Comment on Regenwetter & Robinson (in press): Probabilistic error-free models are also subject to logical fallacies

Clintin P. Davis-Stober

University of Missouri

David Kellen

Syracuse University

Simon Segert

University of Missouri

Author Note

Clintin P. Davis-Stober, Department of Psychology. David Kellen, Department of Psychology. Simon Segert, Department of Psychology. Correspondence should be sent to Clintin P. Davis-Stober (stoberc@missouri.edu) or David Kellen (davekellen@gmail.com). Clintin P. Davis-Stober received support from the National Science Foundation (SES-14-59866) and the National Institutes of Health (K25AA024182). David Kellen received support from the Swiss National Science Foundation Grant 100014_165591.

Abstract

Regenwetter and Robinson (in press) consider a set of heuristics commonly used by psychologists to connect raw data with statistical models. They show that these heuristics are subject to two major reasoning fallacies: the *fallacy of sweeping generalization* and *the fallacy of composition.* To place theory testing on solid, logical ground, Regenwetter and Robinson promoted the use of "probabilistic, error-free" models. We show that this class of models is subject to the very same fallacies considered by Regenwetter and Robinson. We propose a solution for diagnosing when probabilistic error-free models are subject to these fallacies.

Keywords: heterogeneity, data aggregation, probabilistic choice, random preference

Regenwetter and Robinson (in press; henceforth RR) demonstrated that three methods of data aggregation, commonly used by behavioral decision researchers, are subject to the logical fallacies of (i) *composition* and (ii) *sweeping generalization.* Specifically, they examined logical problems intrinsic to: 1) tallying, or averaging, choice responses across multiple participants, 2) examining the modal response in a group of participants, and 3) combining evidence for distinct predictions of a theory across experiments and/or stimuli sets. We agree with RR that such methods of data composition are subject to these logical fallacies and that superior methods of theory testing are sorely needed, not just for decision research but for psychological research more generally.

In light of these challenges, RR offered a "blueprint" to guide researchers in formulating testable theories that avoid these logical fallacies when relying on aggregated (i.e., tallied) responses per decision problem across individuals. The most prominent approach advocated by RR were "probabilistic models without response error" (henceforth PRE), a class of models with a storied history in the decision sciences (e.g., Falmagne, 1978; Loomes & Sugden, 1995). According to these models, individual preferences vary probabilistically across observations, but are expressed in the observed responses without error. The goal of the present theoretical commentary is to demonstrate that the PREs suggested by RR should be used with caution given that they are also vulnerable to the very same logical fallacies they discussed. Like RR, our point is a general one and not necessarily connected to the literature on description- and experience-based decision making that they focused on.[1]

[1] The only comment we will make here with respect to this specific literature is that it often involves individual participants making decisions based on *unique sequences* of experienced outcomes with variable length. Also, in most of these studies there is the possibility that individuals do not experience all the outcomes associated with the options in each decision problem. All of this information is effectively discarded when aggregating data, raising questions on the exact meaning of any subsequent analysis. Because RR's analyses do not address any of these specific problems and take the data at face value, we see their results more as a proof of concept.

Specifying and Testing Probabilistic Error-Free Models

In the case of binary preferences, PRE models are implemented as follows: First, the researcher establishes an experiment with n decision problems that individuals will be confronted with. Each decision problem is comprised of two choice options a and b. Each individual's preference for the ordered pair (a, b) is coded as either 1 (a preferred to b) or 0 (b preferred to a). The space of all possible preference patterns corresponds to $\{0, 1\}^n$. To establish the PRE, the researcher enumerates all of the predicted preference patterns across the n decision problems that are consistent with a given hypothesis of interest, e.g., all preference patterns consistent with over-weighting of probabilities under Cumulative Prospect Theory (Tversky & Kahneman, 1992). Let us denote this set of preference patterns as V_{in}. Note that V_{in} is a subset of $\{0, 1\}^n$. A PRE is deemed testable when there exists a non-empty subset $V_{out} = \{0, 1\}^n - V_{in}$ comprised of preference patterns *not* included in the PRE. Finally, the assumed variability in preferences is introduced by allowing an arbitrary probability distribution over all preference patterns within V_{in}. It follows that the probability of preferring a over b, denoted as $P(a \succ b)$, in a decision problem is given by the probabilities p_v of all preference patterns v for which a is preferred over b:

$$P(a \succ b) = \sum_{v \in V_{in}:\, a \succ b} p_v, \tag{1}$$

with $\sum_{v \in \{0,1\}^n} p_v = 1, p_v \geq 0$. The ability of the PRE to account for the aggregated responses across decision problems can then be assessed by means of the joint binomial likelihood function that follows from it, using a host of classic (Davis-Stober, 2009) and/or Bayesian methods (Myung, Karabatsos & Iverson, 2005; Klugkist & Hoijtink, 2007).

It can be demonstrated that, if individuals in a population have heterogeneous preference patterns and yet all such preferences are consistent with the hypothesis established by the model (e.g., overweighting of small probabilities), then the aggregated/tallied responses will *always satisfy* the PRE model - thus avoiding a fallacy of composition and/or sweeping generalization, depending upon the context. This fortunate result follows from the fact that any PRE, by definition, is a convex set: Said simply, any probabilistic mixture of points from this set (i.e., any aggregation of individual preferences included in a PRE) will necessarily still be included in it.

An alternative case is perhaps more interesting. Depending upon the number of choice pairs and the set of predictions, it is possible for aggregated choice data (tallied or model choice) to satisfy the PRE model, and yet *no single individual* conforms to the hypothesis in question. We show that an improper application of PRE models can lead researchers to commit both logical fallacies:

- *Sweeping Generalization Fallacy*: The aggregate responses satisfy the PRE model so at least some individuals in the population must satisfy it as well.
- *Composition Fallacy*: All individuals violate the model so the aggregate must violate it as well.

For example, consider the subset of decision problems obtained from the Cash II stimuli. As shown in Table 1, the hypothesis of Cumulative Prospect Theory with overweighting encompasses a wide range of preference patterns for these decision problems. However, it is possible to conceive of individuals that do not conform to any of the preferences under this hypothesis. For example, the preferences patterns of the four individuals in Table 1 are not included in V_{in} and yet the aggregate responses are perfectly in line with it (they can be perfectly described as a mixture of patterns 1 and 6). The problem here is that although a mixture of preferences included in the PRE is bound to satisfy the latter, it is possible that a mixture of preferences from outside of the PRE will also satisfy it. The PRE would then provide a perfect account of these aggregated preferences, giving the false idea that the PRE model holds. This problem was first identified by Birnbaum (2011) in the context of testing the transitivity axiom under a specific PRE model.

It then follows that the kind of inferences one can legitimately make from the successes or failures of PREs are very distinct: The failure of a given PRE to account

Table 1
Preference Patterns and Response Data for a Selection of Decision Problems from Cash II

	Decision Problems			
PRE's Preference Pattern v	AD	AE	BC	BD
1	0	0	0	0
2	0	0	1	0
3	1	0	0	0
4	1	0	1	0
5	1	1	1	0
6	1	1	1	1
Non-Conforming Individuals				
Participant 1	0	1	0	0
Participant 2	1	0	1	1
Participant 3	1	1	0	0
Participant 4	0	0	1	1
Aggregate	0.5	0.5	0.5	0.5

Note. Preference patterns associated with Cumulative Prospect Theory with overweighting of small probabilities (for details on the parametric assumptions, see Regenwetter & Robinson, in press). A decision problem *ab* is comprised of options *a* and *b*. A preference value of 1/0 denotes a preference for option *a*/*b*.

for a set of aggregate responses constitutes a scenario of *strong inference* in the sense that the predictions of the model must be violated by at least one or more individuals. In other words, the properties represented by the PRE cannot hold for all individuals. Note that the informativeness of these violations does not hinge on the statistical methods, either classical or Bayesian, adopted by the researcher. Davis-Stober, Morey, Gretton, and Heathcote (2016) provide a more detailed discussion in the context of state-trace analysis.

In contrast, the success of a PRE constitutes an ambiguous scenario, as it is possible that: i) all individuals satisfy the hypothesis, ii) some individuals satisfy the hypothesis, or iii) no individuals satisfy the hypothesis. As before, this ambiguity does not hinge on the statistical methods used. One reason why this ambiguity is likely to be overlooked by researchers is the fact that the PRE being tested might be perceived as extremely parsimonious in these sense that the number of preference patterns contained in V_{in} might be extremely small in comparison to total number of possible preference patterns, 2^n (for a discussion, see Roberts and Pashler, 2000). Although parsimony is a desirable attribute, it does not dismiss the possibility of mimicry such that the observed aggregate responses can be described in terms of the preferences included in V_{out}. For example, Regenewetter, Dana, and Davis-Stober (2011) proposed a PRE that only accounted for $\frac{120}{1024} = 0.12$ of all possible preference patterns, indicating a strong potential of rejection. But as shown by Birnbaum (2011), the observed data could nevertheless be accounted for by a mixture of preference patterns not included in the PRE.[2] In the absence of a formal evaluation of the model and experimental design stating otherwise, it is likely that the possibility of mimicry increases along with the total number of preference patterns contained in V_{out}.[3]

One Remedy

One way to ameliorate this ambiguity problem is to consider the extent to which the aggregate data are compatible with alternative models that allow for preference patterns included in V_{out}. One way to quantify this compatibility for a given vector $\boldsymbol{x}$ of aggregated responses involves computing a *worst-case probability* Q for individual preferences in V_{out}. Formally, $Q(\boldsymbol{x}) = \sup \sum_{v \in V_{out}} p_v$ where the supremum is taken over all probability distributions p_v on $\{0,1\}^n$ such that $\sum_{v \in \{0,1\}^n} p_v v = \boldsymbol{x}$. The value of $Q(\boldsymbol{x})$ is the solution of a linear programming problem in the probability variables p_v, in which we maximize the probability of preference patterns p_v in V_{out}, with the

[2] A discussion on the plausibility of Birnbaum's (2011) argument and its testability is beyond the scope of the present paper and therefore will not be pursued here - see Regenewetter, Dana, Davis-Stober, & Guo (2011) for their reply.

[3] The two inference scenarios discussed here apply equally to within- or between-subject designs. The interpretations are slightly different, but the argument is essentially the same.

requirement that a perfect account of the data vector $\boldsymbol{x}$ is achieved.[4] In canonical form:

$$\begin{aligned}
\texttt{Maximize} \quad & \sum_{v \in V_{out}} p_v, \\
\texttt{subject to} \quad & \sum_{v \in \{0,1\}^n} p_v = 1, p_v \geq 0, \\
\texttt{and} \quad & \sum_{v \in \{0,1\}^n} p_v v = \boldsymbol{x}.
\end{aligned}$$

As an example, for the aggregate responses in Table 1, it turns out that $Q = 1$, indicating that the data could be exclusively characterized by preference patterns included in V_{out}. In contrast, consider the aggregate response vector $\boldsymbol{x} = [.80, .95, .20, .90]$. In this case, $Q = .15$, indicating that in the worst-case scenario we would only expect 15% of the individuals to be at odds with the PRE.[5] Note that our Q function is not a statistical method per se. It could be applied to both aggregate data, e.g., tallied choice frequencies, as well as hypothetical, population-level parameter values. To be clear, large values of Q do not constitute evidence that the aggregated data were formed by preference patterns from V_{out}, only that this possibility exists and the researcher needs to be aware of it from an interpretative standpoint. On the other hand, small values of Q are indicative that these aggregation problems cannot happen (at least not for a majority of individuals), given the preference patterns specified in V_{in}. For PRE models that generate large values of Q, it becomes imperative to compare the existing PRE model to alternative theories that include preference patterns from V_{out}.

Discussion

We agree with RR that PRE models are useful tools for analysis. Our argument isn't intended to "throw the baby out with the bathwater" but to clarify the inferences that can be legitimately made from the models' successes and failures. The rejection of a PRE provides strong evidence against the hypothesis that all individuals are characterized by the preference patterns included in the model. In contrast, the success

[4] Note that for any $r \in [0, 1]$, the set $\{x \in [0, 1]^n : Q(\boldsymbol{x}) \geq r\}$ is a convex polytope with vertex set $((1 - r)V_{in} + rV_{out}) \cup V_{out}$.

[5] The Supplemental Material provides an R implementation for the computation of Q that replicates the examples concerning Table 1.

of a PRE is ambiguous in itself and requires the evaluation of alternative accounts. Researchers are likely to overlook this ambiguity given that it is not in line with the problems typically associated with data aggregation. In most of the scenarios discussed in the literature, aggregation leads to spurious rejections of models rather than spurious "acceptances" (e.g., Estes, 1956; Heathcote, Brown, & Mewhort, 2000; Kellen, Singmann, & Batchelder, in press).

The present difficulties associated with the interpretation of successful PRE models should not be seen as an intrinsic shortcoming of this model class, but part of the nefarious cost of relying on aggregated data. When one only has access to the aggregate response proportions per decision problem rather the individual response profiles across decision problems, one cannot say much about the data. The informativeness of any modeling approach will always be bound by the data it is applied to. But as discussed in detail by RR, there is a widespread naivety with respect to these bounds that enables different kinds of fallacies to take place. When data are not aggregated, a wide range of modeling approaches become available in researchers' toolboxes. For example, there is a variety of individual-subject classification (e.g., Brown, Park, Steinley, & Davis-Stober, in press; Hilbig & Moshagen, 2014; Lee, 2016) and hierarchical-Bayesian modeling methods (e.g., Kellen, Pachur, & Hertwig, 2016) that could also be adopted in order to characterize individual-level response patterns and the latent representations assumed to underlie them. Given these opportunities, RR's critique constitutes a timely invitation for setting old habits aside.

References

Birnbaum, M. (2011). Testing mixture models of transitive preference. comment on Regenwetter, Dana, and Davis-Stober (2011). *Psychological Review*, *118*, 675–683.

Brown, N., Park, S., Steinley, D., & Davis-Stober, C. P. (in press). Modeling between-subject variability in decision strategies via statistical clustering: a p-median approach. *Journal of Behavioral Decision Making*.

Davis-Stober, C. P., Morey, R. D., Gretton, M., & Heathcote, A. (2016). Bayes factors for state-trace analysis. *Journal of Mathematical Psychology*, *72*, 116–129.

Davis-Stober, C. (2009). Analysis of multinomial models under inequality constraints: applications to measurement theory. *Journal of Mathematical Psychology*, *53*, 1–13.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological bulletin*, *53*(2), 134–140.

Falmagne, J.-C. (1978). A representation theorem for finite random scale systems. *18*, 52–72.

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: the case for an exponential law of practice. *Psychonomic bulletin & review*, *7*(2), 185–207.

Hilbig, B. E. & Moshagen, M. (2014). Generalized outcome-based strategy classification: comparing deterministic and probabilistic choice models. *Psychonomic bulletin & review*, *21*, 1431–1443.

Kellen, D., Singmann, H., & Batchelder, W. H. (in press). A classic-probability account of mirrored (quantum-like) order effects in human judgments. *Decision*.

Kellen, D., Pachur, T., & Hertwig, R. (2016). How (in) variant are subjective representations of described and experienced risk and rewards? *Cognition*, *157*, 126–138.

Klugkist, I. & Hoijtink, H. (2007). The bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*(12), 6367–6379.

Lee, M. D. (2016). Bayesian outcome-based strategy classification. *Behavior research methods*, *48*(1), 29–41.

Loomes, G. & Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, *39*, 641–648.

Myung, J., Karabatsos, G., & Iverson, G. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, *49*, 205–225.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*, 42–56.

Regenwetter, M., Dana, J., Davis-Stober, C. P., & Guo, Y. (2011). Parsimonious testing of transitive or intransitive preferences: Reply to Birnbaum (2011). *Psychological Review*, *118*(4), 684–688.

Regenwetter, M. & Robinson, M. (in press). The construct-behavior gap in behavioral decision research: a challenge beyond replicability. *Psychological Review*.

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Supplemental Material

The purpose of this section is to provide a simple R implementation for computing the worst-case probability Q for examples concerning Table 1. Let us assume a model that describes a response vector $\boldsymbol{x}$ for $n = 4$. The model describes this vector as a mixture of the sixteen preference patterns v that constitute the set $\{0, 1\}^n$ of all possible preference patterns in this example. Each preference pattern is associated with a probability p_v, with $\sum_{v \in \{0,1\}^n} p_v = 1$. Note that this model can account for any aggregate data vector $\boldsymbol{x}$ (i.e., it will always produce a perfect fit). Also, note that the model is oversaturated in the sense that there are more free parameters $(2^n - 1)$ than degrees of freedom in the data (n). One consequence is that more than one set of parameter values can produce a perfect fit of the data.

Now, let us reparametrize p_v such that we distinguish between preference patterns belonging to V_{out} and the ones belonging to V_{in}. We will do this by introducing a parameter P_{out} describing the probability of preference patterns included in V_{out}. We will also specify conditional probabilities $q_{v_{out}}$ and $q_{v_{in}}$ concerning the different v in V_{out} and V_{in}, respectively. Altogether, we have $p_v = P_{out} \times q_{v_{out}}$ when v belongs to V_{out}, and $p_v = (1 - P_{out}) \times q_{v_{in}}$ when v belongs to V_{in}. Note that $\sum_{v_{out} \in V_{out}} q_{v_{out}} = 1$ and $\sum_{v_{in} \in V_{in}} q_{v_{in}} = 1$.

As mentioned before, a perfect fit of the data can be achieved under many sets of parameter values. The value that we want to determine, $Q(\boldsymbol{x})$, corresponds to the largest P_{out} value with which we can still obtain a perfect fit of the data. One simple way to do this, exemplified in the R code below, consists of fitting the model under different P_{out} and determine the largest value under which the goodness-of-fit statistic G^2 is zero (i.e., perfect fit).

```
##### R Code #####

# preference patterns in V_in
V_in  <- structure(c(0, 0, 0, 0, 0, 1, 1,
1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1,
0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1), .Dim = c(8L, 4L))

# preference patterns in V_out
V_out <- structure(c(0, 0, 0, 1, 1, 1, 1,
0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0,
0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1), .Dim = c(8L, 4L))

# model function used to fit the model.
Vmodel <- function(Q,x,P_out,V_in,V_out){

  # raw parameter values to be estimated
  Q <- Q[1:14]

  # get the q parameters
  q_v_out <- c(Q[1:7] ,1)/sum(c(Q[1:7] ,1))
  q_v_in  <- c(Q[8:14],1)/sum(c(Q[8:14],1))

  # get model's expected probabilities
  e <- (1-P_out)*matrix(q_v_in,8,4 )*V_in +
          P_out*matrix(q_v_out,8,4)*V_out

  # get complementary probabilities for
  # expected probs. and data
  e <- c(rbind(colSums(e),1-colSums(e)))
  x <- c(rbind(x,1-x))

  # compute and return G^2 statistic
  Gsq <- 2*sum(x[x!=0]*(log(x[x!=0])-log(e[x!=0])))
  return(Gsq)
}

# response vector
# x <- c(.50,.50,.50,.50)
x <- c(.80,.95,.20,.90)

# sequence from 1 to 0 of P_out values
P_out_try <- seq(1,0,-0.01)

# fit model using every value of P_out
# starting from 1 until G^2=0 is produced
fit_try   <- c()
for(ii in 1:length(P_out_try)){
  fit <- nlminb(runif(14,0,5),Vmodel,lower=rep(0,14),
         x=x,V_in=V_in, V_out=V_out,P_out=P_out_try[ii])

  fit_try[ii] <- round(fit$objective,5)
  if(fit_try[ii] == 0) break()
}

# Q(x)
P_out_try[which.min(fit_try)[1]]
```