subgroups. In order to see this, consider two subgroups A and B , with their respective proportions in the data being π and $1 - \pi$, respectively. If QQ-equality holds for both subgroups, then

$$\begin{aligned}
\text{QQ-value} &= \pi \mathbf{P}_{1,1}^A + (1 - \pi) \mathbf{P}_{1,1}^B - \pi \mathbf{Q}_{1,1}^A - (1 - \pi) \mathbf{Q}_{1,1}^B \\
&+ \pi \mathbf{P}_{2,2}^A + (1 - \pi) \mathbf{P}_{2,2}^B - \pi \mathbf{Q}_{2,2}^A - (1 - \pi) \mathbf{Q}_{2,2}^B \\
&= \pi (\mathbf{P}_{1,1}^A - \mathbf{Q}_{1,1}^A + \mathbf{P}_{2,2}^A - \mathbf{Q}_{2,2}^A) + (1 - \pi) (\mathbf{P}_{1,1}^B - \mathbf{Q}_{1,1}^B + \mathbf{P}_{2,2}^B - \mathbf{Q}_{2,2}^B) \\
&= 0.
\end{aligned}$$

This fortunate situation for QQ-equality does not hold for other properties in the data, however. For instance, conditional independence of the two preferences (e.g., opinion on Gore does not depend on the opinion on Clinton and vice versa) can be violated at the aggregate level, even when each subgroup shows conditionally-independent preferences. These violations are well known and are in fact the basis of many prominent data-analytic methods, such as Latent Class Analysis (Clogg, 1995; Lazarsfeld & Henry, 1968). Latent Class Analysis is a widely-used model family in the social sciences

designed to interpret the dependencies between responses in contingency tables defined by several questions. The responses are then described in terms of a mixture of distributions satisfying conditional independence. In these models, it is precisely the lack of i.i.d. responses that provides the signal that is being modeled. Unfortunately, the experimental design underlying the present data corpus do not provide sufficient degrees of freedom to implement the methods used in Latent Class Analysis.

The fact that certain properties in the data do not hold under heterogeneity can be extremely problematic for models that impose constraints above and beyond QQ-equality, as this implies that they can be spuriously rejected. In order to demonstrate the impact that data heterogeneity can have on model selection, let us consider submodel $\mathcal{M}_{1ar,S}/\mathcal{M}_{2ar,S}$, which grossly misfitted the data (see Figure 4 and Table 2). As previously mentioned, this model assumes conditional independence on preference states $S_{i,j}$ and imposes restrictions on both a and r parameters. We generated artificial data from $\mathcal{M}_{1ar,S}/\mathcal{M}_{2ar,S}$ using four different sets of parameter values representing distinct subgroups. For example, let us assume that these data come from a Clinton-Gore poll and that the four groups differ in both their attitudes towards Clinton and Gore and their repeat-choice probabilities: 1) “pure” Democrats (i.e., large probability of responding with “yes” to both questions) with a strong assimilation tendency; 2) Democrats who were put off by Clinton’s extramarital affairs, with moderate contrast tendency, 3) “pure” Republicans in strong opposition to the Democrats and with a strong contrast tendency; 4) Republicans who liked Clinton but did not trust Gore, with moderate contrast tendency. Table 3 gives an overview of proportions and the exact parameters of each group comprising the mixture. We randomly generated one-thousand \mathbf{P} and \mathbf{Q} matrices with 200 respondents per matrix (i.e., 400 in total per simulated dataset) using the mixture probabilities π given in Table 3 and fitted them with both $\mathcal{M}_{1ar,S}/\mathcal{M}_{2ar,S}$ and $\mathcal{M}_{1ar}/\mathcal{M}_{2ar}$. As shown in Figure 8, model fits produced highly-skewed p -value distributions, with statistically-significant misfits ($p < .05$) occurring in 40% of the datasets for $\mathcal{M}_{1ar,S}/\mathcal{M}_{2ar,S}$ (Left Panel) and 39% for $\mathcal{M}_{1ar}/\mathcal{M}_{2ar}$ (Center Panel). As expected, the average QQ-value was 0 (Right

Panel) and the QQ-test model was rejected in approximately 5% of the data sets, in line with the nominal rejection rate. These simulation results show that the presence of individual heterogeneity in the data can lead to an apparent failure of a model, despite the fact that all the responses were generated by processes in line with the assumptions of that model. This situation suggests that the present rejection of constrained models such as $\mathcal{M}_{1ar,S}/\mathcal{M}_{2ar,S}$ and $\mathcal{M}_{1ar}/\mathcal{M}_{2ar}$ should be seen with skepticism.

It would be unwise to interpret the vulnerability of some repeat-choice models to respondent heterogeneity as a shortcoming or as a disadvantage relative to the QQ-test model. Ultimately, we should strive for models to succeed or fail according to their ability to capture theoretically-meaningful information in the data, not distortions introduced by the researchers when incorrectly assuming that there is no individual heterogeneity.

Discussion

The importance of the test of the QQ-equality reported by Wang et al (2014) comes from the expectation that such an a priori constraint would fail when tested with such a large and diverse data corpus. Wang et al.'s result became even more impressive when showing that plausible classic-probability candidates that relied on response repetition and anchoring effects failed to account for such data (see also Busemeyer et al., 2009). The failure of these alternative models indicates that the development of such models is far from being a trivial endeavour. The difficulty here is not necessarily the development of a model that can fit the data corpus, as that can be achieved by specifying an overly flexible model with little or no constraints. Instead, the challenge lies in the development of a model – based on reasonable assumptions – that *expects* the presence of QQ-equality. Upon presenting their results, Wang et al. (2014) urged researchers to develop alternative approaches based on classic probability theory.

The present work answers Wang et al.'s call for alternative models by showing that the QQ-equality constraint also emerges under a classic-probability framework when assuming specific mechanisms that introduce response dependencies. Specifically,

we showed that some repeat-choice models assuming that respondents can change their second response based on a detected similarity between the two questions also predicts QQ-equality as a parameter-free prediction. This prediction holds when the similarity-based dependencies are independent of question order but determined by the respondents' latent preferences (\mathcal{M}_1 and its respective special cases). Furthermore, we also showed that repeat-choice models assuming that response dependencies are moderated by question order and the agreement/disagreement of the latent preferences concerning the two questions (\mathcal{M}_2 and its respective special cases) do not impose QQ-equality but establish it as a highly-likely event and provide an at least as good account as the QQ-test model. The success of such models does not represent a detriment of Wang et al.'s quantum-model account; instead, their success helps to delineate some of the conditions under which an alternative classic-probability account can succeed, and allows for the development of models based on both approaches to occur in tandem. Similar work has been done in the study of superposition effects in memory, where quantum and classic-probability accounts have been proposed (Brainerd, Wang, & Reyna, 2013; Kellen, Singmann, & Klauer, 2014).

In addition to the development of alternative models, we discussed the challenge that the (unaccounted) presence of individual heterogeneity represents for the modeling of this kind of choice data, as some of the simpler models can be spuriously rejected at high rates. Given the possibility of such spurious rejections, it seems critical to first and foremost develop tests that focus on individual data and enable us to characterize the cognitive processes underlying individuals' responses (e.g., estimate individual choice probabilities). For instance, individual-level data would allow us to have a better estimate of how many individuals manifest order effects and whether these effects imply QQ-equality.¹⁰

The use of aggregated data leads to somewhat ambiguous interpretations of

¹⁰It worth noting that an extension of the model to three questions (something already suggested by Newell, Ravenzwaaij, & Donkin, 2013) provides additional degrees of freedom (a simple extension to three questions would result in 42 degrees of freedom) that would allow for certain mixtures to be tested (see Clogg, 1995; Lazarsfeld & Henry, 1968). However, this is still far from a full account of individual differences.

parameters. As previously mentioned, the fact that each person only provided two responses indicates that the probabilities described by the repeat-model parameters are more likely to represent proportions in the sample (e.g., parameter a indicates the proportion of respondents that based their second response on the first one) rather than stochastic processes that occur on an individual level. This ambiguity coming from the use of aggregated data glosses over fundamental issues that were recently discussed by Khrennikov, Basieva, Dzhafarov, and Busemeyer (2014): Khrennikov et al. distinguish between tasks or questions in which responses are expected to be replicated by the same respondent with some probability, and responses that are expected to be replicated with virtual certainty (i.e., probability 1). For example, consider the response to the question “do you like chocolate?”, which is expected to be stable for each respondent across replications. Now, contrast it with the responses individuals provide in a psychophysical task, when attempting to judge whether an auditory signal was just presented, which are expected to differ across replications. This distinction is critical when establishing predictions for a three-question sequence A-B-A: As shown by Khrennikov et al., when respondents are expected to produce the same response to both A questions, the quantum-probability model does not expect order effects to be observed. Such effects are only expected when responses to question A are expected to vary across replications with non-negligible probability. This difference does not hold for the repeat-choice models given that response dependencies only occur at the level of the second question as a way to reconcile or differentiate views given the first response given. It would make little sense to assume that any individual respondent engages in some process of differentiation/reconciliation when producing a second response, only to immediately afterwards change the first response based upon which such differentiation/reconciliation processes were conducted.

The issue of response replicability and the differential predictions coming from it highlight an important theoretical challenge for the quantum-probability model. Specifically, one could argue that the model should not predict order effects in the type questions addressed in the data corpus used by Wang et al. (2014). Khrennikov et al.’s

(2014) discuss different solutions to these problems, but overlooked the issue of data aggregation and individual differences. Note that according to Wang and Busemeyer (2015), the measurement complementarity that results in order effects with QQ-equality is expected to be present when respondents do not have clear preferences and have to construct them on the fly. This means that the QQ-equality-constrained order effects observed in the aggregate data can be entirely due to a proportion of the respondent sample that does not possess stable preferences regarding the domains of the questions being posed (e.g., individuals that do not care much about politics might not have stable preferences concerning Clinton and Gore). One interesting question then is whether we can observe order effects of differing magnitudes in the same sample of respondents, and whether the magnitude of these order effects is related with their familiarity and/or engagement with the questions' topics.

The reliance on aggregated data therefore represents “a plague on both our houses”, although with different symptoms. In the case of classic-probability models, it produces spurious rejections at high rates. In the case of the quantum-probability model it places it in a position of apparent self-contradiction in the sense that the order-effects used to support it occur with questions in which it would be more reasonable for the model to predict no order effects at all. In order to overcome these problems, future work needs to focus on establishing the basic effects on an individual level. One way to achieve this would be to move away from questions like the ones included in the basic data corpus, and instead rely on responses to perceptual categorization tasks that allow for replications to occur under reasonable conditions, and for the similarity between the different stimuli to be carefully controlled for (e.g., Busemeyer et al., 2009).

References



- Bamber, D. & Van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, *44*, 20–40.
doi:10.1006/jmps.1999.1275
- Brainerd, C. J., Wang, Z., & Reyna, V. F. (2013). Superposition of episodic memories: overdistribution and quantum models. *Topics in Cognitive Science*, *5*, 773–799.
doi:10.1111/tops.12039
- Busemeyer, J. R. & Bruza, P. D. (2012). *Quantum models of cognition and decision*. Cambridge University Press.
- Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological review*, *118*(2), 193–218. doi:10.1037/a0022542
- Busemeyer, J. R., Wang, Z., & Lambert-Mogiliansky, A. (2009). Empirical comparison of markov and quantum models of decision making. *Journal of Mathematical Psychology*, *53*(5), 423–433. doi:10.1016/j.jmp.2009.03.002
- Busemeyer, J. R., Wang, Z., Pothos, E. M., & Trueblood, J. S. (2015). The conjunction fallacy, confirmation, and quantum theory: comment on tentori, crupi, and russo (2013). *Journal of Experimental Psychology: General*, *144*, 236–243.
doi:10.1037/xge0000035
- Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & E. Soberl (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York: Plenum.
- Davis-Stober, C. P. (2009). Multinomial models under linear inequality constraints: applications to measurement theory. *Journal of Mathematical Psychology*, *53*, 1–13.
- Davis-Stober, C. P., Park, S., Brown, N., & Regenwetter, M. (2016). Reported violations of rationality may be aggregation artifacts. *Proceedings of the National Academy of Sciences*, *113*(33), E4761–E4763. doi:10.1073/pnas.1606997113

- Duch, R. M., Palmer, H. D., & Anderson, C. J. (2000). Heterogeneity in perceptions of national economic conditions. *American Journal of Political Science*, *44*, 635–652. doi:10.2307/2669272
- Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140. doi:10.1037/h0045156
- Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review*, *20*, 693–719. doi:10.3758/s13423-013-0407-2
- Kellen, D., Singmann, H., & Klauer, K. C. (2014). Modeling source-memory overdistribution. *Journal of Memory and Language*, *76*, 216–236. doi:10.1016/j.jml.2014.07.001
- Khrennikov, A., Basieva, I., Dzhafarov, E. N., & Busemeyer, J. R. (2014). Quantum models for psychological measurements: an unsolved problem. *PloS one*, *9*, e110909. doi:10.1371/journal.pone.0110909
- Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, *71*, 7–31. doi:10.1007/s11336-004-1188-3
- Lazarsfeld, P. F. & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.
- Lee, M. D. (2016). Bayesian outcome-based strategy classification. *Behavior Research Methods*, *48*, 29–41. doi:10.3758/s13428-014-0557-9
- Lee, M. D. & Wagenmakers, E. J. (2013). *Bayesian data analysis for cognitive science: A practical course*. New York: Cambridge University Press.
- Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. *Measurement*, *2*, 201–218. doi:10.1207/s15366359mea0204_1
- Moore, D. W. (2002). Measuring new types of question-order effects: additive and subtractive. *Public Opinion Quarterly*, 80–91. doi:10.1086/338631

- Newell, B. R., van Ravenzwaaij, D., & Donkin, C. (2013). A quantum of truth? querying the alternative benchmark for human cognition. *Behavioral and Brain Sciences*, *36*(03), 300–302. doi:10.1017/s0140525x12003068
- Pothos, E. M. & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, *36*(03), 255–274. doi:10.1017/s0140525x12001525
- Singmann, H. & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models with R. *Behavior Research Methods*, *45*, 560–575. doi:10.3758/s1342801202590
- Smith, J. B. & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*, 713–731. doi:10.3758/PBR.15.4.713
- Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. doi:10.1017/cbo9780511808098.003
- van de Schoot, R., Hoijsink, H., & Deković, M. (2010). Testing inequality constrained hypotheses in sem models. *Structural Equation Modeling*, *17*(3), 443–463. doi:10.1080/10705511.2010.489010
- Wang, Z. & Busemeyer, J. (2015). Reintroducing the concept of complementarity into psychology. *Frontiers in Psychology*, *6*, 1822. doi:10.3389/fpsyg.2015.01822
- Wang, Z., Solloway, T., Shiffrin, R. M., & Busemeyer, J. R. (2014). Context effects produced by question orders reveal quantum nature of human judgments. *Proceedings of the National Academy of Sciences*, *111*(26), 9431–9436. doi:10.1073/pnas.1407756111
- Wulff, D. U. & van den Bos, W. (in press). Modeling choices in delay discounting. *Psychological Science*.

Table 1

Order effects and the QQ-values

Data			Differences		
P	$P_{.,1}$ (Gore “yes”)	$P_{.,2}$ (Gore “no”)	P - Q	$P_{.,1} - Q_{.,1}$	$P_{.,2} - Q_{.,2}$
$P_{1.}$ (Clinton “yes”)	.4899	.0447	$P_{1.} - Q_{1.}$	-.0726	.0192
$P_{2.}$ (Clinton “no”)	.1767	.2886	$P_{2.} - Q_{2.}$	-.0224	.0756
Q	$Q_{.,1}$ (Gore “yes”)	$Q_{.,2}$ (Gore “no”)	P - Q	$P_{.,1} - Q_{.,1}$	$P_{.,2} - Q_{.,2}$
$Q_{1.}$ (Clinton “yes”)	.5625	.0255	$P_{1.} - Q_{1.}$		
$Q_{2.}$ (Clinton “no”)	.1991	.2130	$P_{2.} - Q_{2.}$		

Note. Data from the Clinton (Question A) and Gore (Question B) poll showing order effects. Response matrix **P** concerns question-order Clinton-Gore ($A \rightarrow B$) whereas response matrix **Q** concerns question-order Gore-Clinton ($B \rightarrow A$). The order effect (i.e., the difference between both data tables) is shown on the difference matrix **P - Q** in the upper-right part. The two QQ-values are given by summing the differences along the diagonals as displayed in the lower right table. As the sum of all differences needs to be zero, the two QQ-values are the inverse of each other. For the data shown here the QQ-values are $-.0726 + .0756 = .0030$ and $.0192 - .0224 = -.0032$. In the current manuscript we always use the values of the main diagonal (i.e., .0030 in this example).

Table 2
Goodness of Fit of Repeat-Choice Models

Model	Parameters	Summed G^2	ΔG^2
\mathcal{M}_0	19	0	-75.91
$\mathcal{M}_1^*/\text{QQ-test model}^*$	11	75.91	0
\mathcal{M}_{1a}^*	8	75.91	0
\mathcal{M}_{1r}^*	8	76.20	0.29
\mathcal{M}_2	11	79.04	3.13
\mathcal{M}_{2a}	8	79.04	3.13
\mathcal{M}_{2r}	8	85.72	9.82
\mathcal{M}_{2yn}	7	85.72	9.82
$\mathcal{M}_{2yn,a}$	6	133.75	57.84
$\mathcal{M}_{2yn,r}$	6	88.30	12.39
$\mathcal{M}_{2yn,S}$	6	1178.21	1102.30
$\mathcal{M}_{1ar}^*/\mathcal{M}_{2ar}^*$	5	336.73	260.82
$\mathcal{M}_{1ar,S}^*/\mathcal{M}_{2ar,S}^*$	4	1436.70	1360.78

Note. Models with an asterisk (*) enforce QQ-equality as a parameter-free prediction.

Summed G^2 : Summed G^2 of the model fitted to the 72 datasets reported by Wang et al. (2014). ΔG^2 : Difference between the summed G^2 of a model and the $\mathcal{M}_1/\text{QQ-test}$ model.

Table 3

Hypothetical Parameters of the Restricted Repeat-Choice Model $\mathcal{M}_{1ar,S}$

Parameter	“Pure” Democrats	“Put off” Democrats	“Pure” Republicans	“Clinton” Republicans
$S_{1.}$.90	.50	.10	.90
$S_{.1}$.90	.90	.10	.60
a	.70	.30	.60	.35
r	.80	.30	.90	.20
π	.25	.25	.45	.05

Note. In each step we simulated 200 individual respondents per matrix with the given parameter values (i.e., total $N = 400$ per simulated dataset), $\pi =$ mixture weight.

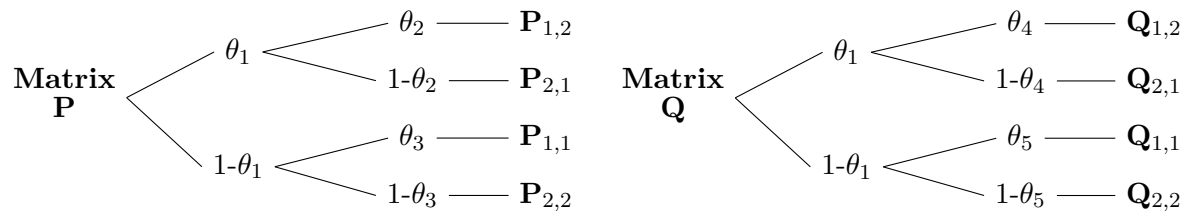


Figure 1. QQ-Test Model. Subscripted parameters θ at the nodes denote branch probabilities, and terminal nodes denote the observed response categories.

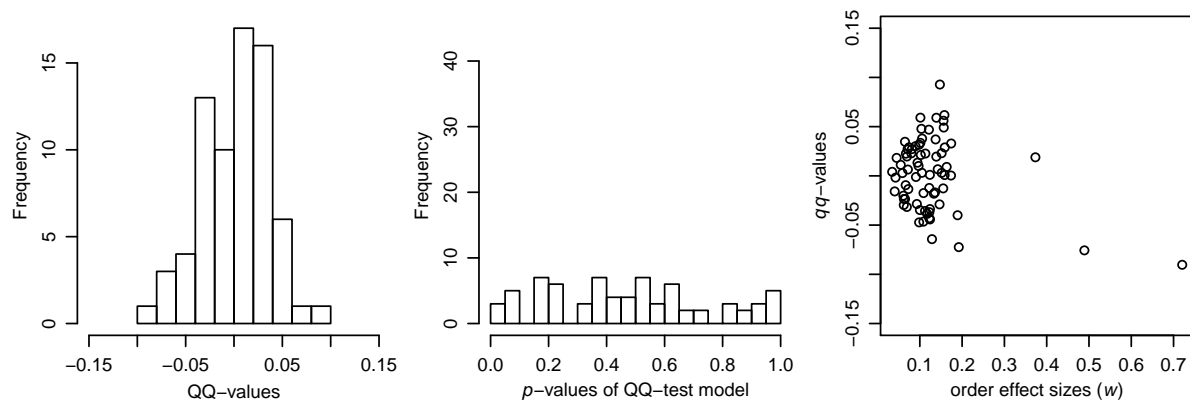


Figure 2. Analysis of the 72 datasets reported by Wang et al. (2014). Left panel: Distribution of QQ-values. Center panel: Distribution of p -values for the QQ-test model. Right panel: QQ-values against size of the order effect (in terms of χ^2 effect size Cohen's w). The three outliers in terms of order effect size are (in decreasing order): the abortion question, one of the two laboratory experiments, and the black-white Gallup poll (see Wang et al., 2014, supporting information).

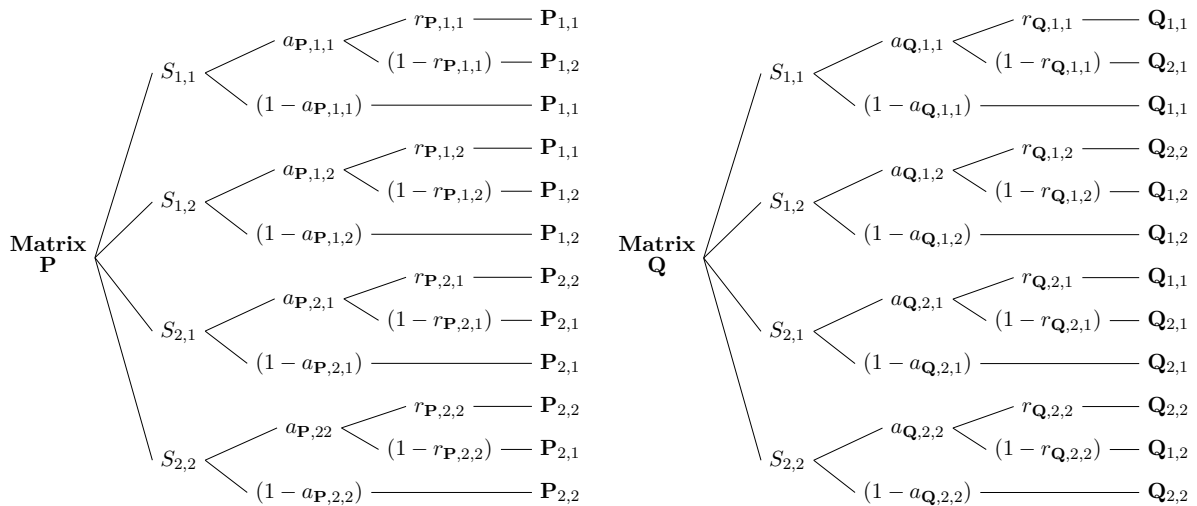


Figure 3. Model \mathcal{M}_0 . Subscripted parameters S , a , and r denote branch probabilities, and terminal nodes denote the observed response categories.

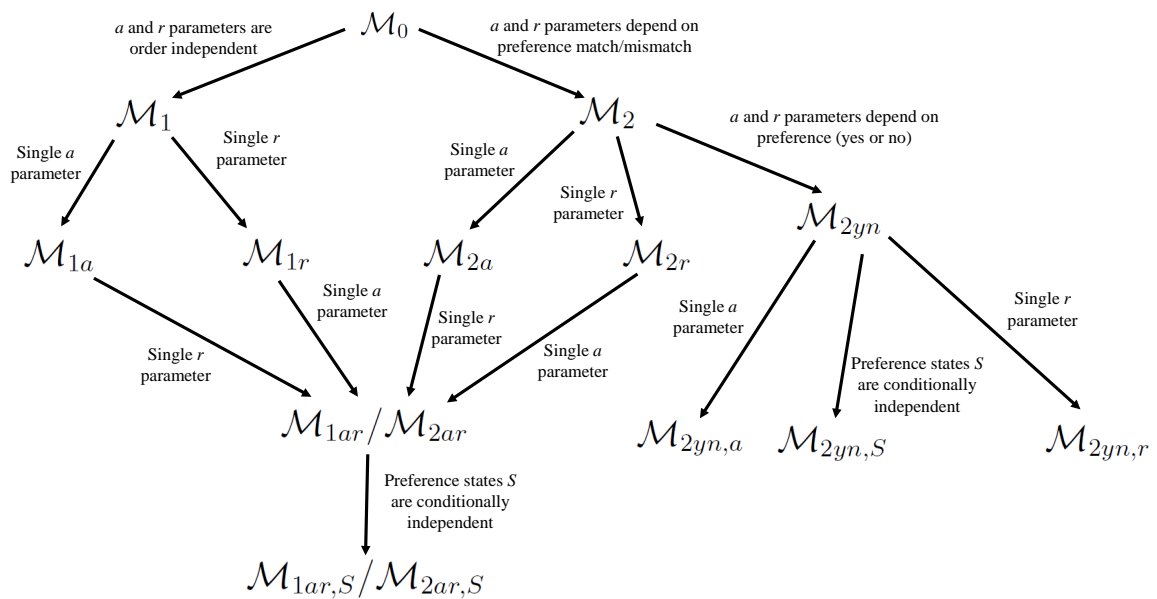


Figure 4. A Non-Exhaustive Hierarchy of Repeat-Choice Models. Parameter subscripts are omitted for purposes of clarity.

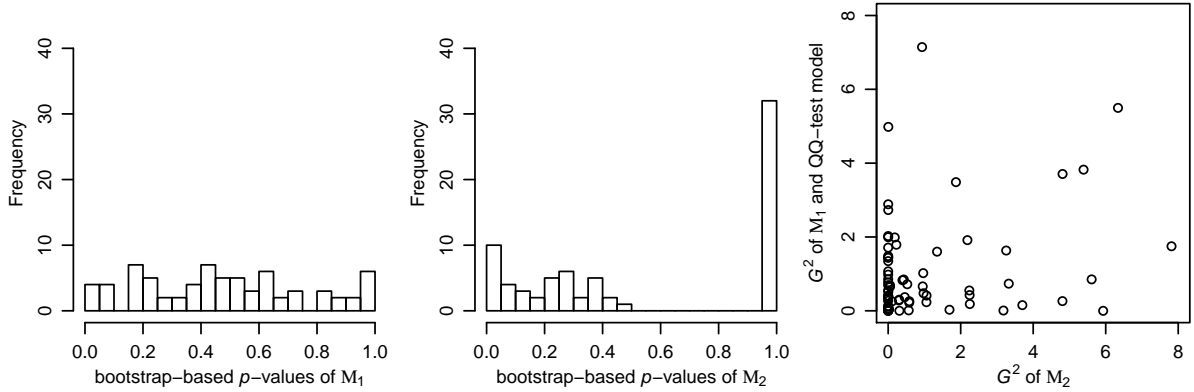


Figure 5. Comparison of repeat-choice models \mathcal{M}_1 and \mathcal{M}_2 fitted to the 72 datasets of Wang et al. (2014). Left panel: (bootstrap-based) p -value distribution of \mathcal{M}_1 (which enforces QQ-equality and is equivalent to the QQ-test model). Center panel: (bootstrap-based) p -value distribution of \mathcal{M}_2 (which does not enforce QQ-equality). Right panel: direct comparison of G^2 fit statistics for both models.

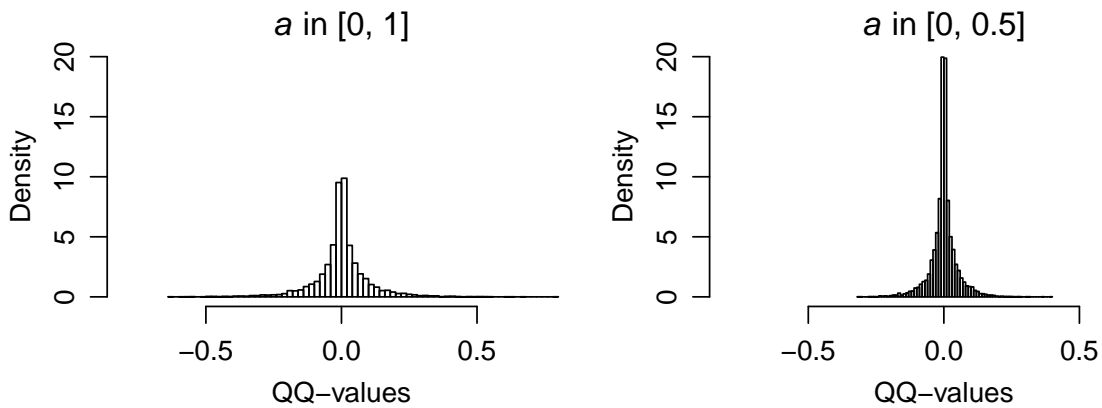


Figure 6. Distribution of 10,000 simulated QQ-values from the \mathcal{M}_2 model. Left panel: using flat priors on all parameters; Right panel: a is drawn uniformly from $[0, 0.5]$, flat priors on all other parameters.

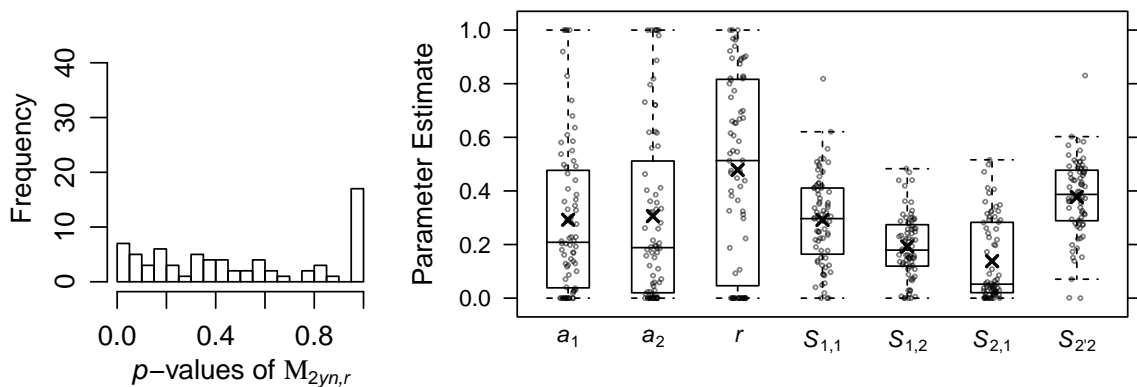


Figure 7. Identifiable submodel $\mathcal{M}_{2yn,r}$ in which repeat-choice probability a are a function of the first response (“yes” versus “no”). Left panel: (bootstrap-based) p -value distribution. Right panel: Parameter estimates. Individual parameter estimates are plotted in the background with horizontal jitter and 50% transparency. The boxplots show the upper and lower quartiles as well as the median, the \times shows the mean.

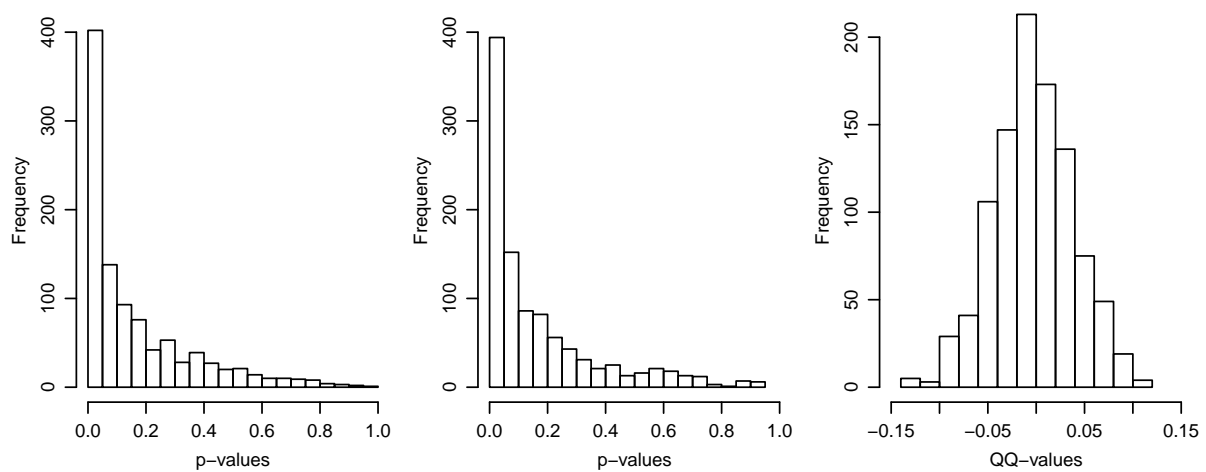


Figure 8. Results obtained with artificial data from mixture distribution (see Table 3). Left panel: Distribution of p -values for the repeat-choice submodel with restricted S , a , and r parameters ($\mathcal{M}_{1ar,S}/\mathcal{M}_{2ar,S}$). Center panel: Distribution of p -values for the repeat-choice submodel with restricted a and r parameters ($\mathcal{M}_{1ar}/\mathcal{M}_{2ar}$). Right panel: Distribution of QQ-values.