Judgment and Decision Making, Vol. 6, No. 8, December 2011, pp. 857–869

A model-based approach for the analysis of the calibration of probability judgments

David V. Budescu* Timothy R. Johnson†

Abstract

The calibration of probability or confidence judgments concerns the association between the judgments and some estimate of the correct probabilities of events. Researchers rely on estimates using relative frequencies computed by aggregating data over observations. We show that this approach creates conceptual problems, and may result in the confounding of explanatory variables or unstable estimates. To circumvent these problems we propose using probability estimates obtained from statistical models—specifically mixed models for binary data—in the analysis of calibration. We illustrate this methodology by re-analyzing data from a published study and comparing the results from this approach to those based on relative frequencies. The model-based estimates avoid problems with confounding variables and provided more precise estimates, resulting in better inferences.

Keywords: calibration, confidence judgments, mixed models, multilevel models, overconfidence, subjective probability.

1 Introduction

There is a substantial literature about the quality of probability and confidence judgments (see Erev, Wallsten, & Budescu, 1994; Griffin & Brenner, 2004; Harvey, 1997; Kahneman, Slovic, & Tversky, 1982; Keren, 1991; McClelland & Bolger, 1994; Murphy & Winkler, 1992; Wallsten & Budescu, 1983). A specific property of the probability judgments—their *calibration*—has been accepted as the "common standard of validity" in the empirical literature (Wallsten & Budescu, 1983). Judgments are said to be calibrated if $p(100)\%$ of all events that are assigned a *subjective* probability of p materialize.[1] This paper focuses on some conceptual and methodological problems associated with standard calibration analyses. After reviewing some of the problems associated with this approach we propose and illustrate an alternative model-based approach to assess the calibration of probability judgments that overcomes these problems.

Both authors contributed equally and the ordering is alphabetical. This work was supported, in part, by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20059. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U. S. Government.

*Department of Psychology, Fordham University, Dealy Hall, 411 East Fordham Road, Bronx, NY, 10458, USA. Email: budescu@fordham.edu.

†Department of Statistics, University of Idaho.

[1] This analysis is applied to future events whose occurrence is unknown and depends on external sources (e.g., "What is the probability that the market will move by 4% tomorrow?"), and factual events about which the judges may be uncertain because of imperfect or incomplete information (e.g., How confident are you that the population of Chile is larger than Peru's?). We do not distinguish between the two.

1.1 Calibration

We define calibration in terms of individual judgments concerning individual events. Let E_{ij} denote the j-th event for which the i-th judge gives a confidence judgment, C_{ij}. Calibration concerns the relationship between the judgment C_{ij} and the probability of E_{ij}. There are two distinct types of calibration, depending on how the probability of the target event is defined (see Budescu, Erev, & Wallsten 1997; Budescu, Wallsten, & Au, 1997; Murphy & Winkler, 1992; Wallsten, 1996). One type concerns the *conditional* probability $P(E_{ij}|C_{ij} = c)$. A judge is calibrated if $P(E_{ij}|C_{ij} = c) = c$ for all c. This means that the probability of the event is c when the judge assigns to it a confidence judgment of c. There are two ways of defining miscalibration or over/underconfidence (Harvey, 1997). The most prevalent definition pertains to the judge's ability to distinguish between true and false events and it is usually applied in forced choice tasks. According to this view, a judge is overconfident if $P(E_{ij}|C_{ij} = c) < c$ and $c > 0.5$, or $P(E_{ij}|C_{ij} = c) > c$ and $c < 0.5$. A judge is considered to be underconfident if $P(E_{ij}|C_{ij} = c) > c$ and $c > 0.5$, or $P(E_{ij}|C_{ij} = c) < c$ and $c < 0.5$ (see Wallsten, Budescu, & Zwick, 1993). A second definition captures the judge's confidence in the truth of an event. Thus overconfidence and underconfidence are implied by the inequalities $P(E_{ij}|C_{ij} = c) < c$ and $P(E_{ij}|C_{ij} = c) > c$, respectively. The (mis)calibration of judgments is often summarized by a "calibration curve" which plots $P(E_{ij}|C_{ij} = c)$ as a function of c.

A second type of calibration concerns the *marginal* probability $P(E_{ij})$. A judge is calibrated if $P(E_{ij}) = C_{ij}$. If we consider the calibration of judges *on aver-*

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

age, then this could be viewed as a reversal of the conditioning argument in the previous definition, in the sense that one is calibrated if $E(C_{ij}|P(E_{ij}) = p) = p$ so that when the probability of the event is p, the expected judgment is also p. This type of calibration can be visualized by plotting C_{ij} against $P(E_{ij})$ as a type of "reversed" calibration curve where the axes are interchanged. This definition is used often in studies of Bayesian updating where the events' probabilities result from a known process of random generation (see Erev, Wallsten, & Budescu, 1994). A judge is overconfident when $C_{ij} > P(E_{ij})$ and $P(E_{ij}) > 0.5$, or $C_{ij} < P(E_{ij})$ when $P(E_{ij}) < 0.5$. Underconfidence occurs when $C_{ij} < P(E_{ij})$ when $P(E_{ij}) > 0.5$, or $C_{ij} < P(E_{ij})$ when $P(E_{ij}) < 0.5$.

Ideally, one would like to quantify the quality *of the judgment provided by any specific judge for any given event*. However, traditional calibration analysis is usually performed *at the group level and across multiple events* since the conditional probabilities, $P(E_{ij}|C_{ij} = c)$, or the marginal probabilities, $P(E_{ij})$, are typically unknown. To estimate $P(E_{ij})$ one computes the proportion of observations in a subset of observations for which the event occurs. To estimate $P(E_{ij}|C_{ij} = c)$, one computes this only for those observations that were assigned a judgment of c. These probability estimates are then used to assess calibration. For example, calibration curves plot p_c against c where p_c is an estimate of $P(E_{ij}|C_{ij} = c)$, which is usually the proportion of observations where the event occurs when the judgment is c. Calibration can then be measured, for example, using the calibration index

$$CI = \frac{\sum_c n_c(p_c - c)^2}{\sum_c n_c},$$

where n_c is the number of observations aggregated to estimate p_c when the judgment is c. For calibration based on marginal probabilities researchers use simple global measures such as $p - \bar{c}$, $|p - \bar{c}|$ and $(p - \bar{c})^2$ where p is a relative frequency and $\bar{c}$ is the average confidence judgment.

The use of relative frequencies is justified by the implicit assumption that the events being aggregated form an equivalence class so that all events in a given aggregated set have a common (conditional) probability. In cases where the observations are not replications of the same process, or not sampled from the same domain, this assumption may be questionable. For example, it may be misleading to combine weather forecasts of different forecasters operating at different locations, as it would be inappropriate to aggregate financial forecasts made in various countries. The practice of combining judgements of many participants regarding general knowledge items from various domains selected arbitrarily, which has been used in many psychological experiments, has been criticized on similar grounds (Gigerenzer, Hoffrage & Kleinbölting, 1991; Juslin, 1994; Juslin, Winman & Olsson, 2000; Winman, 1997). Next we review some important conceptual and statistical concerns over group-level analysis of calibration based on aggregated data.

1.2 Conceptual concerns

The quality of *subjective/personal* judgments regarding *unique/non-repeatable* events relies on a comparison of these judgments with *relative frequencies aggregated across multiple judges, events and occasions*. Paradoxically, the standard of calibration for *subjective* probabilities is based on a *frequentist* approach to probability. Given the diametrically opposed views held by these two schools of thought, this state of affairs should be equally disturbing to all researchers regardless of their stand on the question of the "proper" interpretation of probability (Keren, 1991; Lad, 1984). Another concern is the insensitivity of calibration analysis to individual differences. Researchers (e.g., Gigerenzer et al., 1991; Juslin, 1994; Winman, 1997) have argued that some empirical results are artifacts due to biased selection of events. A similar argument can be made with respect to the selection of judges. The degree of miscalibration in any study is determined, in part, by the expertise of the participants in the domains of interest. Judges who vary in knowledge or expertise may lead researchers to reach different, possibly conflicting, conclusions.

1.3 Statistical concerns

The analysis of calibration based on aggregated observations also has several *statistical* problems. First, if the subsets of observations used to produce relative frequencies are not sufficiently large, the estimated probabilities will be unstable. This undermines the power and precision of statistical inferences based on these estimates. To avoid this problem researchers aggregate observations. But this leads potentially to a second problem, as this may require one to aggregate data over important characteristics of the judges, events, and/or circumstances under which the judgments were elicited. Confounding variables may distort the apparent relationship between the probabilities and these variables. This problem is well-known in statistics and can lead to such phenomenon as Simpson's Paradox (Simpson, 1951) or the Ecological Fallacy (Robinson, 1950). A standard solution to this problem is to avoid aggregation by conditioning on the relevant variables using a statistical model.

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

2 A model-based approach to estimating probabilities

To simplify notation, let $\pi_{ij}^{(c)} = P(E_{ij}|C_{ij} = c)$ and $\pi_{ij} = P(E_{ij})$, where i (still) indexes the judges, j the events, and c a given confidence level. We propose estimating $\pi_{ij}^{(c)}$ and π_{ij} with regression models. Since the outcome (the event) is binary, a natural family of statistical models is generalized linear mixed models for binary variables (see Pendergast, Gange, Newton, Lindstrom, Palta, & Fisher, 1996; Guo & Zhao, 2000). This includes mixed logistic regression models, and extensions thereof. These models are of the form:

$$\pi_{ij}^{(c)} = f(\boldsymbol{\beta}, \mathbf{b}_i, c_{ij}) \quad \text{and} \quad \pi_{ij} = g(\boldsymbol{\beta}, \mathbf{b}_i),$$

where $\boldsymbol{\beta}$ is a vector of parameters representing the effects of explanatory variables that characterize the events or circumstances of the judgments, $\mathbf{b}_i$ is a vector of random judge-specific parameters to allow for individual differences, and f and g are inverse link functions that map the parameters into probabilities.[2] Note that $\pi_{ij}^{(c)}$ is a function of c_{ij}, since the probability is conditional on the confidence judgment, whereas π_{ij} is a marginal probability. Appendix A gives further details on model notation and specification, and the next section provides specific examples. We propose a three steps process:

1. Specify a mixed model for $\pi_{ij}^{(c)}$ and/or π_{ij} to produce the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}_i$.
2. Estimate the event probabilities as $\hat{\pi}_{ij}^{(c)} = f(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}_i, c_{ij})$ and/or $\hat{\pi}_{ij} = g(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}_i)$.
3. Use the estimated probabilities, $\hat{\pi}_{ij}^{(c)}$ and/or $\hat{\pi}_{ij}$, to assess the calibration of the observed judgments, c_{ij}.

Most models can be estimated using standard statistical packages for generalized linear mixed models. We used PROC GLIMMIX in SAS/STAT, Version 9.2 (SAS, 2008). Appendix B gives syntax examples of how to implement these models with software. The specification of a model is an important issue since this approach relies on having valid estimates of the probabilities. We relied on the Akaike information criterion (AIC; Akaike, 1974) to select models, but additional analyses confirmed that our results were reasonably robust to minor changes in the model specification.[3]

Mixed models have already been shown to be useful in analyses of data from research in judgment and decision making (e.g., Merkle, 2010; Merkle, Smithson, Verkuilen, 2010; Stockard, O'Brien, & Peters, 2006). The methodology proposed here can be viewed as an extension of an approach proposed by Merkle (2010) for using mixed models to study calibration. Such models can be used to study systematic trends in data while also accounting for individual differences. This model-based approach overcomes the sparseness of the data by expressing the unknown probabilities as a function of explanatory variables. Instead of relying on relative frequencies from aggregated observations the model provides estimates for the probability of *each* event. Thus the model avoids problems caused by aggregating data. And by using a parametric model we obtain more precise estimates of the probabilities than those based on relative frequencies. In the following section we demonstrate this approach and contrast it to analyses based on aggregation.

2.1 Illustrative example

McGraw, Mellers, and Ritov (2004) report two studies in which subjects gave confidence judgments prior to throwing a basketball at a basket.[4] In the first study 45 subjects threw a basketball at a hoop three times from each of 12 locations that varied in terms of the distance to the basket and side of the court. In the second study 20 subjects were randomly assigned to a control group, and 22 to a "debiased" group where they were instructed to avoid overconfidence. They attempted five shots from each of seven different distances from the basket along the center of the court. McGraw et al. were concerned with the relationship between judgments of pleasure of the outcomes to confidence and calibration, but we will focus only on calibration. The limitations of using relative frequencies to estimate the unknown probabilities becomes clear when we consider that probabilities may vary over distance, side of the court, group, and subject. Each subject made only a few shots from each spot, which is a too small a number of observations to accurately estimate probabilities using relative frequencies. But aggregating across variables, such as distance, can change the results.

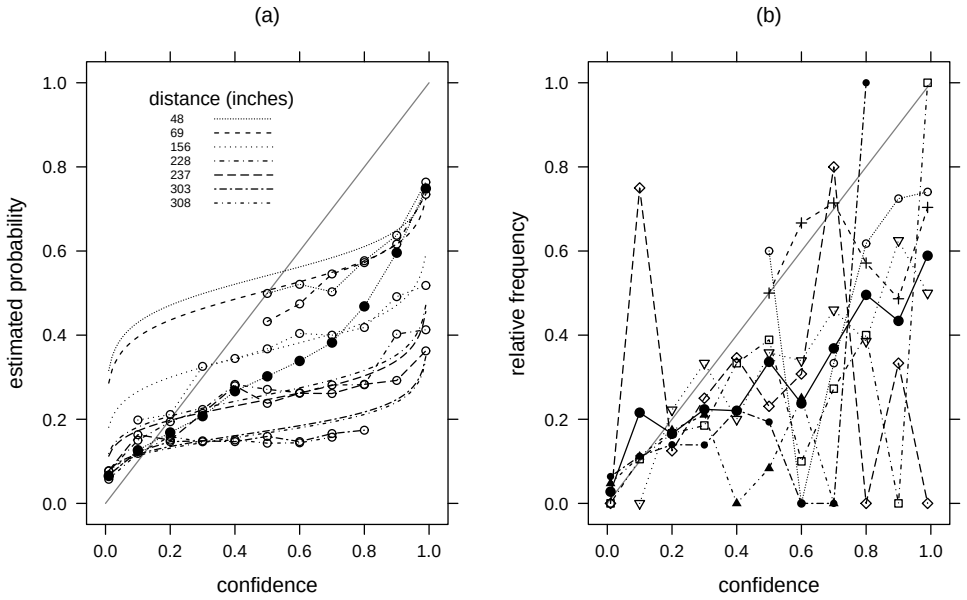
[2] The functions will involve one or more explanatory variables, in addition to c_{ij}, but these have been suppressed here for simplicity.

[3] To evaluate the relative fit of our models we relaxed each model in several ways including introducing higher-order powers of quantitative explanatory variables and interactions among explanatory variables, including interactions with subjects. By generalizing the models far enough in this way we could evaluate the goodness-of-fit fit of our models by (nearly) saturating the models to evaluate goodness-of-fit, although we relied on using AIC rather than goodness-of-fit tests to avoid overfitting. Finally we used graphical methods by plotting the aggregated estimated probabilities against relative frequencies to determine if the model-based estimates agreed with the model-free estimates.

[4] The authors thank Peter McGraw for providing the data.

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

Figure 1: Estimated calibration curves from (a) model-based probability estimates and (b) relative frequencies for the first basketball study. Smooth curves in (a) are mean calibration curves, averaged over subjects. Open points are mean estimated conditional probabilities for each distance and confidence value. Closed points are mean estimated probabilities for each confidence value, averaged over distance.



2.1.1 Analysis based on conditional probabilities

First we consider the conditional probabilities, $\pi_{ij}^{(c)}$. For the first study we used the model

$$\text{logit}(\pi_{ij}^{(c)}) = \beta_0 + \beta_1 \text{DISTANCE}_{ij} + \beta_2 \text{logit}(c_{ij}) + b_{i0}, \quad (1)$$

where $\text{logit}(z) = \ln[z/(1-z)]$, DISTANCE_{ij} is the distance (in inches) to the basket, and c_{ij} is the confidence judgment. It is convenient to transform the probabilities and judgments to log-odds since the model is linear on the log-odds scale, and furthermore when $\beta_0 = \beta_1 = 0$, $\beta_2 = 1$, and $b_{i0} = 0$ then $\text{logit}(\pi_{ij}^{(c)}) = \text{logit}(c_{ij})$ and thus $\pi_{ij}^{(c)} = c_{ij}$, implying perfect calibration.[5] Thus the parameters capture *miscalibration* due to different sources. The discrepancy between the confidence judgment and the conditional probability are represented by β_0 and β_2, the effect of distance by β_1, and b_{i0} is a subject-specific effect. For the second study we specified a similar model but added an effect for the experimental manipulation so that

$$\text{logit}(\pi_{ij}^{(c)}) = \beta_0 + \beta_1 \text{DEBIAS}_i + \beta_2 \text{DISTANCE}_{ij} + \beta_3 \text{logit}(c_{ij}) + b_{i0} \quad (2)$$

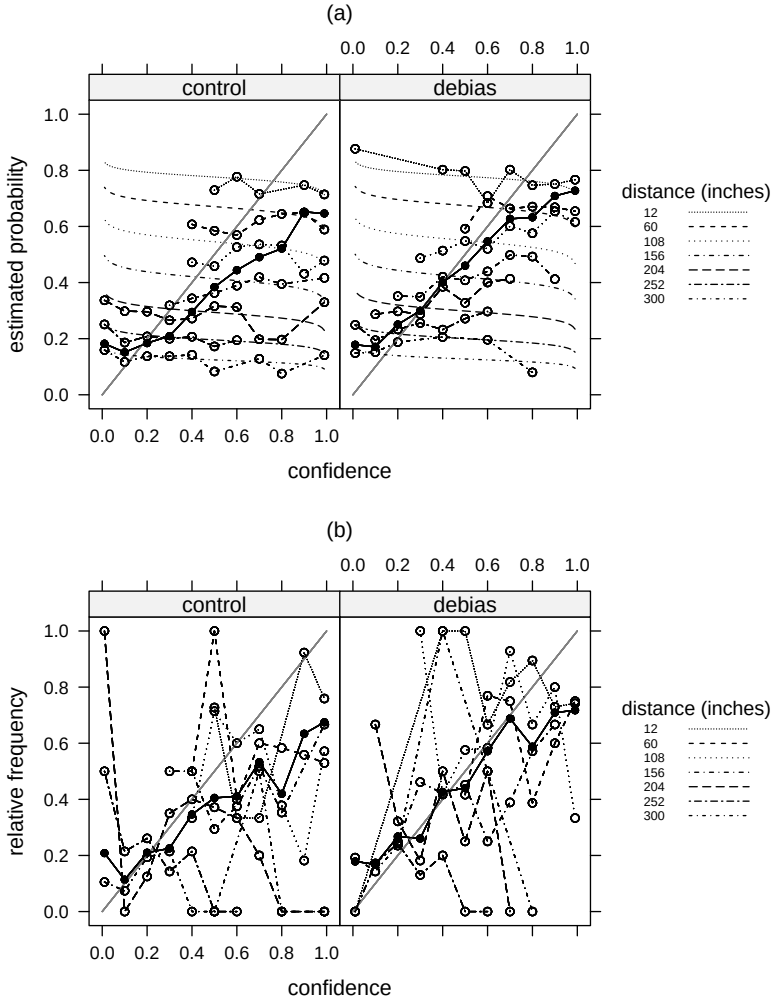
[5] We set $c_{ij} = 0.01$ or $c_{ij} = 0.99$ if c_{ij} was 0 or 1, respectively.

where DEBIAS_i is a binary variable that indicates if the i-th subject was in the debiased group.

Figure 1a plots the estimated calibration curves based on the model in Equation (1). The smooth curves are the mean calibration curves. The open points are mean values of $\pi_{ij}^{(c)}$ grouped by distance and confidence judgment. The model confirmed a significant effect for distance ($\hat{\beta}_1 = -0.007$, $z = -5.61$, $p < 0.001$) which can be seen clearly in the figure: as distance increases, the judges tend to be more overconfident. Figure 2a shows the estimated calibration curves for each distance and group for the second study based on Equation (2). The plot is constructed like Figure 1 but the data are conditioned also on group. Again distance had a significant effect ($\hat{\beta}_2 = -0.01$, $z = -9.24$, $p < 0.001$). However neither the manipulation ($\hat{\beta}_1 = 0.12$, $z = 0.61$, $p = 0.54$) nor the effect of the confidence judgment ($\hat{\beta}_3 = -0.07$, $z = -1.24$, $p = 0.21$) were significant. The lack of apparent effect for the confidence judgment might seem surprising, since it implies flat calibration curves. However it is reasonable that after accounting for distance and the subject, that the judgments themselves would not predict the outcome. In both Figures 1 and 2 the solid points are the mean values of the estimates of $\pi_{ij}^{(c)}$, grouped by confidence judgment and aggregated over distance. This average curve is notably steeper than the estimated calibration curves for each distance. Again the pattern of

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

Figure 2: Estimated calibration curves from (a) model-based probability estimates and (b) relative frequencies for the second basketball study. Smooth curves in (a) are mean calibration curves, averaged over subjects. Open points are mean estimated conditional probabilities for each distance, group, and confidence value. Closed points are mean estimated probabilities for each group and confidence value, averaged over distance.



over- or underconfidence is highly dependent on whether or not one controls for distance.

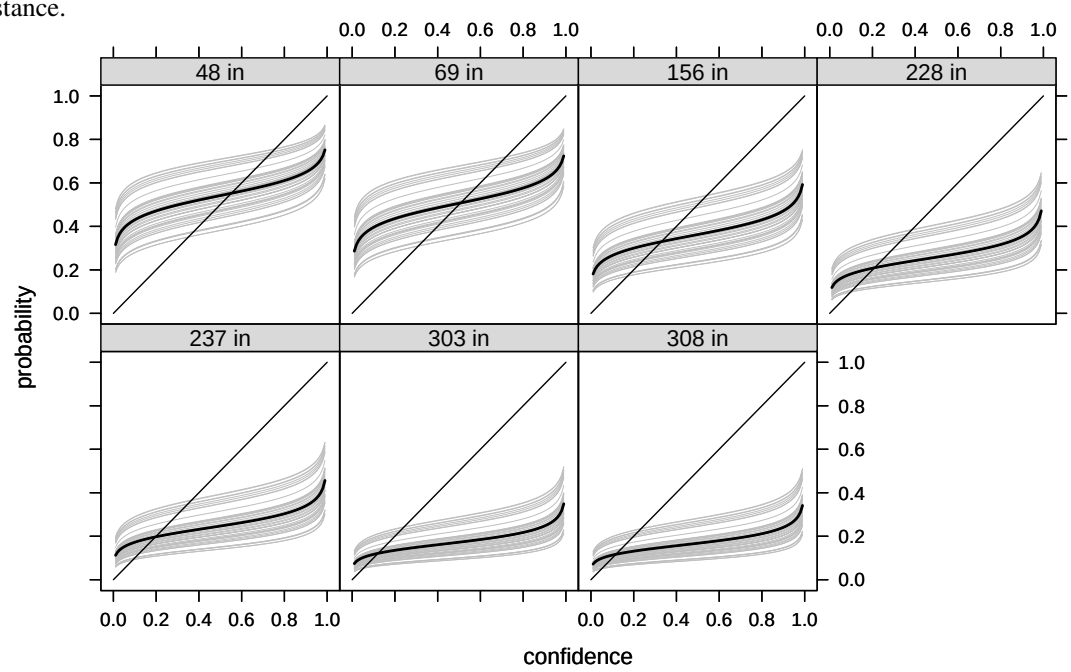
Figure 1b and 2b are similar to Figures 1a and 2a, respectively, but based on relative frequencies. The open points are relative frequencies from aggregating observations for each distance and confidence judgment. The relative frequencies aggregated *within* distances are highly unstable. The closed points are relative frequencies based on aggregating observations *across* distances.[6] These are more stable, but averaging over distance ignores the effect of distance on confidence. The model-based approach provides more stable estimates of the probabilities without ignoring or obscuring the effects of variables.

[6]They correspond to Figures 3 and 6, respectively, in McGraw et al. (2004).

One significant benefit of using a mixed model is that it can also account for individual differences in calibration as well as systematic effects due to explanatory variables such as distance and treatment group. In Equations (1) and (2) these individual differences are modeled through the subject-specific effect represented by b_{i0}. The effect of this parameter can be seen graphically by plotting the estimated calibration curve for *each* subject, for each given distance, as shown in Figure 3. As can be seen in the plot, there is considerable variation across subjects in the calibration curves. The variance of b_{i0} was estimated at approximately 0.25, with a standard error of 0.09. It is useful to note here that one could also permit variability across subjects in the *slope* of the calibration curve, on the log-odds scale, by adding the term $\beta_{i1}\text{logit}(c_{ij})$ to

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

Figure 3: Estimated subject-specific calibration curves for the first basketball study. The light grey curves represent the estimated curves for each of the 45 subjects, at each distance. The black curves are the mean calibration curve at each distance.



(1) or (2), although this did not improve the fit of either model here.

The lack of a significant effect for the experimental manipulation might appear to contradict the analysis by McGraw et al., but their analysis ignores the effect of distance, whereas our analysis controls for it. But the manipulation may also have influenced the judgments.[7] To assess the effect of the manipulation on calibration, we examined its effect on the *joint* distribution of the judgments and the estimated probabilities—specifically the distribution of the discrepancy between them. We computed $(\hat{\pi}_{ij}^{(c)} - c_{ij})^2$ and used it as the response variable in a mixed effects linear model. The main effect for group was significant ($F_{(1,1416)} = 9.07$, $p = 0.003$), showing better calibration indices for the debiased group. The effect for distance was also significant ($F_{(6,1416)} = 5.17$, $p < 0.001$), but the interaction between group and distance was not ($F_{(6,1416)} = 0.82$, $p = 0.55$). Figure 4 shows the means and distributions of the logs of the discrepancy measure as a function of distance by group. A similar analysis based on the relative frequencies did not show a significant effect for group ($F_{(1,615)} = 2.42$, $p = 0.12$), most likely due to the greater variability of the indices computed from relative frequencies based on few observations.

[7] A reviewer noted that another limitation of aggregating data is that the aggregation may also be over *judgments*, which would preclude analyses sensitive to variables influencing the distribution of judgments.

2.1.2 Analysis based on marginal probabilities

Next we estimated the marginal probabilities, π_{ij}. For the first study we specified the model

$$\begin{aligned}\text{logit}(\pi_{ij}) = \quad & \beta_0 + b_{i0} + (\beta_1 + b_{i1})\text{DISTANCE}_{ij} + \\ & \beta_2\text{LEFT}_{ij} + \beta_3\text{RIGHT}_{ij}. \end{aligned} \quad (3)$$

We include the effects of side of the basket with indicator variables, and an additional judge-specific effect for distance, b_{i1}. These effects were not used in the model for the conditional probabilities because they did not improve the fit of the models. To estimate the marginal probabilities in the second study we used the model

$$\begin{aligned}\text{logit}(\pi_{ij}) = \quad & \beta_0 + \beta_1\text{DEBIAS}_i + \\ & \beta_2\text{DISTANCE}_{ij} + b_{i0}. \end{aligned} \quad (4)$$

These models provide for an interesting new analysis by "reversing" the traditional calibration curve to examine the conditional distribution of the confidence judgments given the (estimated) probabilities. This is possible only because the models provide estimates of the probability for *each* observation. This is not possible, or at least cannot be done as finely, by aggregating observations for relative frequencies. Figures 5 and 6 depict the mean confidence judgments and their confidence intervals, conditional on the estimated probabilities, for the first and second study.

We grouped the confidence judgments by the corresponding marginal probability estimate, rounded to the

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

Figure 4: Mean and distribution of calibration indices (log scale) from model-based probability estimates for each group and distance for the second basketball study.

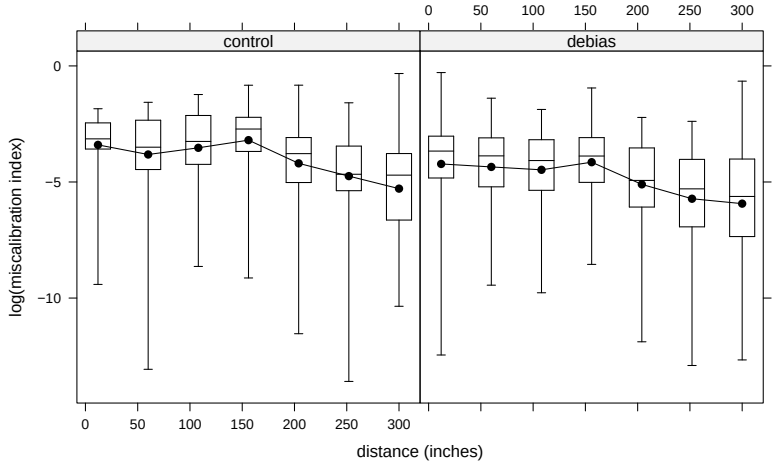
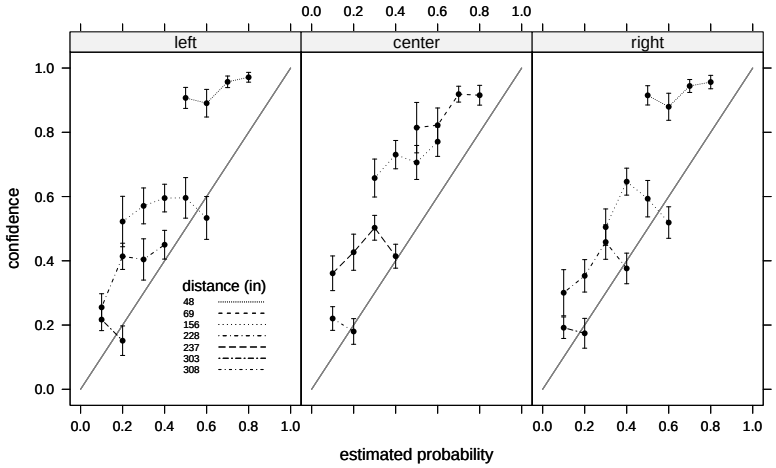


Figure 5: Mean confidence judgments for each side, distance, and estimated marginal probability for the first basketball study. The marginal probabilities have been rounded to the nearest tenth. The error bars represent 95% confidence intervals.



nearest tenth. The plots appear to indicate some tendency to overestimate the probabilities, particularly at shorter distances, when shots were attempted at the center of the court in the first study, and more so in the control group than in the debiased group in the second study.

To further examine apparent trends in the miscalibration of the judgments based on the marginal probabilities, we analyzed the discrepancy measure $c_{ij} - \hat{\pi}_{ij}$. Figures 7a and 8a show the means and distributions of the miscalibration measures by distance and location for the two studies. These plots also show the trends we observed in Figures 5 and 6. Statistical analyses confirmed the trends. In the first study there was a significant interaction between side and distance ($F_{(6,1564)} = 6.17$, $p < 0.001$). The tendency to overestimate the probability decreased with distance when shooting from the side, but the trend is curvilinear when shooting from the center. For the second study we found significant main effects for both distance ($F_{(6,1416)} = 4.01$, $p = 0.005$) and group ($F_{(1,1416)} = 8.97$, $p = 0.003$), but not their interaction ($F_{(6,1416)} = 1.88$, $p = 0.08$). When controlling for distance, the analysis reveals a significant effect for the debiasing manipulation. It significantly improved calibration overall.

To compare these analyses with a model-free approach based on the raw data, we analyzed the discrepancy

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

Figure 6: Mean confidence judgments for each group, distance, and estimated marginal probability for the second basketball study. The marginal probabilities have been rounded to the nearest tenth. The error bars represent 95% confidence intervals.

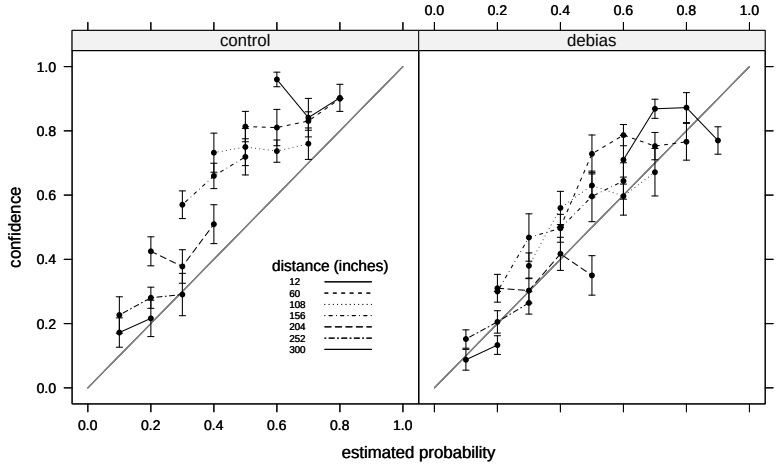
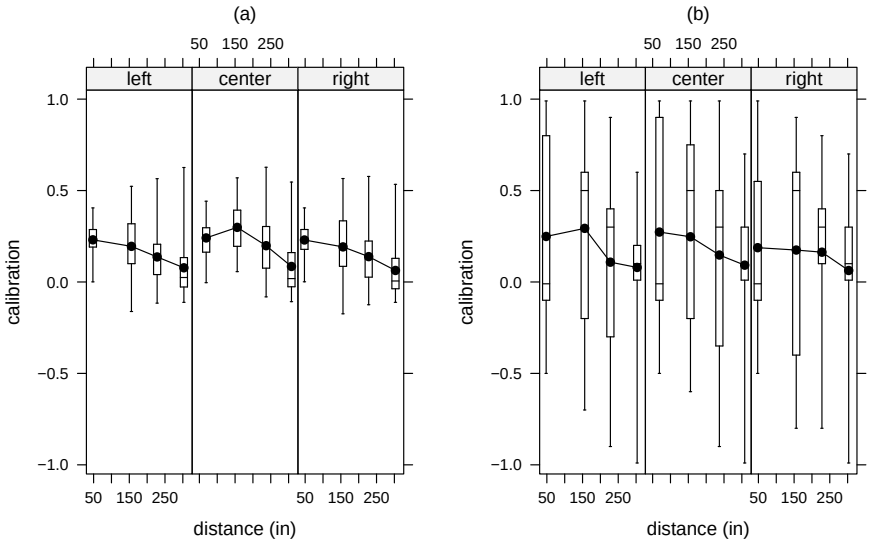


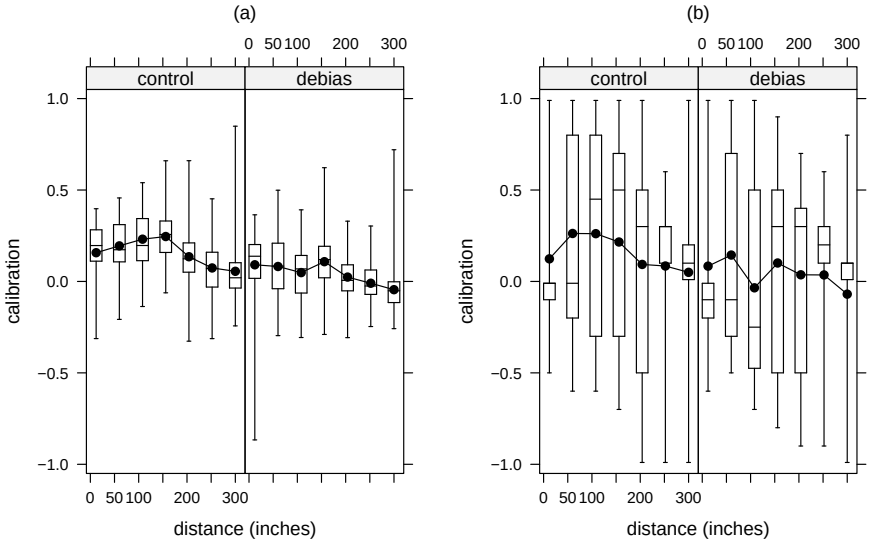
Figure 7: Mean and distribution of calibration measures from (a) the model-based probability estimates and (b) the raw data, for each side and distance, for the first basketball study.



measure $c_{ij} - I(E_{ij})$, where $I(E_{ij})$ indicates whether event E_{ij} (a basket) occurred. We used $c_{ij} - I(E_{ij})$ rather than aggregating observations to estimate π_{ij} since $E[c_{ij} - I(E_{ij})] = E(c_{ij}) - \pi_{ij}$. Figures 7b and 8b show the means and distributions for this measure for the first and second study, respectively. Note the greater variability of this measure. This instability hinders statistical analyses. We failed to detect a significant interaction ($F_{(6,1564)} = 1.23$, $p = 0.29$), main effect for distance ($F_{(3,1565)} = 1.85$, $p = 0.14$), or main effect for side ($F_{(2,1564)} = 1.06$, $p = 0.35$). For the second study we confirmed significant main effects for the manipulation ($F_{(1,1416)} = 8.97$, $p = 0.003$), distance ($F_{(6,1416)} = 4.01$, $p = 0.001$), but not the interaction ($F_{(6,1416)} = 1.88$, $p = 0.081$). While both analyses estimated the same mean difference of the calibration measure between the control and debiased groups (0.11), the standard error was approximately 30% larger than in the analysis using the model-based probability estimates. While it is possible to analyze calibration based on

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

Figure 8: Mean and distribution of calibration measures from (a) the model-based probability estimates and (b) the raw data, for each group and distance, for the second basketball study.



the marginal probabilities without aggregation, a model-based approach can provide more stable estimates and thus more precise inferences.

3 Discussion

We have argued and demonstrated with examples that the standard calibration analysis has significant limitations. Conceptually it is inconsistent because it makes relative frequencies the standard of evaluation of the judges' subjective probabilities. Statistically it is problematic because it can lead to biased estimates when aggregation is over important variables, and imprecise estimates when aggregation is over too few observations.

We proposed and demonstrated the use of mixed models for binary data to estimate the probabilities of specific events for the purpose of analyzing the calibration of specific judges. With a good model, one can estimate the probabilities of individual events accurately, without resorting to indiscriminate data aggregation. Instead of comparing the judgments to a set of relative frequencies our approach uses probabilities derived from a model that captures empirical regularities, and incorporates relevant individual differences. Thus the standard of comparison for any given event is personal, as it is explicitly tailored to each judge. This fact addresses, at least in part, the conceptual concern about using a frequentist analysis to estimate the quality of subjective judgments. We demonstrated that the model-based approach provides superior results in that it can address issues of confounding explanatory variables while providing more precise probability estimates which translate into higher precision and power in statistical inferences, and allows new informative analyses (e.g., "reverse calibration" curves) and at different levels (individual events and judges), which are not possible in the traditional approach.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-39. http://CRAN.R-project.org/package=lme4

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997). On the importance of random error in the study of probability judgments. Part I: New theoretical developments. *Journal of Behavioral Decision Making*, *10*, 157–172.

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997). On the importance of random error in the study of probability judgments. Part II: Applying the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making*, *10*, 173–188.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-confidence and conservatism in judg-

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

ment: Implications for research and practice. *Psychological Review*, *101*, 519–527.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*,506–528.

Goldstein, H. (2010). *Multilevel statistical models* (4th Edition). West Sussex, UK: Wiley.

Griffin, D. & Brenner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koehler and N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making*, pp. 177–199.

Guo, G. & Zhao, H. (2000). Multilevel modeling for binary data. *Annual Review of Sociology*, *26*, 411–462.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences*, *1*, 78–82.

Juslin, P. (1994). The over-confidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, *57*, 226–246.

Juslin, P., Winman, A., & Olsson, H. (2000). Naïve empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *107*, 384–396.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgment under uncertainty: Heuristics and biases,* Cambridge, England: Cambridge University Press.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, *77*, 217–273.

McClelland, A. G. R. & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980-1994, In G. Wright & P. Ayton (Eds.) *Subjective probability*, (pp. 453–484). Chichester: Wiley.

McGraw, A. P., Mellers, B. A., & Ritov, I. (2004). The affective cost of overconfidence. *Journal of Behavioral Decision Making*, *17*, 281–295.

Merkle, E. C. (2010). Calibrating subjective probabilities using hierarchical Bayesian models. In Chai, S.-K., Salerno, J. J., & Mabry, P. L. (Eds.), *Social Computing, Behavioral Modeling, and Prediction (SBP) 2010* (pp. 13-22). Lecture Notes in Computer Science 6007.

Merkle, E. C., Smithson, M., & Verkulien, J. (2010). Hierarchical models of simple mechanisms underlying confidence in decision making. *Journal of Mathematical Psychology*, *55*, 57–67.

Murphy, A. H., & Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, *7*, 435–455.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., & Fisher, M. R. (1996). A survey of methods for analyzing clustered binary data. *International Statistical Review*, *64*, 89–118.

R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Raudenbush, S. W. & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods* (2nd Edition). Thousand Oaks, CA: Sage.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357.

SAS Institute Inc. (2008). SAS/STAT 9.1 users's guide: The GLIMMIX procedure. Cary, NC: SAS Publishing.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, *13*, 238–241.

Stockard, J., O'Brien, R. M., & Peters, E. (2007). The use of mixed models in a modified Iowa Gambling Task and a prisoner's dilemma game. *Judgment and Decision Making*, *2*, 9–22.

Thomas, A, O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R News*, *6*, 12–17.

Wallsten, T. S. (1996). Commentary: An analysis of judgment research analyses. *Organizational Behavior and Human Decision Processes*, *65*, 220–226.

Wallsten, T. S. & Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, *29*, 151–173.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probabilistic judgments. *Management Science*, *39,* 176–190.

Winman, A. (1997). The importance of item selection in "knew it all along" studies of general knowledge. *Scandinavian Journal of Psychology*, *38*, 63–72.

Appendix A: Model Parameterization

We introduced a model-based approach to estimating the conditional and marginal probabilities by using the general models $\pi_{ij}^{(c)} = f(\boldsymbol{\beta}, \mathbf{b}_i, c_{ij})$ and $\pi_{ij} = g(\boldsymbol{\beta}, \mathbf{b}_i)$, respectively. In this appendix we discuss in more detail how these models might be specified. A common parameterization is the generalized linear mixed model

$$h(\pi_{ij}^{(c)}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i \quad \text{or} \quad h(\pi_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i,$$

where h is a link function, such as the log-odds or "logit" as we used, $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are vectors of observed design variables and covariates corresponding to the fixed and random effects, respectively, $\boldsymbol{\beta}$ is a vector of unknown parameters, and $\mathbf{b}_i$ is a random vector of unknown subject-specific parameters. In the first model the probability $\pi_{ij}^{(c)}$ is conditional on the confidence judgment, so $\mathbf{x}_{ij}$ and possibly $\mathbf{z}_{ij}$ will contain c_{ij}, or some function of thereof.

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

We showed that it is convenient to use the logit link function so that this function is $h(c_{ij})$ and thus the parameters $\boldsymbol{\beta}$ and $\mathbf{b}_i$ capture *miscalibration*. The random subject-specific parameters are assumed to have a specified distribution, such as a multivariate normal distribution such that all $\mathbf{b}_i$ are independently and identically distributed as $N(\mathbf{0}, \boldsymbol{\Sigma})$, which is what we assumed in our analyses.

An alternative parameterization is to write the model as a multilevel or hierarchical generalized linear model as described in, for example, Goldstein (2010) and Raudenbush and Bryk (2001), respectively. Here we can write the 2-level multilevel model in two stages where the level-1 within-subjects model is

$$h(\pi_{ij}^{(c)}) = \mathbf{v}_{ij}'\boldsymbol{\delta}_i \quad \text{or} \quad h(\pi_{ij}) = \mathbf{v}_{ij}'\boldsymbol{\delta}_i,$$

and the level-2 between-subjects model is $\boldsymbol{\delta}_i = \boldsymbol{\Gamma}\mathbf{w}_i + \mathbf{u}_i$, where $\mathbf{v}_{ij}$ and $\mathbf{w}_i$ are vectors of observed design variables or covariates that vary within subjects or between subjects, respectively.[8] For estimating $\pi_{ij}^{(c)}$ the vector $\mathbf{v}_{ij}$ would contain $h(c_{ij})$. This model can be written in the mixed model parameterization given earlier by substituting the level-2 model for δ_i into the level-1 model, although note that in some cases some of the elements of $\boldsymbol{\delta}_i$ will be fixed so that $\mathbf{u}_i$ has a degenerate distribution in which case $\mathbf{b}_i$ is a sub-vector of $\mathbf{u}_i$. To give a concrete example of both parameterizations, Equation (2) can be written as a generalized linear mixed model with $\mathbf{x}_{ij}' = (1, \text{DEBIAS}_i, \text{DISTANCE}_{ij}, \text{logit}(c_{ij}))$, $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2, \beta_3)$, $\mathbf{z}_{ij} = 1$, and $\mathbf{b}_i = b_{i0}$ in the generalized linear mixed model parameterization, and as $\mathbf{v}_{ij}' = (1, \text{DISTANCE}_{ij}, \text{logit}(c_{ij}))$, $\mathbf{w}_i' = (1, \text{DEBIAS}_i)$,

$$\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{00} & \gamma_{01} \\ \gamma_{10} & 0 \\ \gamma_{20} & 0 \end{pmatrix},$$

and $\mathbf{u}_i = (u_{i0}, 0, 0)'$ using the multilevel model parameterization. Note that the multilevel model is then

$$\begin{aligned} \pi_{ij}^{(c)} = \gamma_{00} + \gamma_{01}\text{DEBIAS}_i + \quad & \gamma_{10}\text{DISTANCE}_{ij} + \\ & \gamma_{20}\text{logit}(c_{ij}) + u_{i0} \end{aligned}$$

which is equivalent to Equation (2) except for the change in notation. The choice of parameterization is largely a matter of preference and the software used.

There is a fairly large literature on inference based on generalized linear mixed or multilevel models. Our approach was to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ using maximum likelihood, approximating the integral in the likelihood using adaptive quadrature. Estimates of $\mathbf{b}_i$ can be obtained using an empirical Bayes approach. The estimates of $\hat{\pi}_{ij}^{(c)}$ and $\hat{\pi}_{ij}$ are then obtained by replacing $\boldsymbol{\beta}$ and $\mathbf{b}_i$ by their estimates in the model. Another potentially useful approach would be to specify a Bayesian probability model by specifying a prior distribution for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ to make inferences concerning the posterior distribution of $\pi_{ij}^{(c)}$ or π_{ij} using simulation-based methods.

The methodological approach described in this paper is quite general. One could potentially specify a useful model beyond the families of generalized linear mixed or multilevel models described here. For example, one might find it useful to consider models that are not linear on the scale of $h(\pi_{ij}^{(c)})$ or $h(\pi_{ij})$, or alternative distributions for the the subject-specific parameters such as a mixture distribution. All that is necessary is to have a viable statistical model that provides good estimates of the conditional or marginal probabilities.

[8] We should note that in the multilevel modeling literature it is traditional to reverse the indices so that i refers to the level-1 unit (trial) and j refers to the level-2 unit (subject), but we have kept the use of indices here consistent with our earlier notation.

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

Appendix B: Software Implementation

This appendix gives the syntax for PROC GLIMMIX in SAS/STAT, Version 9.2 (SAS, 2008) for estimating the conditional and marginal probabilities (i.e., $\pi_{ij}^{(c)}$ and π_{ij}, respectively) for the two studies in McGraw et al. (2004). These models can also be implemented using the `glmer()` function in the lme4 package (Bates, Maechler, & Bolker, 2011) for R (R Development Core Team, 2011). We have included the corresponding syntax for `glmer()` for each model as well. The data are assumed to be in "long-form" where each observation/row in the data file corresponds to a one trial for a given subject. The response variable `result` is a binary indicator variable for a successful basket. The explanatory variables `distance`, `logitc`, and `debias` correspond to DISTANCE$_{ij}$, logit(c_{ij}), and DEBIAS$_i$, respectively, in Equations 1-4. The variable `side` indicates the side of the basket (left, center, or right) and generates the indicator variables LEFT$_{ij}$, CENTER$_{ij}$, and RIGHT$_{ij}$ as shown in Equation (3). Subjects are identified by `id`. The output variables `probc` and `probm` are $\hat{\pi}_{ij}^{(c)}$ and $\hat{\pi}_{ij}$, respectively.

Model 1: Estimating $\pi_{ij}^{(c)}$ for Study 1

PROC GLIMMIX Syntax

```
proc glimmix method = quad(qmin = 21);
model result = distance logitc /
  solution chisq
  link = logit
  dist = binomial;
random int / subject = id solution;
output out = study1 pred(ilink) = probc;
```

`glmer()` Syntax

```
model1 <- glmer(result ~ distance + logitc
  + (1 | id), data = study1, family = binomial)
probc <- fitted(model1)
```

Model 2: Estimating $\pi_{ij}^{(c)}$ for Study 2

PROC GLIMMIX Syntax

```
proc glimmix method = quad(qmin = 21);
class debias;
model result = debias distance logitc /
  solution chisq
  link = logit
  dist = binomial;
random int / subject = id solution;
output out = study2 pred(ilink) = probc;
```

`glmer()` Syntax

```
model2 <- glmer(result ~ debias + distance + logitc
  + (1 | id), data = study2, family = binomial)
probc <- fitted(model2)
```

Model 3: Estimating π_{ij} for Study 1

PROC GLIMMIX Syntax

```
proc glimmix method = quad(qmin = 21);
```

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press

```
class side;
model result = distance side /
  solution chisq
  link = logit
  dist = binomial;
random int distance / subject = id type = chol solution;
output out = study1 pred(ilink) = probm;
```

`glmer()` Syntax

```
model3 <- glmer(result ~ distance + side
  + (distance | id), data = study1, family = binomial)
probm <- fitted(model3)
```

Model 4: Estimating π_{ij} for Study 2

PROC GLIMMIX Syntax

```
proc glimmix method = quad(qmin = 21);
class debias;
model result = debias distance /
  solution chisq
  link = logit
  dist = binomial;
random int / subject = id solution;
output out = study2 pred(ilink) = probm;
```

`glmer()` Syntax

```
model4 <- glmer(result ~ debias + distance
  + (1 | id), data = study2, family = binomial)
probm <- fitted(model4)
```

For making inferences concerning the posterior distribution of $\pi_{ij}^{(c)}$ or π_{ij}, or functions thereof, we would suggest using OpenBUGS (Thomas, O'Hara, Ligges, & Sturtz, 2006). For an example we give below one possible way to specify the probability model corresponding to Equation (1).

```
model {
  for (i in 1:45) {
    for (j in 1:12) {
      logit(p[i,j]) = beta0 + beta1 * distance[i,j]
        + beta2 * logitc[i,j] + b0[i]
      y[i,j] ~ dbern(p[i,j])
    }
    b0[i] ~ dnorm(0, tau)
  }

  beta0 ~ dnorm(0, 0.001)
  beta1 ~ dnorm(0, 0.001)
  beta2 ~ dnorm(0, 0.001)

  tau ~ dgamma(0.001, 0.001)
}
```

https://doi.org/10.1017/S1930297500004277 Published online by Cambridge University Press