



# AN INTRODUCTION TO THE GENERAL MONOTONE MODEL WITH APPLICATION TO TWO PROBLEMATIC DATA SETS

*Michael R. Dougherty\**

*Rick P. Thomas<sup>†</sup>*

*Ryan P. Brown<sup>‡</sup>*

*Jeffrey S. Chrabaszcz\**

*Joe W. Tidwell\**

## Abstract

*We argue that the mismatch between data and analytical methods, along with common practices for dealing with “messy” data, can lead to inaccurate conclusions. Specifically, using previously published data on racial bias and culture of honor, we show that manifest effects, and therefore theoretical conclusions, are highly dependent on how researchers decide to handle extreme scores and nonlinearities when data are analyzed with traditional approaches. Within LS approaches, statistical effects appeared or disappeared on the basis of the inclusion or exclusion of as little as 1.5% (3 of 198) of the data, and highly predictive variables were masked by nonlinearities. We then demonstrate a new statistical modeling technique called the*

---

\*University of Maryland, College Park, MD, USA

<sup>†</sup>Georgia Institute of Technology, Atlanta, GA, USA

<sup>‡</sup>University of Oklahoma, Norman, OK, USA

## Corresponding Author:

Michael R. Dougherty, University of Maryland, Department of Psychology, College Park, MD 20742, USA.

Email: [mdougher@umd.edu](mailto:mdougher@umd.edu)

*general monotone model (GeMM) and show that it has a number of desirable properties that may make it more appropriate for modeling messy data: It is more robust to extreme scores, less affected by outlier analyses, and more robust to violations of linearity on both the response and predictor variables compared with a variety of well-established statistical algorithms and frequently possesses greater statistical power. We argue that using procedures that make fewer assumptions about the data, such as GeMM, can lessen the need for researchers to use data-editing strategies (e.g., to apply transformations or to engage outlier analyses) on their data to satisfy often unrealistic statistical assumptions, leading to more consistent and accurate conclusions about data than traditional approaches of data analysis.*

### **Keywords**

*data editing, monotone regression, maximum rank correlation estimator, culture of honor, racial bias*

## **1. INTRODUCTION**

Although recent high-profile cases of fraud have brought unwelcome attention to social sciences, these cases offer an opportunity to reflect on the state of our sciences as well as currently accepted practices (Crocker 2011; Fang, Steen, and Casadevall 2012). To be sure, sociologists have been somewhat ahead of the curve in addressing issues related to data quality, reproducibility (Freese 2007; Hauser 1987), replicability (King 1995), and publication bias (Gerber and Malhotra 2008; Leahey 2005). Of these, data quality arguably ranks as the foremost problem for social scientists because so much, including reproducibility and replication, depends on having good-quality data. Unfortunately, much of the data within the social sciences are messy, and they often require a good amount of editing (e.g., transformation, replacement of missing values, outlier removal) prior to analysis when used with traditional metric statistics. Data editing, however, enables the researcher to capitalize on chance, a problem that is compounded by the fact that there are not well-accepted (or followed) guidelines for how and when to use particular data-editing strategies (Leahey 2008; Leahey, Entwisle, and Einaudi 2003; Sana and Weinreb 2008). The plethora of available strategies, even for something as simple as outlier analysis, can promote flexibility in data analysis. Unfortunately, different approaches to data editing can yield different substantive conclusions, meaning that replications depend not only on the data but also on the specific choices one makes in data editing.

The use of data-editing strategies is just one end of the spectrum of the flexibility afforded to researchers. Modern computers and an ever expanding toolbox of available statistical algorithms permit researchers to easily explore their data in a variety of different ways under different modeling assumptions prior to settling on the subset of analyses that are to be reported (Ho et al. 2007). Coupled with methodological issues surrounding the use of data editing, flexibility in analysis techniques has been a major concern within the social sciences, leading some to call for open-source documentation of data analysis techniques (Freese 2007; Simonsohn 2013). Although there are many reasons to demand open-source documentation, it does not address the problem of flexibility of analysis; it only makes the use of flexible analysis methods public and open to scrutiny.

The work presented here has two related goals. The first is to illustrate the problem with implementing accepted practices on how to deal with messy data, showing just how sensitive substantive conclusions can be to different choices made in data analysis. The second goal is to provide an alternative approach to modeling messy data that reduces or eliminates the need for researchers to make such decisions. With regard to the first goal, using data on racial prejudice (Siegel, Dougherty, and Huber 2012) and culture of honor (Henry 2009), we show that the use of least squares (LS) regression techniques yields inconsistent conclusions across various accepted methods for dealing with messy data. These inconsistencies call into question the validity of statistical conclusions based on LS approaches and in general render the data less interpretable. We argue that the mismatch between the nature of one's data and standard statistical approaches can deceive researchers into drawing invalid conclusions, no matter how well intentioned or diligent the researchers are.

Turning to the second goal, we introduce a new statistical algorithm, the general monotone model (GeMM; Dougherty and Thomas 2012) that makes weaker assumptions than LS approaches about scale of measurement and the functional relationships among manifest variables. GeMM provides relatively more consistent statistical outcomes across several criteria for inclusion or exclusion of extreme scores and the presence of nonlinearities. We show that GeMM is more robust to extreme scores, it is unaffected by nonlinear monotone relationships, and it has superior predictive accuracy and better statistical power when compared with a variety of procedures based on LS. Our application of GeMM in

this article goes beyond previous published applications. Specifically, our analyses evaluate the stability or robustness of GeMM relative to alternative modeling techniques under a variety of realistic conditions that might otherwise entice researchers to make tough decisions about how to handle nonlinear or nonnormal data or the presence of extreme scores. We argue that GeMM provides a promising solution to flexibility in data analysis by greatly reducing both the need for and the impact of data editing.

## 2. MESSY DATA AND TOUGH DECISIONS

Rarely do data neatly conform to the assumptions required for carrying out standard statistical procedures. For instance, it is well recognized that real data typically deviate, often nontrivially, from normality (Micceri 1989), which can result in violations of assumptions underlying standard statistical techniques. Real data are messy. As researchers, we are taught to be vigilant to aberrations in our data, and even to remove them through the use of transformations or “outlier” analyses. For example, Hays (1994) stated that

the data should be inspected for unusually skewed or artificially restricted distributions, missing data, and the presence of unusually deviant cases or outliers. . . . Fortunately, even messy data can often be cleaned up enough to be used, but doing so requires many choices. (p. 721)

Many textbooks contain similar advice—advice that instructs researchers to clean their data through transformation and outlier deletion techniques. These techniques, which we refer to collectively as *data-editing strategies* (Leahey 2008; Leahey et al. 2003), allow researchers to clean and/or reexpress the data in a form that more closely conforms to the assumptions of the statistical model. However, the same textbooks that offer advice on how to handle nonnormalities and outliers also point out that standard LS estimation procedures and their robust implementations often perform reasonably well even when their assumptions are not met (see Howell 2002). This type of back-and-forth between prescribing data-editing strategies and touting robustness is typical.

The fact that many analysis techniques make strong assumptions about distributional (e.g., multivariate normality) and functional (e.g., linear) forms can present researchers with a potentially important

dilemma: Should they engage in data editing to bring the data in line with the assumptions of the analytical procedure, recognizing that the statistical conclusions are conditional on the particular data-editing strategies used? Or should they analyze the data “as is,” recognizing that the statistical conclusions are conditional on potential violations of assumptions? Obviously, the best-case scenario is that statistical conclusions are invariant across various data-editing strategies and methodologies. However, there may be cases in which researchers’ statistical, and therefore theoretical, claims depend on *whether* or *how* they have transformed or trimmed the data. Indeed, in investigating Diederik Stapel’s infamous body of work for instances of deceptive research practices, an investigatory panel specifically noted how the elimination or inclusion of extreme scores affected the statistical conclusions:

On the one hand, “outliers” (extreme scores on usually the dependent variable) were removed from the analysis where no significant results were obtained. This elimination reduces the variance of the dependent variable and makes it more likely that “statistically significant” findings will emerge. . . . Conversely, the Committees also observed that extreme scores of one or two experimental subjects were kept in the analysis where their elimination would have changed significant differences into insignificant ones; there was no mention anywhere of the fact that the significance relied on just one or a few subjects. (Levelt Committee, Noort Committee, and Drenth Committee 2012:49)

Obviously, it strikes us as problematic when statistical and theoretical conclusions are dependent not on the data *per se* but on the creative use (or misuse) of statistical methods and data-editing strategies—what Simmons, Nelson, and Simonsohn (2011) have referred to as “experimenter degrees of freedom.” Although Stapel may have been guilty of not disclosing his decisions to include or exclude participants (and outright fraud in other cases), the fact that he sometimes engaged in outlier elimination (and other times chose not to) is not inconsistent with standard practices. In fact, the authors of the Stapel report even seem conflicted about whether it was appropriate to eliminate extreme scores. The bottom line is that decisions about whether to engage in data editing that are based on whether the data meet the assumptions of the statistical model leave the researcher in a precarious position: damned if you do and damned if you don’t.

Although there have been several documented cases of inappropriate data editing within the psychological literature (e.g., that of Diedrick Stapel), the issue of data editing is clearly of concern across all of the social sciences, including sociology (John, Loewenstein, and Prelec 2012; Leahey et al. 2003). The tension surrounding the appropriateness of eliminating outliers was illuminated by an exchange between Kahn and Udry (1986) and Jasso (1986) in the *American Sociological Review* regarding an analysis of intercourse frequency among married couples: Kahn and Udry criticized Jasso's original analysis by arguing that her inclusion of outliers was inappropriate and biased the statistical results; Jasso countered by arguing that the exclusion of outliers in Kahn and Udry's reanalysis produced "sample truncation bias." This divergence on the inclusion of outliers highlights a common predicament: There is not always a clear solution to the presence of outliers, and decisions to include or exclude them often come down to a judgment call.

The scope of the data-editing problem for statistical inference is difficult to assess from published work, in part because there is little oversight or consistency in regard to how data-editing procedures are carried out (Leahey 2008) and in part because few articles include serious discussion of how specific data-editing decisions affect statistical conclusions. Nevertheless, it is clear that data editing is a relatively common component of statistical analysis. Notable examples from the literature include the common use of logarithmic transformations for analyses that include estimates of income (e.g., Olsen and Dahl 2007; Semyonov and Lewin-Epstein 2011) and homicide rates (Lederman, Loayza, and Menéndez, 2002). Although decisions regarding whether to transform variables are presumably based on the need to bring the data in line with modeling assumptions, these decisions represent an important source of flexibility in data analysis—a flexibility that can be exploited either intentionally or unintentionally (Simmons et al. 2011).

The exploitation of flexible analysis techniques is a problem for science. However, the critical question concerns the precise nature of this problem: Is it that people fail to report faithfully the many decisions that ultimately exploit this flexibility? Or is it that there is too much flexibility with data analysis techniques to begin with? Depending on how we perceive the problem, it suggests different solutions. If the problem is that people do not faithfully report the many decisions that exploit the flexibility of available statistical algorithms, then the obvious solution is to require full disclosure of data analysis methods in an open-source

forum, as suggested by Freese (2007). However, if the problem is that there is too much flexibility to begin with, then the solution would seem to lie in the development (or use of) procedures that reduce this flexibility (Ho et al. 2007). Thus, although full disclosure is important, we believe that the more fundamental problem lies with the use of standard statistical techniques, which permit, and in some cases demand, that researchers engage in data editing. Assuming this is the case, then one reasonable approach is to use analysis techniques that are robust to the types of decisions that researchers would otherwise be compelled to make in order to bring their data in line with the modeling assumptions (cf. Beck and Jackman 1998).

### 3. THE GENERAL MONOTONE MODEL

Fundamentally, the GeMM is an algorithm for detecting and modeling monotone statistical relationships in regression contexts. The primary difference between GeMM and standard LS approaches lies in the fitness function. In LS regression, the goal is to find the regression coefficients that minimize the sum of the squared differences between the observed and the predicted values. In contrast, in GeMM the goal is to find the regression coefficients that minimize the difference in the ordinal correspondence (i.e., that minimize the number of rank-order inversions) between the observed and predicted values, as defined by Kendall's (1938)  $\tau$ . In this way, GeMM attempts to find the solution that provides the best *monotonic* (i.e., rank-order) fit to the data, as opposed to finding the best *linear* LS fit to the data. Thus, GeMM is a variant of the maximum rank correlation estimator (Cavanagh and Sherman, 1998; Han 1987). As demonstrated below, GeMM has superior statistical power relative to ordinary LS (OLS) to detect nonlinear but monotone statistical relationships, without requiring the researcher to model the nonlinearity directly or engage in data editing. The reason for this is that the rank-order correlation  $\tau$ , on which GeMM is based, is invariant to monotone transformation on the criterion variable. It is also important to note that GeMM suffers little loss in statistical power compared with OLS when the statistical relationship is linear and the data satisfy standard OLS assumptions (Dougherty and Thomas 2012). Because GeMM is invariant to transformation on the criteria, unaffected by nonlinearities, and should be less sensitive to extreme scores (a property we demonstrate below), it provides a new tool for modeling messy

data that would otherwise require editing or more specialized statistical algorithms.

In its simplest form, GeMM consists of a one-parameter model (i.e., one predictor), which is used to predict the criterion variable<sup>1</sup> of interest. In this context, GeMM is actually identical to Kendall's (1938)  $\tau$  correlation coefficient, but it is expressed in a model form. Rather than expressing the relationship between  $X$  and  $Y$  directly, we substitute  $\hat{Y}$  for  $X$  to show the model-form equivalence of  $\tau$  for a single predictor:

$$\hat{Y} = \beta X. \quad (1)$$

In equation (1), we wish to find a value for  $\beta$  that minimizes the *incorrectly* predicted paired comparisons, as defined by equations (2) to (6):

$$\tau(\hat{Y}, Y) = (C - D) / \text{sqrt}[(\text{Pairs} - T_p) * (\text{Pairs} - T_c)], \quad (2)$$

$$C = \text{Prop}(Y_i > Y_j \cap \hat{Y}_i > \hat{Y}_j) + \text{Prop}(Y_i < Y_j \cap \hat{Y}_i < \hat{Y}_j), \quad (3)$$

$$D = \text{Prop}(Y_i > Y_j \cap \hat{Y}_i < \hat{Y}_j) + \text{Prop}(Y_i < Y_j \cap \hat{Y}_i > \hat{Y}_j), \quad (4)$$

$$T_p = \text{Prop}(Y_i \geq Y_j \cap \hat{Y}_i = \hat{Y}_j) + \text{Prop}(Y_i \leq Y_j \cap \hat{Y}_i = \hat{Y}_j), \quad (5)$$

and

$$T_c = \text{Prop}(Y_i = Y_j \cap \hat{Y}_i \leq \hat{Y}_j) + \text{Prop}(Y_i = Y_j \cap \hat{Y}_i \geq \hat{Y}_j), \quad (6)$$

where  $\text{Pairs} = N(N-1)/2$ , the number of unique paired comparisons;  $C$  is the number of concordant paired comparisons;  $D$  is the number of discordant pairs;  $T_p$  is the number of ties on the predictor; and  $T_c$  is the number of ties on the criterion. With only one predictor, only the sign of  $\beta$  matters, which provides the direction of the relationship between  $\hat{Y}$  and  $Y$ . Thus, for the one-predictor case, the specific value of  $\beta$  is irrelevant, and the strength of the predictor is defined by the value of  $\tau$ . Note that there is no intercept parameter in equation (1), because it is not necessary for predicting the ordered relationship.

Equation (1) can be generalized to the multiple predictor case:

$$\hat{Y} = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k. \quad (7)$$

In equation (7), the different coefficients are estimated to maximize model fit and can therefore take on any real number, which allows the



variables to differentially contribute to the overall fit between the data,  $Y$ , and the model estimates,  $\hat{Y}$ . In this context, the magnitudes of the  $\beta$  values are interpreted as the *relative* contribution of each predictor for predicting the ordinal values of  $Y$ . In contexts in which predictors are uncorrelated, the  $\beta$  weights can be viewed as the relative importance of each variable for characterizing the ordinal values of  $Y$ .

Parameter estimation is achieved computationally, rather than analytically, because there are no currently available methods for deriving optimal weights to maximize the rank-order correspondence between a model and the data. In the present analyses, we used a genetic algorithm to search the parameter space for the best-fit parameter estimates. Prior work (Dougherty and Thomas 2012) illustrated that genetic search works well for estimating the optimal weights for simulated data with known parameters.

In the analyses that follow, we fit data within the context of minimizing model complexity. This was achieved by using a variant of the Bayesian information criterion (BIC). Raftery (1995) showed that the BIC could be estimated from

$$\text{BIC} = N \log(1 - R^2) + k \log(N), \quad (8)$$

where  $N$  is the sample size,  $R^2$  is the squared multiple correlation, and  $k$  is the number of parameters. One problem with applying equation (8) directly is that GeMM is designed to predict rank orders. However, Kendall and Gibbons (1990) showed that under bivariate normality,<sup>2</sup> Pearson's  $r$  could be estimated from  $\tau$  using

$$r - \tau = \sin(\pi/2\tau). \quad (9)$$

Substituting equation (9) for the value of  $R^2$  in (8) yields equation (10):

$$\text{BIC}_\tau = N \log(1 - (\sin[\pi/2\tau])^2) + k \log(N). \quad (10)$$

Equation (10) is the value of the BIC estimated from the  $\tau$ -to- $r$  transformation. However, because the value of  $r_\tau$  shows greater variability than  $r$  (Rupinski and Dunlap 1996), we use an adjusted form of  $r_\tau$  based on sample size and the number of predictors used in the regression. Specifically, we define  $r'_\tau$  as

$$r'_\tau = \sin[\pi/2\tau\omega], \quad (11)$$

where

$$\omega = (N - P - 1)/N, \quad (12)$$

where  $\omega$  is a weighting function based on the number predictors,  $P$ , used in the regression and sample size,  $N$ . Because  $\omega$  serves to deweight the value of  $\tau$  for smaller sample sizes, it reduces the variance of the  $\tau$ -to- $r$  transformation. Because  $\omega$  goes to 1.0 as  $N$  increases, the asymptotic value of the  $\tau$ -to- $r$  transformation is preserved. Substituting  $r'_\tau$  into equation (12) gives

$$\text{BIC}'_\tau = N \log(1 - r'^2_\tau) + k \log(N). \quad (13)$$

Model selection based on equation (13) ( $\text{BIC}'_\tau$ ) is assessed on the fit of the model to the data as given by the degree of monotonic relationship expressed by the  $\tau$ -to- $r$  transformation, adjusted for model complexity. Dougherty and Thomas (2012) showed that model fitting based on  $r'^2_\tau$  is invariant to monotone transformation on  $y$ , whereas model fit based on the linear  $r^2$  can suffer from considerable loss of power when statistical relations deviate from strict linearity. Furthermore, Dougherty and Thomas (2012) illustrated that GeMM's estimated parameters approximated the metric population values, and they were unaffected by nonlinearities. This later result occurs for the same reason that ordinal multidimensional scaling solutions approximate metric properties of the data: The number of constraints on the rank-order solution increases exponentially as sample size increases (Dougherty and Thomas 2012; see also Shepard 1962, 1966).

The base GeMM algorithm described above and in Dougherty and Thomas (2012) searches the parameter space to find coefficients that maximize the value of  $\tau$ . However, a simple modification to this process involves maximizing the linear fit ( $R^2$ ), conditional on the optimal ordinal fit. This can be achieved in GeMM by sorting all models with equivalent (maximal) ordinal fit by their corresponding values of  $R^2$ . This yields the vector of  $\beta$  values that optimize the *linear* fit, conditional on the set of coefficients that maximize ordinal fit. Note that the coefficients derived from this process are scale independent and are not directly comparable with coefficients derived from OLS, because there is an infinite number of parameter values that will yield an equivalent solution. This is because GeMM lacks an intercept term and because the fit statistic,  $\tau$ , is invariant to monotone transformation. However, we

may obtain a comparable LS model, one that is conditioned on maximizing  $\tau$ , by regressing the criterion value  $Y$  on the predicted values of  $Y$  obtained from GeMM. In other words, we can use the OLS machinery to rescale the GeMM fitted weights to the LS solution that simultaneously maximizes the rank-order correspondence between the criterion and fitted values. We refer to this procedure as order-constrained least-squared optimization (OCLO; Tidwell et al. 2014). In principle, the OCLO solution is a special case of the base GeMM model in which weights are rescaled to minimize the sum of squared errors, conditional on the optimized ordinal fit. The end result of applying OCLO is a set of  $\beta$  coefficients that are directly comparable with those obtained via OLS regression.

#### **4. REDUCING FLEXIBILITY IN ANALYSIS: AN ILLUSTRATION OF GEMM ON TWO DATA SETS**

Flexibility in data analysis presents an appreciable challenge when different analysis techniques or data-editing decisions change the substantive conclusions. Here, we argue that GeMM offers a promising approach for reducing this flexibility. GeMM assumes that the predictors are interval scale, permitting the model to take the traditional additive form, but it treats the criterion variable as ordinal—allowing ordinal, interval, ratio, and even nominal (in some cases) scale variables to serve as the criterion. A key feature of GeMM is that it is designed to model the monotone relations of the data. This feature means that GeMM is invariant to transformation on  $Y$  and should be relatively robust to extreme scores, or outliers, compared with LS procedures. Consequently, GeMM's solution should be relatively stable across different methods for identifying and eliminating extreme scores. In contrast, because LS procedures seek to maximize linear fit, extreme scores can exert undue influence on LS solutions, even when only a small number of scores are extreme. Below, we demonstrate that a small number of extreme scores can sometimes drive manifest effects, and other times hide effects when data are analyzed using LS procedures. In addition, we illustrate that different methods for identifying and eliminating extreme scores and nonlinearities can lead to inconsistent statistical conclusions when analyzed with LS approaches. In contrast, GeMM provides more consistent statistical conclusions across multiple data-editing strategies in our demonstrations.

#### 4.1. *When Extreme Scores Drive Effects: The Case of Racial Bias*

What is the relationship between explicit measures of racial bias, implicit measures of racial bias, and motivation to control prejudice? Prior work on this topic suggests that explicit measures of racial bias capture some element of a person's true underlying attitude but that they are subject to response biases on the part of the participant (e.g., Dunton and Fazio 1997; Fazio et al. 1995). For example, how people respond on the Attitudes Toward Blacks (ATB) scale appears to be moderated by people's motivation to control prejudice (Plant and Devine 1998). Plant and Devine (1998) identified two separate forms of such motivation: an internal motivational factor and an external motivational factor. The internal factor tests for motivations stemming from the belief that stereotypes are morally wrong or personally unacceptable. The external factor tests for motivations stemming from the desire to avoid social censure—in other words, the belief that *other people* believe that stereotypes are morally wrong or unacceptable. Either type of motivation could lead to similar self-censoring of socially unpopular attitudes, but that similarity belies the important differences between people who are driven by one versus the other motive type.

Partly to deal with this problem of self-censoring, considerable research has validated the use of implicit measures of racial bias. Perhaps the most well-known implicit measure is the implicit association test (IAT; Greenwald, McGhee, and Schwartz 1998), a measure that uses response times to assess the difficulty respondents have classifying white or black faces simultaneous to categorizing other stimuli as good or bad. More recently, other implicit measures have been developed that do not rely on response times. For example, Payne and colleagues (Payne et al. 2005; Payne, Burkley, and Stokes 2008) developed the affect misattribution procedure (AMP), which involves showing people a stimulus word or picture that they are told to ignore, followed by a Chinese pictograph. Participants are instructed to rate how pleasant the pictograph is, ignoring the stimulus that precedes it. However, the affect associated with the first stimulus is expected to “bleed over” to the pictograph, revealing how positively or negatively respondents *actually* feel about that *first* stimulus, which they are supposed to be ignoring. Payne and colleagues showed that scores on the AMP reflect subtle in-group preferences among both white and black respondents and that this in-group bias occurs whether or not participants are warned to avoid

being biased on the measure (an external motivation to control prejudice). In contrast, participants who reported strong *internal* desires to avoid prejudice appeared to modify their *explicit* racial attitudes. Consequently, the self-reported attitudes of these participants hardly correlated at all with their scores on the AMP. Among participants who reported weaker internal desires to avoid prejudice, AMP scores were highly correlated with explicit prejudice.

An important question regarding the measurement of racial attitudes is the degree to which explicit measures of racial attitudes capture one's true attitude and the degree to which they are subject to people's motivation to control their expression of their attitude. This problem is reflected in the results found by Payne et al. (2005), as well as by many other researchers (e.g., Devine et al. 2002; Dunton and Fazio 1997; Plant and Devine 1998; Plant, Devine, and Brazy, 2003). Theoretically, a case can be made for both the inclusion and exclusion of external and internal motivation to control prejudice as predictors of racial attitudes. On one hand, it makes sense that participants would wish to avoid social censure (an external motivation) as a consequence of openly admitting that they are racially biased. For this reason, it is clear that explicit motivations should play an important role in how participants respond on the ATB scale and other such explicit attitude measures. On the other hand, the belief that racism is morally wrong (an internal motivation) might lead them to explicitly state more positive attitudes toward blacks than they actually hold. Either way, researchers who want to know people's *true* attitudes would seem to do well by accounting for these types of motivations in studies of prejudice or other socially sensitive topics.

**4.1.1. Data and Analyses.** We reanalyzed data initially published by Siegel et al. (2012). The original sample included 213 University of Maryland undergraduate students (128 women). Of these, 15 participants were missing data on one or more measures and were therefore excluded from the analysis. Each participant was measured on 10 variables, including three measures of racial attitudes (the ATB scale, the Race AMP [Race-AMP], and the Racism IAT [Race-IAT]), the motivation to control prejudice subscales (the External Motivation Scale [EMS] and the Internal Motivation Scale [IMS]), two measures of cognitive control (the Stroop test and the Stop Signal Task), and three measures of political attitudes (explicit political attitudes [EPA], a Political

AMP [Pol-AMP], and a Political IAT [Pol-IAT]). Additional details of the study, including how the various tasks were constructed, administered, and scored, are provided in Siegel et al. (2012).<sup>3</sup>

Siegel et al. (2012) were concerned primarily with understanding the relationship between the IAT and the measures of cognitive control. Using factor analyses, they showed that both the Pol-IAT and the Race-IAT loaded on two factors: their respective attitude factor and a cognitive control factor. That is, performance on the IAT appeared to be predicted best by a model that assumed that the IAT measures both the target attitude and cognitive control. Although the Race-IAT was unrelated to the explicit ATB scale, it was highly related to the Race-AMP. Moreover, the ATB scale was correlated with the Race-AMP and both the EMS and IMS. This pattern of correlations suggests that scores on the ATB are dependent on an (implicit) attitude factor and both forms of motivation to control prejudice. However, Siegel et al. did not explore these relationships in depth. Thus, the substantive goal of our reanalysis was to identify the best predictors of scores on the ATB from the collection of variables included in the study by Siegel et al. There were two methodological goals: (1) to demonstrate that the substantive conclusions could change depending on how extreme scores were identified and treated and (2) to test whether GeMM was less sensitive to the treatment of outliers.

Using LS regression, we tested the hypothesis that both internal and external motivations to control prejudice were negatively related to participants' self-reported (explicit) racial biases, as measured by the ATB scale, independent of participants' implicit racial bias, as measured by the Race-AMP. Using the classical null hypothesis significance testing (NHST) approach with  $\alpha = .05$ , we found the predicted relationship: The ATB scale was significantly and positively related to the Race-AMP, and the ATB scale was negatively related to both the EMS and IMS. Summary statistics for this analysis are presented in Tables 1 and 2, in the top row, labeled "Full data." Overall, these three variables accounted for 13.2% of the variance in ATB scores, with the rank-order correlation between the predicted and the actual values of the ATB yielding a value of  $\tau = .239$ . Thus, on the basis of this analysis, it seems that we are justified in supporting the theory that self-reported (explicit) racial bias is a function of people's implicit racial bias, their internal motivations to control racial bias, and their external motivations to avoid being seen as racially biased. Or are we?

**Table 1.** Fit Indices from the Various Models

	BIC' <sub>τ</sub>	BIC	τ	R	k
<b>OLS-NHST</b>					
Full data (N = 198)	-9.905	-12.283	.239	.364	3
Univariate (N = 195)	-12.837	-6.282	.23	.288	2
DFFITS (N = 191)	-15.321	-10.142	.245	.321	2
Cook's D (N = 185)	-22.196	-24.986	.299	.444	3
<b>Robust regression</b>					
Huber	-15.651	-8.965	.241	.307	2
Bisquare	-15.005	-9.132	.238	.308	2
Hampel	-10.646	-11.886	.243	.362	3
<b>OLS-BIC</b>					
Full data (N = 198)	-9.242	-12.445	.236	.365	3
Univariate (N = 195)	-11.472	-6.551	.223	.29	2
DFFITS (N = 191)	-10.982	-11.093	.224	.328	2
Cook's D (N = 185)	-21.074	-25.253	.295	.445	3
<b>Ordered logistic</b>					
Full data (N = 198)	162.602	166.442	.277	.3	35
Univariate (N = 195)	153.567	157.535	.266	.285	33
DFFITS (N = 191)	147.438	150.248	.268	.298	32
Cook's D (N = 185)	132.55	131.86	.377	.443	33
<b>GeMM</b>					
Full data (N = 198)	-17.835	-5.878	.251	.282	2
Univariate (N = 195)	-15.389	-3.89	.242	.267	2
DFFITS (N = 191)	-17.348	-7.524	.254	.301	2
Cook's D (N = 185)	-23.914	-24.109	.306	.44	3

Note: k is the number of significant or retained parameters. For ordered logistic, k includes the number of significant threshold parameters. Thus, for k = 35, there are three significant predictors (External Motivation Scale, Race Attitude Misattribution Procedure, and Internal Motivation Scale) and 32 significant threshold parameters. BIC = Bayesian information criterion; GeMM = general monotone model; NHST = null hypothesis significance testing; OLS = ordinary least squares.

Figure 1 plots the histograms for the 10 variables in the study, and Figure 2 provides the bivariate scattergrams for each predictor (x-axis) plotted against the ATB scale. Three findings should be evident from inspection of the graphs. First, the relationships identified by linear regression are not easily discernible from the bivariate plots, although by itself, this fact might not be terribly concerning—subtle associations do not always yield their secrets to the naked eye. Second, many of the predictors are poorly distributed, which is somewhat more concerning, given the assumptions underlying LS regression. Third, there appears to be a small number of extreme scores (outliers?) in the distribution of the

**Table 2.** Standardized Regression Coefficients Revealed for Each Model

	P-EXP	Pol-IAT	Pol-AMP	Race-AMP	EMS	IMS	Race-IAT	Stroop	Stop Signal	Task
<b>OLS-NHST</b>										
Full data ( $N = 198$ )	-.054 (.074)	-.083 (.092)	.016 (.081)	.17 <sup>a</sup> (.073)	-.217 <sup>a</sup> (.069)	-.172 <sup>a</sup> (.074)	.016 (.082)	.03 (.086)	-.027 (.074)	
Univariate ( $N = 195$ )	-.059 (.070)	-.076 (.088)	.028 (.077)	.152 <sup>a</sup> (.070)	-.194 <sup>a</sup> (.067)	-.11 (.070)	.009 (.072)	.035 (.081)	-.034 (.069)	
DFFITs ( $N = 191$ )	-.103 (.068)	-.017 (.083)	.081 (.073)	.147 <sup>a</sup> (.070)	-.199 <sup>a</sup> (.063)	-.127 (.069)	.025 (.073)	.014 (.079)	-.02 (.067)	
Cook's $D$ ( $N = 185$ )	-.04 (.066)	-.054 (.079)	.007 (.070)	.167 <sup>a</sup> (.067)	-.244 <sup>a</sup> (.064)	-.17 <sup>a</sup> (.068)	.049 (.072)	.045 (.077)	-.039 (.064)	
<b>Robust regression</b>										
Huber ( $N = 198$ )	-.059 (.071)	-.078 (.088)	.018 (.078)	.19 <sup>a</sup> (.070)	-.229 <sup>a</sup> (.066)	-.13 (.070)	.034 (.078)	.05 (.082)	-.049 (.070)	
Bisquare ( $N = 198$ )	-.064 (.074)	-.076 (.092)	.02 (.081)	.201 <sup>a</sup> (.081)	-.222 <sup>a</sup> (.073)	-.125 (.073)	.025 (.081)	.047 (.085)	-.044 (.073)	
Hampel ( $N = 198$ )	-.063 (.073)	-.078 (.090)	.023 (.080)	.177 <sup>a</sup> (.072)	-.21 <sup>a</sup> (.068)	-.144 <sup>a</sup> (.072)	.025 (.080)	.034 (.084)	-.036 (.072)	
<b>OLS-BIC</b>										
Full data ( $N = 198$ )	—	—	—	.168 <sup>a</sup> (.070)	-.213 <sup>a</sup> (.068)	-.200 <sup>a</sup> (.069)	—	—	—	
Univariate ( $N = 195$ )	—	—	—	.181 <sup>a</sup> (.066)	-.172 <sup>a</sup> (.065)	—	—	—	—	
DFFITs ( $N = 191$ )	—	—	—	—	-.233 <sup>a</sup> (.061)	-.196 <sup>a</sup> (.063)	—	—	—	
Cook's $D$ ( $N = 185$ )	—	—	—	.168 <sup>a</sup> (.063)	-.241 <sup>a</sup> (.062)	-.206 <sup>a</sup> (.063)	—	—	—	

(continued)

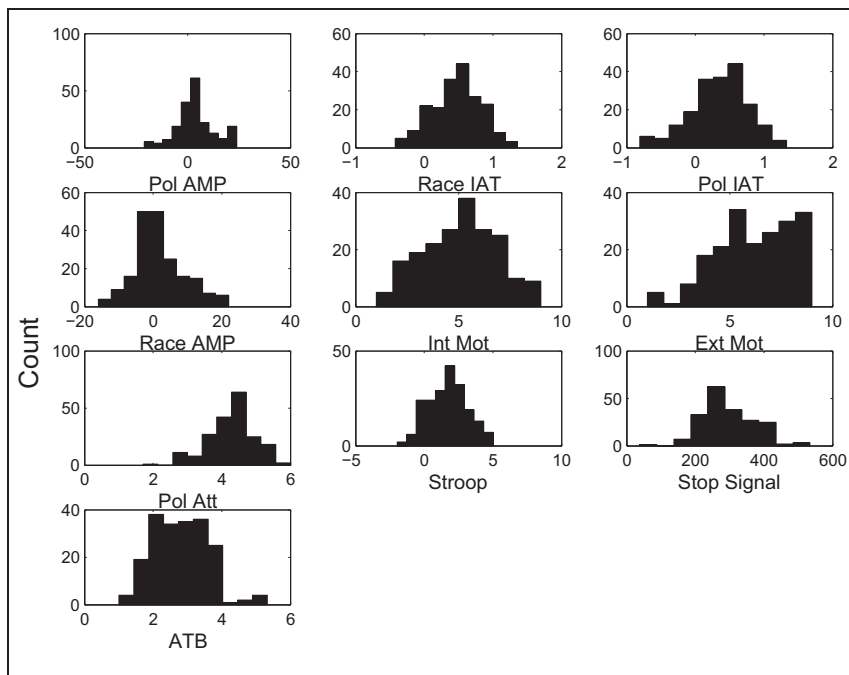


**Table 2.** (continued)

	P-EXP	Pol-IAT	Pol-AMP	Race-AMP	EMS	IMS	Race-IAT	Stroop	Stop Signal Task
<b>Ordered logistic</b>									
Full data ( $N = 198$ )	-.098 (.138)	-.15 (.161)	.032 (.144)	<b>.383<sup>a</sup></b> (.142)	<b>-.447<sup>a</sup></b> (.141)	<b>-.292<sup>a</sup></b> (.144)	.05 (.155)	.071 (.158)	-.061 (.135)
Univariate ( $N = 195$ )	-.108 (.139)	-.141 (.162)	.048 (.145)	<b>.375<sup>a</sup></b> (.144)	<b>-.431<sup>a</sup></b> (.142)	-.230 (.145)	.043 (.156)	.078 (.159)	-.07 (.137)
DFFITs ( $N = 191$ )	-.187 (.144)	-.052 (.165)	.147 (.148)	<b>.338<sup>a</sup></b> (.148)	<b>-.466<sup>a</sup></b> (.145)	<b>-.326<sup>a</sup></b> (.151)	.058 (.160)	.031 (.164)	-.024 (.141)
Cook's $D$ ( $N = 185$ )	-.058 (.144)	-.136 (.167)	.005 (.150)	<b>.436<sup>a</sup></b> (.150)	<b>-.573<sup>a</sup></b> (.150)	<b>-.409<sup>a</sup></b> (.155)	.067 (.164)	.107 (.168)	-.046 (.142)
<b>GeMM</b>									
Full data ( $N = 198$ )	—	—	—	<b>.133<sup>a</sup></b> (.077)	<b>-.236<sup>a</sup></b> (.085)	—	—	—	—
Univariate ( $N = 195$ )	—	—	—	<b>.113<sup>a</sup></b> (.071)	<b>-.217<sup>a</sup></b> (.084)	—	—	—	—
DFFITs ( $N = 191$ )	—	—	—	<b>.112<sup>a</sup></b> (.067)	<b>-.224<sup>a</sup></b> (.074)	—	—	—	—
Cook's $D$ ( $N = 185$ )	—	—	—	<b>.146<sup>a</sup></b> (.076)	<b>-.270<sup>a</sup></b> (.079)	<b>-.147<sup>a</sup></b> (.098)	—	—	—

*Note:* Dashes indicate that the predictor variable was not included in the model. Coefficients and standard errors listed for GeMM were computed on the basis of 1,000 bootstrap runs. BIC = Bayesian information criterion; EMS = External Motivation Scale; GeMM = general monotone model; IMS = Internal Motivation Scale; NHST = null hypothesis significance testing; OLS = ordinary least squares; P-EXP = explicit political attitude; Pol-AMP = Political Affect Misattribution Procedure; Pol-IAT = Political Implicit Association Test; Race-AMP = Race Attitude Misattribution Procedure; Race-IAT = Racism Implicit Association Test.

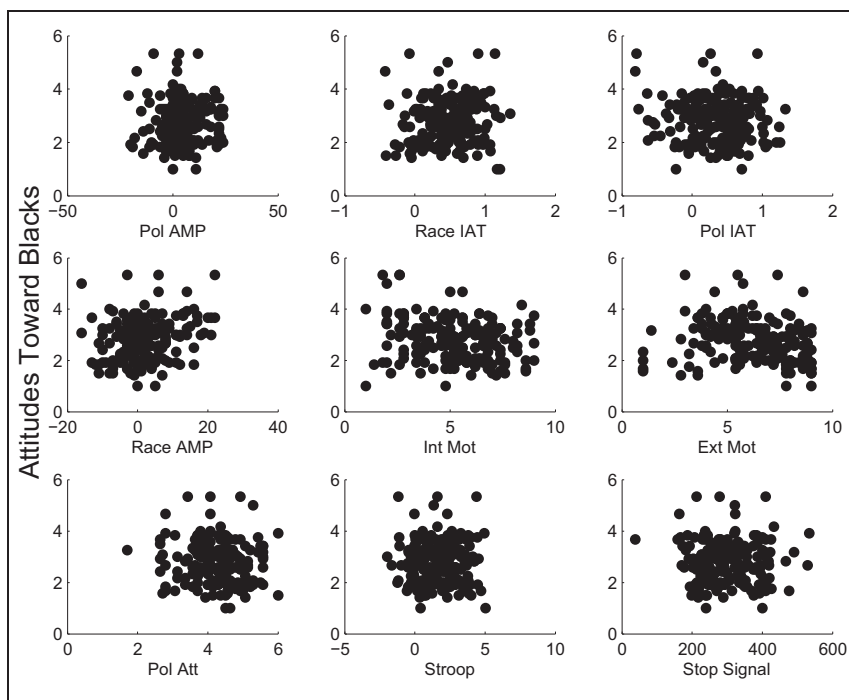
<sup>a</sup>predictor retained by the model using BIC (OLS-BIC and GeMM) or predictor was significant at  $p \leq .05$  (OLS-NHST, robust regression, ordered logit).



**Figure 1.** Histograms for the 10 variables reported in Siegel et al. (2012).

*Note:* AMP = affect misattribution procedure; Ext = external; IAT = implicit association test; Int = internal; Mot = motivation; Pol = political; Pol Att = explicit political attitude.

ATB scale, which could prove to be especially problematic for standard regression techniques and might even “require” data editing prior to analyzing the data with OLS. Given the presence of extreme scores and the nonnormality of the distributions, we conducted a series of follow-up analyses to determine the robustness of the conclusions to different methods for reducing the influence of violations of the assumptions of linear LS regression. The first approach was to conduct outlier analyses to identify and eliminate potentially problematic data points. There are a variety of outlier detection methods, but we confined ourselves to three techniques: (1) univariate outlier analysis, (2) Cook’s *D*, and (3) DFFITS.<sup>4</sup> Application of these three approaches to the data set resulted in the identification of 3, 13, and 8 extreme scores, respectively. After trimming these data points out of the sample, we reanalyzed the data, again using OLS with an  $\alpha$  value of .05.



**Figure 2.** Scattergram plotting attitudes toward blacks ( $y$ -axis) against nine predictor variables.

*Note:* AMP = affect misattribution procedure; Ext Mot = external motivation to control prejudice; IAT = implicit association test; Int Mot = internal motivation to control prejudice; Pol = political; Pol Att = explicit political attitude.

The results of the analyses after eliminating these extreme data points are also presented in Tables 1 and 2. As can be seen, two approaches to eliminating extreme scores revealed that two predictors were significant, and one approach revealed three significant predictors. Surprisingly, the elimination of a mere 1.5% of the data (three data points) was sufficient to knock out internal motivation as a significant predictor. This was not just a matter of the  $p$  value's hovering around .05 and bouncing back and forth over the threshold, as the  $p$  value for IMS was .02 for the full data set, but it jumped to nearly .12 after eliminating only three data points. Thus, the decision to exclude IMS for the univariate trimmed data is not an inconvenient by-product of the conventional, yet arbitrary, value of  $\alpha = .05$ . Combined with the analyses using the full data set,

there appears to be no clear “winner” regarding which statistical conclusions are most appropriate.

The fact that different methods for dealing with extreme scores resulted in different statistical models is problematic for the purposes of theory testing. Therefore, we conducted a series of analyses using three variations on robust LS regression and ordinal logistic regression. Robust statistics are designed to deweight extreme scores on the basis of their distance from the mean, and therefore they are purported to have better statistical properties when distributional assumptions are violated. The methods used here are the Huber, bisquare, and Hampel methods, which were implemented again using NHST. Ordinal logistic regression treats the criterion variable as an ordered category, and it estimates thresholds for each category. For our purposes, we modeled the raw data rather than creating binned responses.<sup>5</sup> In addition, we also reanalyzed the full and trimmed data sets using the BIC as a model selection method. The BIC model selection method has the advantage of not relying on the arbitrary .05 threshold for statistical significance.

The results of the robust, OLS-BIC model selected, and ordinal regression analyses are presented in the middle portions of Tables 1 and 2. Once again, the results are inconclusive, with two of the robust approaches (Huber and bisquare) yielding two significant predictors and one approach (Hampel) yielding three significant predictors. Model selection using the BIC to select predictors was even more inconsistent, as it yielded *two different* two-predictor models, as well as a three-predictor model. The results of the ordinal logistic model are a bit more complicated. This model fits the ordinal properties of the data, but to do so, it estimates thresholds for each of the ordered categories using the full model with all predictors. As can be seen, this model fits the ordinal properties quite well, but at the expense of a considerable increase in model complexity due to the need to estimate the threshold parameters. Even so, this method also produced different models between the full data set and the univariate trimmed data set in which only three observations were eliminated: three predictors were significant on the full data set, but only two were significant on the reduced (univariate trimmed) data set.

The inconsistency across outlier and data analysis methods is undesirable for many reasons, but principally *because it allows the researcher the freedom to choose which theoretical conclusions to draw from the data*, rather than forcing theoretical conclusions to be

constrained by the data—a principle at the heart of basic science. Given these inconsistencies, we reanalyzed the data using GeMM. In contrast to traditional LS approaches, GeMM models data at the level of paired comparisons, as we explained earlier. Because GeMM does not model data using a distance metric and makes less stringent assumptions of the data, it should be more robust to the presence of extreme scores and nonlinearities.

The results using GeMM are presented at the bottom of Tables 1 and 2.<sup>6</sup> As can be seen, GeMM resulted in a two-parameter solution when applied to the full data set, and this solution was consistent for both the univariate and DFFITS methods for eliminating outliers. Note that the two-predictor solutions include the same predictors (EMS and Race-AMP) identified as significant by the Huber and bisquare procedures. GeMM was not completely insensitive to outlier deletion methods, as it identified a three-predictor solution when the 13 observations were trimmed using Cook's  $D$ . However, the fact that it was stable for both the univariate and DFFITS methods (which required deleting only 3 and 8 observations) suggests that it is relatively more robust than OLS. In fact, further analyses on these data indicated that the OLS solution changed from a three- to a two-predictor model even after eliminating just one data point, the single most extreme value on the ATB scale. This pattern of analyses suggests that GeMM has a much greater tolerance for extreme scores than OLS. Coincidentally, the robust regression procedures also resulted in a three-predictor model when applied to the Cook's  $D$ -trimmed data.

If we consider the full data set, is the two-parameter GeMM solution preferable to the three-predictor LS solutions, and are we justified in accepting the two-parameter model over one that includes three predictors? There are two ways to address this question: (1) compare the fit indices for GeMM with those of OLS and (2) conduct cross-validation analyses. We consider both in turn.

*4.1.2. Comparing Fit and Cross-validation.* Inspection of the fit indices indicates that the two-parameter GeMM solution actually provides a better fit to the data in terms of accurately capturing the ordinal properties of the data than all of the other approaches except ordinal regression, even the models that included three parameters, as shown by the values for  $BIC_{\tau}'$  and  $\tau$ . Although the LS solutions fit the data better when evaluated in terms of the multiple  $R$  and BIC, these indices are

highly suspect because they require the assumption of linearity: Inasmuch as the linear (LS) solution is relatively poor at capturing the monotonic relations of the data (as given by  $\tau$  and  $BIC_{\tau}'$ ), we must be wary of interpreting a solution that makes the stronger assumptions of normality and linearity. Although ordinal logistic regression had a higher value of  $\tau$ , this came with considerable increase in model complexity. As we show below, this increase in model complexity can lead to overfitting.

One interesting aspect of these fit indices is that although the LS versions (ordinary and robust regression) provide better fit to the data in terms of  $R^2$ , this fit comes at a cost of accurately capturing the ordinal properties of the data. For instance, for the full sample, OLS accounts for 13.2% of the variance ( $R^2 = .132$ ), but it has a rank-order correlation of only .239. In contrast, when GeMM is applied to the same data it accounts for only 8.0% of the variance ( $R^2 = .080$ ), but it is better able to account for the ordinal properties of the data, with a rank-order correlation of .251. This pattern also holds for all three methods for trimming outliers.

We used split-half cross-validation to evaluate out-of-sample prediction: Which statistical algorithm provides the best predictive accuracy when the estimated parameters are used to predict new observations? This approach has the advantage that it directly addresses the problem of overfitting, in which statistical models tend to show poorer accuracy (i.e., shrinkage) at predicting new observations compared with the fit to the original estimation sample. The cross-validation approach has the added benefit, however, of allowing us to evaluate statistical power, or the probability that each of the predictor variables will be identified as a “significant” predictor (or included in the selected model). We conducted a split-half cross-validation using the full data set ( $N = 198$ ), in which half of the data were randomly sampled and used to estimate model parameters. The remaining half of the data were used as the holdout sample. For each “replication” of this procedure, we recorded for each algorithm which parameters were recovered, fit indices, and  $\beta$  weights. For methods using NHST, a parameter was classified as recovered if it was significant at the .05 level using a  $t$  test on the regression coefficient. Out-of-sample predictive accuracy was assessed by applying the recovered statistical model to the holdout sample (i.e., the  $\beta$  weights for nonsignificant predictors were set to zero). We computed the

multiple  $R$ ,  $\tau$ , and the corresponding values of BIC and  $BIC_{\tau}'$ . This procedure was repeated 500 times for each statistical model.

The results of the cross-validation analyses are presented in Tables 3 and 4. Table 3 shows the probability of recovering each predictor when each algorithm is provided half of the data. Recall that on the full sample, OLS recovered a three-predictor model consisting of the Race-AMP, the IMS, and the EMS, whereas GeMM recovered a two-predictor model consisting of the Race-AMP and the EMS. Overall, GeMM was more likely to recover both the AMP and the EMS than OLS, indicating that GeMM had more power to detect these effects. The remaining models are less straightforward, but on balance GeMM showed recovery rates that were either approximately equal to (Robust LS-Huber, ordinal logistic) or better than the other alternatives.

Perhaps more instructive are the fit statistics provided in Table 3, which illustrate the average fit (top half) and average cross-validation accuracy (bottom half). GeMM provided better out-of-sample predictive accuracy than all of the alternatives in terms of  $\tau$  and even outperformed many of the alternatives in terms of the multiple  $R$ . Note that logistic regression showed the worst out-of-sample prediction in terms of  $R$  and second worst in terms of  $\tau$ , despite the fact that it showed the best performance in terms of  $\tau$  (and second best in terms of  $R$ ) on the estimation sample.

To summarize, on the basis of the statistical fit and predictive accuracy of the various statistical models, it is clear that the best and most defensible conclusion to draw from the data is that responses on the ATB scale in Siegel et al.'s (2012) study are best accounted for by both implicit racial prejudices (as measured by the Race-AMP) and external motivations to control prejudice (EMS), but not internal motivation to control prejudice (IMS). However, the bigger point to be made from these analyses is that statistical conclusions based on LS approaches proved to be highly suspect, a situation often due to a very small number of observations. Removing merely three of the 198 data points was sufficient to change the statistical conclusions, and the use of robust procedures only muddled the picture. The main problem, as we see it, is that the labile nature of LS procedures and their sensitivity to the removal or downweighting of extreme scores *licenses the researcher to choose which theory to support via the selection of a data-analytic strategy*. Thus, rather than the data constraining the theory, the theory can constrain the data in the name of making sure the data adhere to statistical

**Table 3.** Probability of Recovering Each Model Coefficient Given Half ( $N/2$ ) the Full Sample, for Each Algorithm

	P-EXP	Pol-IAT	Pol-AMP	Race-AMP	EMS	IMS	Race-IAT	Stroop	Stop Signal Task
GeMM	.062	.028	.014	<b>.336</b>	<b>.716</b>	.198	.02	.000	.008
OLS-NHST	.024	.016	.002	<b>.306</b>	<b>.596</b>	<b>.368</b>	.032	.006	.008
OLS-BIC	.068	.068	.008	<b>.402</b>	<b>.566</b>	<b>.486</b>	.036	.012	.004
RLS-bisquare	.028	.012	.002	<b>.458</b>	<b>.576</b>	.164	.014	.012	.010
RLS-Huber	.034	.016	.004	<b>.422</b>	<b>.642</b>	.214	.016	.012	.010
RLS-Hampel	.032	.012	.000	<b>.336</b>	<b>.572</b>	<b>.268</b>	.028	.010	.010
Ordinal logistic	.030	.020	.004	<b>.432</b>	<b>.636</b>	<b>.266</b>	.026	.012	.018

*Note:* Boldface values correspond to predictors that were recovered on the full sample as indicated in Table 2. BIC = Bayesian information criterion; EMS = External Motivation Scale; GeMM = general monotone model; IMS = Internal Motivation Scale; NHST = null hypothesis significance testing; OLS = ordinary least squares; P-EXP = explicit political attitude; Race-AMP = Political Affect Misattribution Procedure; Pol-IAT = Political Implicit Association Test; Race-AMP = Race Attitude Misattribution Procedure; Race-IAT = Racism Implicit Association Test; RLS = recursive least squares.



**Table 4.** Cross-validation Results for Analyses Predicting Attitudes toward Blacks

Cross-validation Using Selected (Best Fit) Models					
	$BIC'_\tau$	BIC	$\tau$	$R$	$k$
Estimation					
GeMM	-7.847	-4.881	.259	.314	1.397
OLS-BIC	-5.293	-7.321	.245	.363	1.647
Bisquare	-6.686	-5.787	.255	.337	1.542
Huber	-6.796	-6.041	.259	.344	1.604
Hampel	-6.864	-6.755	.255	.348	1.513
OLS	-6.209	-6.894	.251	.352	1.559
Ordered logit	120.919	114.465	.299	.354	28.042
Cross-validation					
GeMM			.155	.18	
OLS-BIC			.134	.178	
Bisquare			.145	.18	
Huber			.149	.183	
Hampel			.141	.169	
OLS			.147	.179	
Ordered logit			.137	.157	

*Note:* BIC = Bayesian information criterion; GeMM = general monotone model; OLS = ordinary least squares.

assumptions. GeMM appears to be more resistant to outliers, which means it will be less affected by decisions to eliminate them.

#### 4.2. *When Nonlinearities Mask Effects: The Case of Homicide Rates and the Culture of Honor*

A recent topic of interest in social-psychological research concerns cultures of honor, which are societies in which defense of reputation is a central organizing theme (Nisbett 1993; Nisbett and Cohen 1996). Such societies are especially common, according to Nisbett and colleagues, where scarce resources are highly portable (hence, easily stolen) and where the rule of law is weak or altogether absent (see also Brown and Osterman 2012). Nisbett (1993) argued that this combination is quite common in societies whose economies are based on herding rather than agriculture or industry. Because herding societies tend to be resource poor, their resources are quite portable, and they tend to be poorly managed by law enforcement, the latter due in part to the fact that herders

are, by necessity, spread out. Under such conditions, people are especially vulnerable to social predation, both from within (via internal competition for scarce resources) and from without (via attack from other groups). This vulnerability, over long periods of time, has a tendency to breed the beliefs, values, and social norms that characterize honor cultures, such as a hypervigilance to reputational threats and aggressive responses to perceived honor violations.

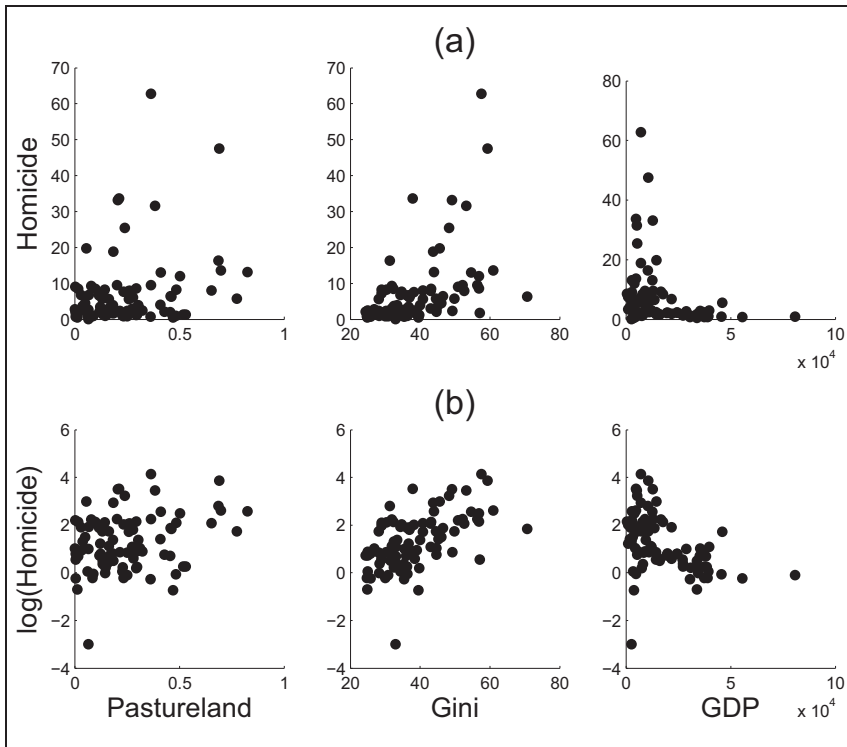
Honor cultures tend to stress strength and toughness as primary qualities of value for men, and loyalty and purity as primary qualities of value for women (Nisbett and Cohen 1996; Vandello and Cohen 2003). These qualities are pursued vigorously by men and women in such societies, as they help protect them from their key sources of vulnerability. For instance, men who are known to be strong and brave are not likely to be targeted for attack, as long as there are other targets available. Arguably, a man does not have to be *absolutely* strong and brave to protect himself or his family—he has to be known only as being *relatively* stronger and braver than other men, as someone who should not be disturbed or “messed with.” As long as he maintains his reputation for pugnacity, he can reduce the odds of predation from his neighbors and from hostile out-groups. Because of this combination of an extreme emphasis on reputation management and the types of reputations that are idealized for men and women, honor cultures tend to exhibit higher than average rates of argument-based homicides (Nisbett and Cohen 1996). In addition, research has shown that U.S. states classified as “honor states” (in the South and West) display higher levels of school violence (Brown, Osterman, and Barnes 2009), higher rates of suicide (Osterman and Brown 2011), and excessive levels of risk taking that lead to higher rates of accidental deaths (Barnes, Brown, and Tamborski 2012), compared with “nonhonor states.”

In a series of studies, Henry (2009) argued that one of the reasons that herding cultures tend to develop honor norms, as Nisbett and Cohen (1996) suggested they do, is that such cultures tend to be characterized by strong status disparities. When a society has a large status hierarchy, with relatively few people controlling a relatively large amount of that society’s resources, people at the bottom of the status hierarchy may feel especially vulnerable to social devaluation and be prone to hypervigilance and hyperreactance to status threats (see also Daly and Wilson 2010). Aggression in the face of insults is one prime example of the type of reaction that might be especially prevalent in members of low-status

groups in such unequal societies. Henry tested this notion in part by showing that homicide rates were higher in cultures whose economies tended to be based heavily on herding, where (theory suggests) honor-related beliefs and values will tend to proliferate. Important, Henry showed that elevated homicide rates in herding-oriented countries were statistically accounted for by levels of social *wealth disparity* within those countries, independent of a country's overall level of wealth. Henry also expected to replicate past findings that overall wealth would independently predict homicide rates, which he showed in study 1 (at the county level) but failed to show in study 2 (at the country level).

4.2.1. *Data and Analyses.* We reanalyzed the data used for study 2 of Henry (2009). Our use of this data set was a matter of convenience, and it was motivated by Henry's failure to replicate the association between overall wealth and homicide rates obtained in his study 1 and other prior work (Nisbett and Cohen 1996). Using OLS regression, we were able to reproduce his international results: Countries with larger proportions of their lands devoted to uncultivated pastures and meadows appropriate to herding (hereafter *pastureland*) tended to exhibit higher homicide rates, but this association was largely accounted for by within-country levels of wealth disparity (as indexed by the Gini coefficient of income inequality, hereafter *Gini*), independent of overall levels of wealth across those countries (as indexed by gross domestic product per capita, adjusted for purchasing power parity, hereafter *GDP*). Replicating Henry (2009), GDP was not a significant predictor ( $p = .36$ ), which remains as surprising to us as it did to Henry. However, a key question is raised: are our statistical conclusions robust?

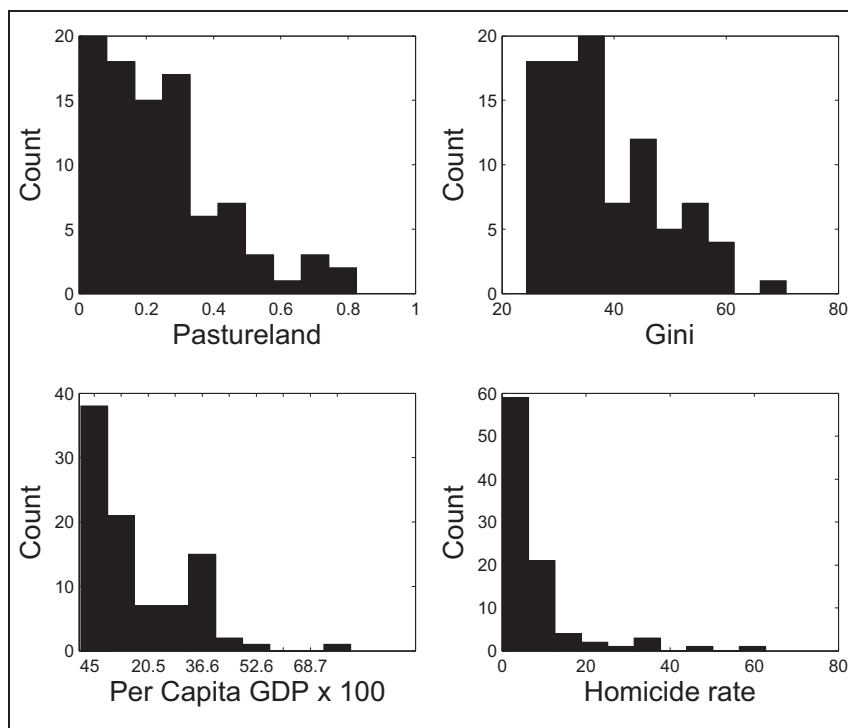
Figures 3a and 4 show the bivariate scattergrams and histograms for the four variables: homicide rates, percentage pastureland, Gini, and GDP. As is clear, the data are poorly distributed, yet there is obvious structure in the bivariate scattergrams. In particular, there appears to be a monotone but nonlinear relationship between GDP and homicide rates. Indeed, in terms of Kendall's  $\tau$ , the strength of the relationship between Gini and homicide ( $\tau = +.39$ ), is virtually identical to the strength of the relationship between GDP and homicide ( $\tau = -.36$ ). In contrast, the pattern of correlations obtained using Pearson's  $r$  yields a much stronger relation between Gini and homicide ( $r = +.50$ ) compared with GDP and homicide ( $r = -.30$ ). There appears to be not only substantial



**Figure 3.** Scattergram showing homicide rates per 100,000 residents ( $y$ -axis) plotted against three predictor variables used by Henry (2009). (a) Untransformed data. (b) Data after applying the log transformation to the homicide rate (per 100,000 residents).

nonlinearity in the data but also a small number of extreme scores and substantial nonnormality.

Given the obvious violations of assumptions for the linear model, it is likely that OLS regression is ill equipped to model these data accurately.<sup>7</sup> But, how *should* the data be modeled? Henry (2009) modeled homicide rates in their raw form using OLS, but other researchers interested in understanding factors contributing to homicide rates have used different approaches. For example, in testing the social capital theory of cross-national homicide rates, Lederman et al. (2002) modeled the natural logarithm of homicide. In a replication of this study, Robbins and Pettinicchio (2012) used negative binomial regression, which they argued more accurately captures the modeled distribution. Because the



**Figure 4.** Histograms for three predictor variables and the criterion used by Henry (2009).

data are transformations on count data (homicides per 100,000), both Poisson and quasi-Poisson regression are logical alternatives as well. The fact that there are multiple potential analysis techniques raises two questions: (1) Which method is “most” appropriate? and (2) Do the different methods yield different substantive conclusions? The question of which method is most appropriate is debatable, though addressing the second question seems straightforward. To start, we reanalyzed the data using two reasonable and common transformations: the natural logarithm and the square root. We used these transformations in two ways: first where only the criterion variable (homicide rate) was transformed and second where all of the variables were transformed. As an illustrative example, Figure 3b plots the bivariate scattergram after applying the log transformation to homicide rate. As can be seen, the nonlinear

**Table 5.** Fit Indices for Models

Transformation	BIC' <sub><math>\tau</math></sub>	BIC	$\tau$	R	k
OLS-NHST					
None	-29.332	-22.289	.391	.503	1
sqrt(homicide)	-41.419	-30.888	.471	.593	2
sqrt(all)	-29.332	-30.682	.391	.564	1
Log(homicide)	-42.672	-32.096	.476	.6	2
Log(all)	-29.332	-28.687	.391	.55	1
OLS-BIC					
None	-32.709	-22.289	.391	.503	1
sqrt(homicide)	-32.709	-31.062	.391	.566	1
sqrt(all)	-32.709	-30.682	.391	.564	1
Log(homicide)	-46.769	-32.129	.471	.601	2
Log(all)	-32.709	-28.687	.391	.55	1
GeMM					
None	-45.712	-13.636	.488	.467	2
sqrt(homicide)	-45.712	-27.531	.488	.573	2
sqrt(all)	-42.328	-26.307	.475	.565	2
Log(homicide)	-45.712	-30.677	.488	.592	2
Log(all)	-38.647	-25.71	.459	.561	2
GLM					
Poisson	-35.999	-12.302	.467	.495	3
Quasi-Poisson	-40.521	-16.824	.467	.495	2
Negative binomial	-40.298	-16.267	.466	.49	2

*Note:* In all cases in which  $k = 1$ , the predictor included in the model (or identified as significant) was Gini. In all cases in which  $k = 2$ , the predictors included in the model or identified as significant were both Gini and GDP. BIC = Bayesian information criterion; GDP = gross domestic product per capita, adjusted for purchasing power parity; GeMM = general monotone model; GLM = generalized linear model; NHST = null hypothesis significance testing; OLS = ordinary least squares.

relationships in the raw data are mostly linearized after the transformation.

Table 5 provides the results of the analyses using LS regression and GeMM both on the original (raw) data and on the transformed data. As should be evident, only GeMM provided a consistent model form across the various transformations. In particular, both versions of LS regression (OLS-NHST and OLS-BIC) recovered a one-predictor model consisting of Gini when applied to the raw data but a two-predictor model consisting of Gini and GDP when the criterion variable was log-transformed ( $p$  values  $< .001$  across methods for both Gini and GDP). When all of the variables were transformed, however, both OLS-NHST and OLS-BIC again recovered the single-predictor model consisting of Gini. The

square-root transformation also yielded inconsistent findings across methods. GeMM recovered a two-parameter model (Gini and GDP), and this was consistent across all of the transformations. Also included in Table 5 are the results from using three variants from the generalized linear model (GLM) family. Poisson regression identified all three predictors as significant, whereas both quasi-Poisson and negative binomial regression identified both Gini and GDP as significant.

Arguably, given the distributions presented in Figure 4, the data could legitimately be transformed to remove the skew prior to using traditional LS regression. However, whether the transformation should apply only to the criterion variable (homicide rate) or to all variables is a matter of debate and an existing “researcher degree of freedom” under traditional analysis methods. Although explicit transformations are unnecessary for negative binomial and the two Poisson regressions, they are implicitly carried out via the link function within GLM, of which researchers have many options. In contrast, with GeMM there is no need to transform the criterion variable because the rank-order correlation,  $\tau$ , is invariant to monotone transformation. Thus, whether the homicide rate is transformed by taking the logarithm, square root, or any other monotonic function or left untransformed is immaterial for GeMM’s solution and therefore removes this potentially important researcher degree of freedom.

The analyses presented above indicate that LS regression procedures are sensitive to decisions about whether (and how) the data are transformed. This should not be too surprising, because LS procedures fit distance information and because the distance information changes under different transformations. But just how distorted can it get? To explore this sensitivity, we analyzed the data again, but this time after adding a constant before applying the logarithmic transformation. The need to add a constant to the data prior to taking the logarithm arises when responses take on the value of 0 or are negative. Negative values are likewise problematic for the square-root transformation, but so are positive values less than 1 (as a square-root transformation on values between 0 and 1 will *increase* these values, while *decreasing* all values greater than 1; adding a constant to raise all raw values to a number greater than 1 eliminates this transformation disequilibrium). If OLS-NHST is used, adding any constant between .2 and 1.4 leads to both Gini and GDP identified as significant. Adding any constant above 1.4 or below .2 results in only Gini as statistically significant. The LS

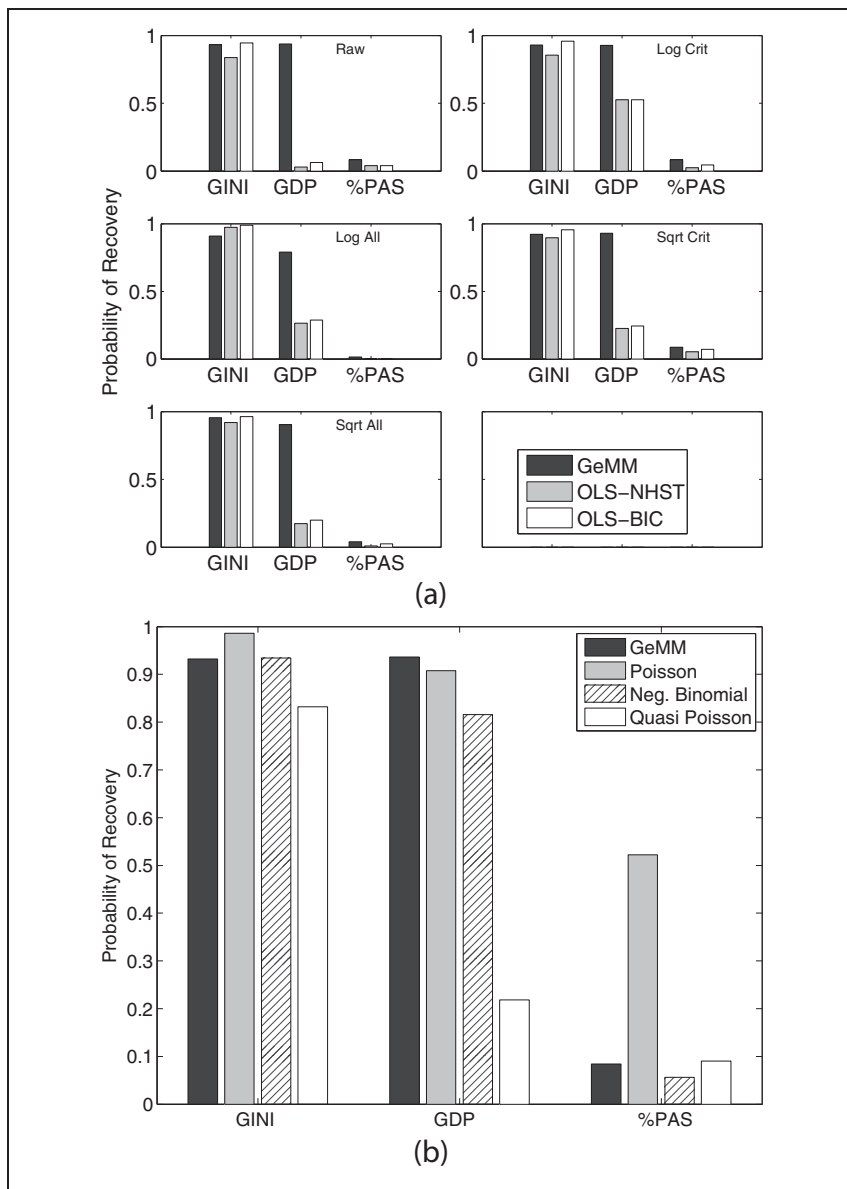
models yield different models with different additive constants; GeMM does not. The use of the negative binomial and Poisson regression models from the GLM family does not really solve the underlying issue: For these models, whether the number of homicides in each country is conceptualized as a count or a rate problem can actually change the form of the statistical model. Furthermore, if homicides are interpreted as a rate (number of homicides per unit of population), the model form can also depend on the choice of scaling constant. For example, both Pearson's  $r$  and rank-order correlation ( $\tau$ ) between the fitted values and the data vary depending on whether the homicide variable is expressed per 1,000, per 100,000, or per 10,000,000. How likely is it that researchers are aware of these sources of variation when they choose to add or divide by a constant as part of their data transformation routines?

As mentioned above, GeMM provides a two-predictor model regardless of transformation. But how well does this solution succeed when evaluated in terms of fit indices and cross-validation?

**4.2.2. Comparing Fit and Cross-validation.** A comparison of the relative fit indices favors the solution identified by GeMM. First, consider the results of the OLS. The one-predictor solution on the raw data has the highest value for  $R$  among the various procedures. However, despite having the best *metric* fit, this model is much poorer at capturing the ordinal properties of the data compared with GeMM and the GLMs. That is, to fit the ordinal properties of the data, it is necessary to give up a little accuracy in predicting the metric properties. Both the GeMM and the GLMs do just this. Comparing GeMM with the GLMs, however, also reveals that GeMM performs favorably in terms of  $\tau$  and  $BIC_{\tau}$ . GeMM's fit to the metric properties is somewhat poorer than that of the GLMs.

Using the same split-half methodology described in the discussion of the racial bias data, we evaluated the predictive accuracy and statistical power of GeMM relative to the various LS procedures. Figure 5a plots the probability of recovering each predictor using  $N/2$  for raw and transformed data for OLS-NHST, OLS-BIC, and GeMM. Table 6 provides the fit and out-of-sample predictive accuracy for all of the models. As is strikingly clear, GeMM recovers each of the two predictors (Gini and GDP) identified in the full sample on approximately 95% of runs, with statistical power remaining high across the various transformations. For comparison, for the full data set all three procedures recovered Gini and





**Figure 5.** Probability of recovering each predictor on half the total sample. (a) Comparison with least squares procedures across various transformations. (b) Comparison with various forms of generalized linear modeling. GDP = gross domestic product per capita, adjusted for purchasing power parity; GINI = wealth disparity index (Gini coefficient of income inequality); %PAS = percentage pastureland.

**Table 6.** Estimation and Cross-validation Analyses for the Culture of Honor Data Using Various Transformations

Transform	Model	BIC <sub>τ</sub>	BIC	τ	R	Mean k
<b>Estimation</b>						
None	GeMM	-18.626	-4.903	.495	.48	1.952
None	OLS	-12.554	-11.398	.4	.527	1.051
None	OLS-BIC	-12.344	-10.501	.397	.512	1.05
None	Poisson	-14.289	-8	.473	.54	2.416
None	Quasi-Poisson	-19.113	-13.078	.474	.546	1.197
None	Negative binomial	-17.556	-8.458	.481	.513	1.806
Sqrt(homicide)	GeMM	-18.616	-11.686	.495	.578	1.936
Sqrt(homicide)	OLS	-13.771	-15.292	.418	.585	1.202
Sqrt(homicide)	OLS-BIC	-13.736	-15.184	.421	.587	1.268
Sqrt(all)	GeMM	-17.356	-11.472	.482	.573	1.9
Sqrt(all)	OLS	-13.533	-14.895	.413	.578	1.129
Sqrt(all)	OLS-BIC	-13.463	-14.732	.415	.579	1.186
Log(homicide)	GeMM	-18.619	-14.427	.495	.603	1.94
Log(homicide)	OLS	-16.088	-16.689	.452	.608	1.474
Log(homicide)	OLS-BIC	-15.702	-16.321	.449	.605	1.528
Log(all)	GeMM	-15.673	-12.401	.459	.568	1.712
Log(all)	OLS	-13.541	-14.623	.418	.576	1.248
Log(all)	OLS-BIC	-13.566	-14.547	.419	.576	1.278
<b>Cross-validation</b>						
None	GeMM			.453	.463	
None	OLS			.379	.48	
None	OLS-BIC			.382	.487	
None	Poisson			.442	.428	
None	Quasi-Poisson			.373	.405	

(continued)

**Table 6.** (continued)

Transform	Model	BIC <sub>τ</sub>	BIC	τ	R	Mean k
None	Negative binomial					
Sqrt(homicide)	GeMM			.438	.436	
Sqrt(homicide)	OLS			.452	.554	
Sqrt(homicide)	OLS-BIC			.382	.532	
Sqrt(all)	GeMM			.393	.545	
Sqrt(all)	OLS			.44	.547	
Sqrt(all)	OLS-BIC			.385	.536	
Log(homicide)	GeMM			.391	.543	
Log(homicide)	OLS			.453	.567	
Log(homicide)	OLS-BIC			.402	.526	
Log(all)	GeMM			.416	.541	
Log(all)	OLS			.406	.516	
Log(all)	OLS-BIC			.392	.529	
				.395	.531	

*Note:* BIC = Bayesian information criterion; GeMM = general monotone model; OLS = ordinary least squares.

GDP when homicides were log transformed, but it is notable that GeMM substantially outperforms both versions of OLS in terms of recovering these predictors when provided half the data, in particular for GDP. Thus, not only does GeMM accurately recover GDP on the full data set, it does so with much higher power compared with OLS, even under conditions in which the data are transformed to make them more suitable for OLS.

Given the relatively poor showing of the LS procedures in recovering the predictors identified on the full data set, it should not be surprising that GeMM substantially outperformed its LS competitors in out-of-sample prediction. Indeed, even if we restrict our analyses to only the subset of nonnull models identified by the LS procedures, which we have done here in Table 6, it is clear that GeMM is the hands-down winner of the cross-validation contest. GeMM uniformly outperforms OLS and OLS-BIC in terms of predicting the rank order of homicide rates across nations, and in some cases even outpredicts OLS in terms of Pearson's  $R$ , for example,  $\sqrt{\text{homicides}}$ , and  $\log(\text{homicides})$ .

A comparison of GeMM with the GLMs is a bit more complicated. GeMM clearly outperforms quasi-Poisson in probability of recovery (Figure 5b) and out-of-sample predictive accuracy (Table 6). Standard Poisson regression recovered both Gini and GDP at the approximate level of GeMM, but it also recovered percentage pastureland more than 50% of the time. The inclusion of percentage pastureland in the model is particularly problematic here, because adding it to the model actually decreases ordinal predictive accuracy. Negative binomial regression performed nearly identical to GeMM across the board, with GeMM having a modest advantage in probability of recovery and a small (.015) advantage in terms of  $\tau$ . Thus, overall GeMM performed better than all three of the models from the GLM family.

To summarize, we argue that homicide rates across the 92 countries analyzed in this data set are best accounted for by both wealth disparity and a country's overall wealth per capita. Although the finding was tangential to Henry's (2009) main theoretical conclusions, it nevertheless explains a failure to replicate a classic finding in one of his studies, that of the relationship between GDP and murder rates. This finding is consistent with Henry's original prediction, which presumably was masked by the substantial nonlinearity present in the data. GeMM was able to accurately capture both Gini and GDP as important predictors of homicide rates without transformation and without requiring specific

assumptions about the form of the underlying distribution of homicides. In contrast, within OLS, the decision to include GDP in the statistical model was contingent on how the data were transformed, and within GLM, it was contingent on which distribution was assumed.

## 5. GENERAL DISCUSSION

The analyses presented in this article identified two important problems faced by behavioral and social scientists in their use of standard and robust LS procedures and a possible solution to these problems. First, LS regression procedures are highly sensitive to violations of assumptions and the presence of extreme scores. In our reanalysis of the racial bias data, we illustrated that a small number of extreme scores was sufficient to drive or mask statistical effects. Eliminating a mere 1.5% of the data was sufficient to render internal motivation to control prejudice as unnecessary to predict explicit attitudes toward blacks. In contrast, for the culture of honor data, violations of the linearity assumption and/or the presence of extreme scores resulted in the failure of LS regression to identify expected patterns for which there was structure in the data. Taken together, these results suggest that nuances within one's data can either drive effects or mask them when using LS procedures. The fact that violations of assumptions and messy (which is to say, real) data can undermine statistical conclusions is not a new insight, of course. What is new, we believe, is that accepted procedures for dealing with messy data offer no real solutions to the problem, which leads to the second finding.

The second finding identified by our analyses is that accepted methods for dealing with messy data do not uniformly converge on a consistent statistical, and therefore theoretical, conclusion. This is especially problematic because the failure to find consistency across methods leaves too much decision-making power in the hands of the scientist. Unfortunately, scientists are not always unbiased observers of their data, and they are probably most likely to use the data-editing strategies that result in outcomes supportive of their theories, although they might not be aware that they are doing so. Thus, standard practices for dealing with messy data increase the number of researcher degrees of freedom (cf. Simmons et al. 2011), which we argue can undermine the search for valid scientific conclusions and hamper scientific progress.

Our solution to these two problems is to advocate for statistical procedures that reduce or eliminate the need for conducting outlier analyses and data transformations. As we showed throughout this article, GeMM provides a promising new approach that maximizes fit at the ordinal level. To illustrate the fundamental importance of modeling the ordinal level of data, imagine that a new scoring system were proposed for use at the Olympic Games. This scoring system, statisticians show, does a good job of accounting for variance in athletes' past scores (analogous to a high  $R^2$ ), although it does not do particularly well at recovering ordinal outcomes, in other words, in postdicting who came in first, second, or third place. We cannot imagine that such a scoring system would ever see the light of day, and whoever proposed it would be laughed out of a career in statistics. Nonetheless, that is essentially what the present two studies suggest is happening with LS procedures when it comes to modeling ordinary, messy data in the behavioral sciences. As we have shown, GeMM's solution was relatively more robust across a variety of reasonable methods for identifying and eliminating extreme and influential scores. This is a major advantage of GeMM, as it removes some of the degrees of freedom that researchers have to make the results "turn out" in favor of their hypotheses (Simmons et al. 2011).

As a side note, it is interesting to comment on what constitutes an outlier in the traditional sense. Outliers are typically identified by their distance from the center point of a distribution of scores, or how much influence they have on the fit of a regression model. Measures of influence, such as Cook's  $D$  and DFFITS, are defined within a LS function and provide a metric for how influential a particular data point is on the overall LS fit of a model. Thus, the more extreme an observation is, the more influence it exerts on the LS solution. In contrast, within the GeMM framework, a score that is 3 standard deviations from the mean is treated as no different than a score that is 100 standard deviations from the mean. Indeed, the only influence an extreme score has on the overall fit of the model is gauged by how many inversions it creates in the predicted rank orders when included in the data set. This implies a need for influence statistics that operate in ordinal, rather than metric, space. Because GeMM models data on an ordinal level, it has a higher bar in terms of what constitutes an outlier.

Reconceptualizing data through the lens of ordinality redefines the meaning of outliers as those observations that have undue influence on the rank-order fit of the model. These observations may be true

aberrations—data points that represent illegitimate responses given the measurement instruments (e.g., a response of 12, when the scale is bounded at 10)—or they may be real observations. For example, there are many cases in which extreme scores might be produced by data-entry errors, distracted subjects, or other processes external to an experiment. However, in the great majority of cases there is no ground truth by which researchers can determine whether an extreme score is a legitimate member of the population distribution or an aberration due to an external factor. The uncertainty surrounding the cause of an extreme score is problematic for justifying its exclusion. If the decision to exclude is based on the need to meet the assumptions of a statistical algorithm, this strikes us as a poor justification and is tantamount to forcing a round peg into a square hole.

Obviously, there are a number of alternative regression procedures not included in our modeling competition, and one might take issue with our focus on LS regression. However, we believe that this focus is warranted given the widespread use of the OLS (and its robust implementations) across the social sciences. Still it is quite possible that other models might perform better than GeMM, though the appropriate candidates for the two data sets presented here (ordinal logistic, negative binomial, Poisson and quasi-Poisson regressions) did not offer any performance advantages over GeMM, and in most cases underperformed relative to GeMM.

At the same time, one might argue that decisions regarding whether to transform one's data should be based on sound justification and the need to do so prior to engaging in data analysis. We agree, of course, but also argue that transformation for the purpose of analyzing a particular data set seems potentially opportunistic. Hence, we suggest that decisions to transform a data set in a particular way should be based on an understanding of the population distribution and driven by theory, not based merely on characteristics of the sample distribution. In the absence of theoretically justified reasons for transformation, we suggest that procedures such as GeMM are more appropriate for handling data where there are even slight departures from linearity, except where the form of the nonlinearity is of theoretical interest.<sup>8</sup>

Substantively, the findings based on GeMM for the racial bias and culture of honor data were at odds with what were found using traditional LS approaches. First, analysis of the race data suggests that responses on the ATB scale are a function of two variables: an

unconscious racial attitude, as measured subtly by the AMP, and an external motivation to control prejudice. The AMP was positively predictive of people's responses on the ATB scale, whereas external motivation to control prejudice was negatively related to people's responses on the ATB. This pattern supports the idea that individuals are motivated to conceal their racial attitudes because they know that racial prejudice is socially unacceptable.

The fact that internal motivation to control prejudice was not included in the GeMM contradicts the conclusions drawn by Plant and Devine (1998) and more recent findings of Payne et al. (2005). There are many possible reasons that our findings are at odds with these prior studies, including the fact that racial attitudes likely differ across geographical regions (i.e., attitudes toward blacks may differ across different subject populations) and change over time (i.e., the data collected by Plant and Devine are at least 15 years old). We therefore do not question the validity of these prior findings. Rather, the critical point for the present purposes is that *the statistical, and therefore theoretical, conclusions drawn from our data were heavily dependent on decisions about how to deal with its messiness.*

Second, for the culture-of-honor data, we showed that homicide rates are predicted by both wealth disparity (Gini) and overall country wealth (GDP). Wealthier countries experience fewer homicides, whereas countries with greater *wealth disparity* experience *more* homicides. These variables are theoretically independent of one another, as a country could be poor but exhibit complete social equality in its distribution of its few resources (not likely, but theoretically possible), or a country could be wealthy and exhibit a similar degree of social equality. Indeed, developed nations with high GDPs per capita differ widely in terms of how their overall wealth is distributed across their people. This potential independence of GDP and Gini, however, is largely theoretical, as overall wealth and wealth disparity are, in fact, negatively correlated in analyses at the level of nations, states, and even counties within states (e.g., Henry 2009). In poorer countries, resources are more likely to be controlled by a few powerful people, compared with the more abundant resources of wealthier countries. Because of this typical association, researchers studying wealth or wealth disparities must consider both of these variables if they want to avoid confounding one with the other.

According to the analyses presented here, *how* a researcher decides to handle messy data can have an enormous impact on whether or to



what extent variables (e.g., GDP, internal motivations to control prejudice) reveal their influences. Because of both nonlinear patterns and the influence of extreme scores, traditional LS analyses will sometimes overestimate a variable's influence, as in the case of internal motivations to control prejudice as a predictor of racial attitudes. Traditional LS analyses can also *underestimate* a variable's influence, as is the case in the association between a country's wealth and homicide rates, due to nonlinear relations and extreme scores in the data.

### 5.1. *What Are the Practical Advantages of GeMM?*

These substantive issues aside, what might compel one to use GeMM in lieu of traditional LS regression? As with other regression techniques, GeMM is a tool for prediction, inference, and data mining and exploration, though we believe that it offers some practical advantages over standard LS techniques. We articulate these next.

5.1.1. *GeMM as a Tool for Prediction.* As demonstrated with the two data sets presented in this article, GeMM provides a computational algorithm for optimizing rank-order prediction that can outperform more complex algorithms on the basis of LS. The trade-off, of course, is that GeMM is not guaranteed and likely will not optimize prediction of metric values. However, we believe that this trade-off is warranted in many contexts. For example, consider any task that entails a selection decision on the criterion or outcome variable, such as selecting among job applications, choosing graduate applicants (if you are a faculty member), or choosing graduate programs (if you are a student). In all of these cases, the goal of the decision maker is to predict the relative ordering on the criterion, rather than to predict a specific quantitative value. As should be clear from the two example data sets presented here, GeMM generally showed greater accuracy for out-of-sample prediction when assessed in terms of predicting the ordinal values. Inasmuch as one of the principal goals of the social and health sciences is to predict real-world behaviors, having statistical models that can, first and foremost, accurately predict ordered relations is important: what good is a statistical model with a high  $R^2$  value if it does poorly in predicting the relative ordering of the criterion variable?

5.1.2. *GeMM as a Tool for Inference.* In an ideal world, inferences drawn from data should be invariant across data-editing strategies. The

problem, of course, is that there is theoretically an infinite number of ways in which data can be transformed, and numerous justified ways of identifying outliers. Although it is certainly possible to explore a variety of potential data-editing strategies to assess the robustness of the conclusions, it would be virtually impossible to explore all possible transformations and outlier deletion methods. In this respect, GeMM offers many practical advantages over standard techniques: It is (1) invariant to transformation on the criterion variable, (2) more robust to transformation on the predictors, and (3) more robust to outliers. These advantages follow from the use of tau as the fit metric, which, unlike Pearson's  $r$ , is invariant to monotone transformation. Because transformation on the predictors can affect the additive form of the predicted values, GeMM can still be affected by transforming the predictors, but only if the transformation results in changes in the ordinal properties of the additive model. In contrast, the use of transformation on the predictors is *guaranteed* to affect the LS fit. In other words, many of the decisions that could be exploited for analysis on the basis of LS approaches are unnecessary for analyses based on GeMM. Furthermore, unlike linear LS, GeMM does not lose statistical power under deviations from linearity.

As an example, consider our analysis of the culture-of-honor data. In this analysis, we illustrated that GeMM was relatively insensitive to transformation and had higher statistical power than linear LS. Thus, making fewer assumptions about one's data can pay off in an increased likelihood of detecting effects and more robust conclusions that are not conditional on having met specific model assumptions or on particular data-editing strategies. Importantly, the conditions in which researchers are most inclined to engage in data editing are precisely those conditions in which the data are unlikely to satisfy metric statistical assumptions.

On the flip side, GeMM's strength as a method for identifying monotone relationships limits the specificity of the inferences that can be drawn from the data. Although it can identify any nonlinear monotone relationship with equal probability without the need to transform the data, it cannot characterize the nature of those relationships. Thus, if researchers are interested in modeling the specific functional relationship between a set of variables, then GeMM would not be an appropriate tool. It should be noted though that the application of GeMM does not preclude them from further exploring these functional relationships

with nonlinear LS methods, if they are comfortable drawing conclusions that go beyond the ordinal properties.

*5.1.3. GeMM as a Tool for Exploration.* As with traditional LS methods, GeMM can also be used in the context of data exploration. Note, however, that in this context the fact that GeMM relaxes assumptions about functional form can be advantageous. Consider, for example, a data set in which a researcher has no *a priori* hypotheses about which variables should be related to the criterion. In these cases, it is even less likely that the researcher has any *a priori* guess about the form of the functional relationships that might exist therein. The problem with using traditional LS regression approaches in these contexts is that they require either that the researcher commit to modeling specific functional forms, engage in a great deal of data editing, or explore various alternative modeling approaches. With GeMM, identifying potentially interesting statistical relations can be accomplished with minimal data editing and without loss of power when those relations are nonlinear.

## *5.2. Interpreting the Output of Regression Coefficients within GeMM*

The most straightforward interpretation of GeMM is in its model form, wherein the GeMM returns the model that best accounts for the rank-ordered properties of the criterion. The regression coefficients derived from GeMM have the exact same interpretation as those obtained from OLS once the OCLO solution is obtained, with one caveat. The OLS solution minimizes LS, whereas the rescaled OCLO-GeMM weights minimize LS conditional on maximizing ordinal fit.

Although in many cases the actual parameter values derived from GeMM may be close in magnitude to those obtained from other statistical procedures, there may be cases in which the relative magnitudes of the parameters differ in important ways. For example, for the homicide data set, the standardized regression coefficients derived from OLS yielded  $|B_{\text{Gini}}| > |B_{\text{GDP}}|$  (.42 vs. -.09), but the GeMM solution yielded  $|B_{\text{Gini}}| < |B_{\text{GDP}}|$  (.25 vs. -.29). This is informative because it tells us that the relative contributions of GDP and Gini are different if we are interested in using these variables to predict the rank order of homicide rates (GeMM) versus predicting the metric values of homicide rates (OLS). The implications of the GeMM solution compared with the OLS

solution could be rather important. For example, a policymaker who wishes to reduce homicide rates would make different policy decisions if using OLS as the basis of that decision than if GeMM were used as the basis of that decision: the OLS solution implies that efforts at reducing homicide rates should focus primarily on decreasing wealth disparity (Gini), whereas the GeMM solution implies both that wealth disparity should be decreased and overall wealth (GDP) increased. This is not to suggest that GDP or wealth disparity cause homicides but rather to highlight the two very different policies that could result from using OLS versus GeMM.

### 5.3. Availability and Extensions

The bulk of this article has focused on the application of GeMM in contexts in which we must deal with messy data in one way or another. To facilitate the use of GeMM, we have developed versions in MATLAB, Mathematica, SAS, and R. MATLAB code and an accompanying user's guide are available at the first author's Web site (<http://www.damlab.umd.edu/gemm.html>); Mathematica and SAS code is available upon request. The development version of the GeMM package for R, and associated code and data used in this article, are available for free from the authors. The R package will be posted to Cran when completed. In its present form, the R package automatically produces the OCLO solution proposed in Tidwell et al. (2014).

We have a number of active lines of work aimed at extending the GeMM framework. A key limitation of GeMM thus far is that it is constrained to modeling monotonic relationships and therefore is not applicable to data sets that include nonmonotonic relationships. To address this, we have begun developing a version of GeMM that permits inflection points between the criterion and the modeled data, where an inflection point implies a change in the direction (sign) of the modeled relationship (Lawrence, Thomas, and Dougherty 2014).

A second area of work motivated by GeMM involves the development of leverage or influence statistics that identify outliers in ordinal space. Although GeMM should in principle be more robust to many different types of extreme scores, it will still be sensitive to extreme scores that create a large number of rank-order inversions. This is likely the reason that GeMM showed some sensitivity to the outlier deletion in racial bias data set. Although these types of extreme scores might be

identifiable with traditional leverage statistics such as Cook's  $D$ , we imagine that alternative methods for identifying highly influential scores in ordinal space will be required.

## 6. SUMMARY

The existence of "uncooperative" and messy data poses a major challenge for behavioral and social science researchers. Unfortunately, within the standard approaches, traditional methods for handling nonlinearities, nonnormalities, and outliers provide the data analyst with a great deal of freedom for reconditioning the data to remove these properties, a freedom that can be exploited, intentionally or otherwise, to tell the preferred story. The more freedom allotted to the data analyst to make decisions that are not well justified, the more likely it is that the stories that get told are little more than myths. The goal of discovering fundamental facts about nature should not lead us to treat data and data analysis as if it were fine art requiring delicate hands. Rather, it should compel us to approach data analysis the way an engineer approaches the development of a new jetliner, which is to ensure that the plane flies even under nonideal conditions. As a public good that informs social and health policy, we argue that the same standard should operate for scientific claims. GeMM provides a new tool that we believe can help ensure that scientific claims are robust and invariant to data-editing strategies.

### Notes

1. We use the term *criterion variable* to refer to the outcome or dependent variable.
2. The assumption of bivariate normality is not crucial for the operation of GeMM. One way to conceptualize the  $\tau$ -to- $r$  transformation is that it allows one to estimate the value of  $r$  under any order-preserving transformation of the data, without actually needing to transform the data. When assumptions of bivariate normality and linearity are met, then the  $\tau$ -to- $r$  transformation should closely approximate the value of  $r$  on the untransformed data.
3. Siegel et al. (2012) used structural equation modeling to examine the factor structure of the various measures of attitude and cognitive ability. For that analysis, the absolute (unsigned) scores were used.
4. The univariate outliers were identified by observations  $\pm 3$  standard deviations from the mean. Cook's  $D$  and DFFITS are standard leverage statistics that quantify the influence of each individual point on the regression solution. Observations were trimmed from the data set if the value of Cook's  $D$  exceeded  $4/N$  and if the value of

- DFFITS exceeded  $2[\text{sqrt}(p/N)]$ , where  $p$  is the number of predictors in the regression.
5. On the basis of the full sample, there are 37 distinct response categories, for which ordered logistic regression must fit 36 threshold parameters. For the full sample, only 32 of these thresholds were statistically significant at  $p < .05$ .
  6. Model fitting for GeMM consisted of a two-step process in which we first fit GeMM to the full sample to find the subset of predictors that minimized  $\text{BIC}_\tau'$ . We then ran 1,000 bootstrap samples to estimate the standard errors of the coefficients. The coefficients listed in Table 2 correspond to the mean coefficients (and corresponding standard errors) from the 1,000 bootstrap samples. Model fits listed in Table 1 are based on the analysis of the full sample.
  7. Both the Henze-Zirkler and Mardia tests of multivariate normality revealed significant departures from multivariate normality, a finding that held for both the untransformed and transformed data.
  8. However, we suggest that in most cases in the social sciences, theories are not specified in such detail and instead are expressed largely as ordinal predictions (see also Cliff 1996).

## References

- Barnes, Collin D., Ryan P. Brown, and Michael Tamborski. 2012. "Living Dangerously: Culture of Honor, Risk-Taking, and the Nonrandomness of 'Accidental' Deaths." *Social Psychological and Personality Science* 3(1):100–107.
- Beck, Nathaniel, and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42(2):596–627.
- Brown, Ryan P., Lindsey L. Osterman, and Collin D. Barnes. 2009. "School Violence and the Culture of Honor." *Psychological Science* 20(11):1400–405.
- Brown, Ryan P., and Lindsey L. Osterman. 2012. "Culture of Honor, Violence, and Homicide." Pp. 218–32 in *Oxford Handbook of Evolutionary Perspectives on Violence, Homicide, and War*, edited by T. Shackelford and V. W. Shackelford. New York: Oxford University Press.
- Cavanagh, Christopher, and Robert P. Sherman. 1998. "Rank Estimators for Monotonic Index Models." *Journal of Econometrics* 84(2):351–81.
- Cliff, Norman. 1996. *Ordinal Methods for Behavioral Data Analysis*. New York: Psychology Press.
- Crocker, Jennifer. 2011. "The Road to Fraud Starts with a Single Step." *Nature* 479(7372):151.
- Daly, Martin, and Margo Wilson. 2010. "Cultural Inertia, Economic Incentives, and the Persistence of "Southern Violence." Pp. 229–41 in *Evolution, Culture, and the Human Mind*, edited by Mark Schaller, Ara Norenzayan, Steven J. Heine, Toshio Yamagishi, and Tatsuya Kameda. New York: Taylor & Francis.
- Devine, Patricia G., E. Ashby Plant, David M. Amodio, Eddie Harmon-Jones, and Stephanie L. Vance. 2002. "The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond without Prejudice." *Journal of Personality and Social Psychology* 82(5):835.

- Dougherty, Michael R., and Rick P. Thomas. 2012. "Robust Decision Making in a Nonlinear World." *Psychological Review* 119(2):321.
- Dunton, Bridget C., and Russell H. Fazio. 1997. "An Individual Difference Measure of Motivation to Control Prejudiced Reactions." *Personality and Social Psychology Bulletin* 23(3):316–26.
- Fang, Ferric C., R. Grant Steen, and Arturo Casadevall. 2012. "Misconduct Accounts for the Majority of Retracted Scientific Publications." *Proceedings of the National Academy of Sciences of the United States of America* 109(42):17028–33.
- Fazio, Russel H., Joni R. Jackson, Bridget C. Dunton, and Carol J. Williams. 1995. "Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline?" *Journal of Personality and Social Psychology* 69(6):1013–27.
- Freese, Jeremy. 2007. "Reproducibility Standards in Quantitative Social Science: Why Not Sociology?" *Sociological Methods and Research* 36:153–72
- Gerber, Alan S., and Neil Malhotra. 2008. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods and Research* 37(1):3–30.
- Greenwald, Anthony G., Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74(6):1464–80.
- Han, Aaron K. 1987. "Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator." *Journal of Econometrics* 35(2–3): 303–16.
- Hauser, Robert M. 1987. "Sharing Data: It's Time for ASA Journals to Follow the Folkways of a Scientific Sociology." *American Sociological Review* 52(6):vi–vii.
- Hays, William. 1994. *Statistics*. 5th ed. Belmont, CA: Wadsworth.
- Henry, P. J. 2009. "Low-status Compensation: A Theory for Understanding the Role of Status in Cultures of Honor." *Journal of Personality and Social Psychology* 97(3): 451–66.
- Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199–236.
- Howell, David C. 2002. *Statistical Methods for Psychology*. 5th ed. Pacific Grove, CA: Duxbury/Thomson Learning.
- Jasso, Guillermina. 1986. "Is It Outlier Deletion or Is It Sample Truncation? Notes on Science and Sexuality." *American Sociological Review* 51(5):738–42.
- John, Leslie K., George Loewenstein, and Drazen Prelec. 2012. "Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling." *Psychological Science* 23(5):524–32.
- Kahn, Joan R., and J. Richard Udry. 1986. "Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions." *American Sociological Review* 51(5):734–37.
- Kendall, Maurice. 1938. "A New Measure of Rank Correlation." *Biometrika* 30(1/2): 81–93.
- Kendall, Maurice, and Jean D. Gibbons. 1990. *Rank Correlation Methods*. 5th ed. New York: Oxford University Press.



- King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28(3): 444–52.
- Lawrence, Ashley, Rick P. Thomas, and Michael R. Dougherty. 2014. "A Non-monotonic Approach to Ordinal Prediction." Unpublished manuscript.
- Leahey, Erin. 2005. "Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology." *Social Forces* 84(1):1–24.
- Leahey, Erin. 2008. "Overseeing Research Practice: The Case of Data Editing." *Science, Technology, and Human Values* 33(5):605–30.
- Leahey, Erin, Barbara Entwisle, and Peter Einaudi. 2003. "Diversity in Everyday Research Practice: The Case of Data Editing." *Sociological Methods and Research* 32(1):64–89.
- Lederman, Daniel, Norman Loayza, and Ana Maria Menéndez. 2002. "Violent Crime: Does Social Capital Matter?" *Economic Development and Cultural Change* 50(3): 509–39.
- Levelt Committee, Noort Committee, and Drenth Committee. 2012. *Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel*. Commissioned by the Tilburg University, University of Amsterdam, and the University of Groningen.
- Micceri, Theodore. 1989. "The Unicorn, the Normal Curve, and Other Improbable Creatures." *Psychological Bulletin* 105(1):156.
- Nisbett, Jon, and Dov Cohen. 1996. *Psychology of Violence in the South*. Boulder, CO: Westview.
- Nisbett, Richard E. 1993. "Violence and U.S. Regional Culture." *American Psychologist* 48(4):441–49.
- Olsen, Karen M., and Sverre Dahl. 2007. "Health Differences between European Countries." *Social Science and Medicine* 64(8):1665–78.
- Osterman, Lindsey L., and Ryan P. Brown. 2011. "Culture of Honor and Violence Against the Self." *Personality and Social Psychology Bulletin* 37(12):1611–23.
- Payne, B. Keith, Melissa A. Burkley, and Mark B. Stokes. 2008. "Why Do Implicit and Explicit Attitude Tests Diverge? The Role of Structural Fit." *Journal of Personality and Social Psychology* 94(1):16.
- Payne, B. Keith, Clara Michelle Cheng, Olesya Govorun, and Brandon D. Stewart. 2005. "An Inkblot for Attitudes: Affect Misattribution as Implicit Measurement." *Journal of Personality and Social Psychology* 89(3):277.
- Plant, E. A., Patricia G. Devine, and Paige C. Brazy. 2003. "The Bogus Pipeline and Motivations to Respond without Prejudice: Revisiting the Fading and Faking of Racial Prejudice." *Group Processes and Intergroup Relations* 6(2):187–200.
- Plant, E. Ashby, and Patricia G. Devine. 1998. "Internal and External Motivation to Respond without Prejudice." *Journal of Personality and Social Psychology* 75(3):811.
- Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." Pp. 111–63 in *Sociological Methodology*, Vol. 25, edited by Peter V. Marsden. Cambridge, MA: Blackwell.
- Robbins, Blaine, and David Pettinicchio. 2012. "Social Capital, Economic Development, and Homicide: A Cross-national Investigation." *Social Indicators Research* 105(3):519–40.



- Rupinski, Melvin T., and William P. Dunlap. 1996. "Approximating Pearson Product-moment Correlations from Kendall's Tau and Spearman's Rho." *Educational and Psychological Measurement* 56(3):419–29.
- Sana, Mariono, and Alexander A. Weinreb. 2008. "Insiders, Outsiders, and the Editing of Inconsistent Survey Data." *Sociological Methods Research* 36(4):515–54.
- Semyonov, Moshe, and Noah Lewin-Epstein. 2011. "Wealth Inequality: Ethnic Disparities in Israeli Society." *Social Forces* 89(3):935–59.
- Shepard, Roger N. 1962. "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. Part I." *Psychometrika* 27:125–40.
- Shepard, Roger. 1966. "Metric Structures in Ordinal Data." *Journal of Mathematical Psychology* 3:287–315.
- Siegel, Eric F., Michael R. Dougherty, and David E. Huber. 2012. "Manipulating the Role of Cognitive Control While Taking the Implicit Association Test." *Journal of Experimental Social Psychology* 48(5):1057–68.
- Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11):1359–66.
- Simonsohn, Uri. 2013. "Just Post It: The Lesson from Two Cases of Fabricated Data Detected by Statistics Alone." *Psychological Science* 24(10):1875–88.
- Tidwell, Joe W., Michael R. Dougherty, Jeffery S. Chrabaszcz, and Rick P. Thomas. 2014. "Order Constrained Linear Optimization." Unpublished manuscript.
- Vandello, Joseph A., and Dov Cohen. 2003. "Male Honor and Female Fidelity: Implicit Cultural Scripts that Perpetuate Domestic Violence." *Journal of Personality and Social Psychology* 84(5):997–1010.

### Author Biographies

**Michael R. Dougherty** is a professor of psychology at the University of Maryland, College Park. His areas of interest include research methods, computational and mathematical modeling, cognitive decision theory, and memory theory.

**Rick P. Thomas** is an associate professor of psychology at the Georgia Institute of Technology. His areas of interest include computational and mathematical modeling, cognitive decision theory, and engineering psychology.

**Ryan P. Brown** is a professor of psychology at the University of Oklahoma. His area of research includes understanding factors contributing to honor cultures and the impact honor cultures have on outcomes ranging from school violence to terrorism.

**Jeffrey S. Chrabaszcz** is a PhD student at the University of Maryland. His interests include research methods, computational and mathematical modeling, judgment and decision making, and anxiety.

**Joe W. Tidwell** is a PhD student at the University of Maryland. His interests include research methods, computational and mathematical modeling, judgment and decision making, and forecasting.