

Determining the Internal Consistency of Attitude Attributions

Kyle E. Jennings (jennings@berkeley.edu)

Department of Psychology, University of California, Berkeley
Berkeley, CA 94720-1650 USA

Abstract

In the attitude attribution paradigm, observers must estimate the true attitude of an author who was assigned to advocate a particular position. Observers' tendency to attribute an attitude in line with the expressed position despite its having been assigned is called the correspondence bias. While there is strong evidence that such attributions are externally invalid, it is less clear whether they are internally consistent. This research develops a Bayesian model that specifies what attitude an observer should attribute, given assumptions about the prior attitude distribution, and perceptions of the degree of compliance shown in the essay and the strength of the situation. The model reproduces classical findings regarding chosen vs. assigned positions, prior attitude probability, and degree of compliance, and also fits newly collected data. The results suggest that future research should examine observers' assumptions and perceptions, and focus less on the reasoning process itself.

Keywords: Correspondence bias; Attitude attribution; Normative standard; Bayesian modeling.

People's tendency to neglect situational influences on behavior has been a subject of long-standing interest to social psychologists. Many of the earliest and most famous demonstrations of this error make use of the attitude attribution paradigm (Jones & Harris, 1967), wherein participants read an essay that expresses an opinion on an issue, and must estimate the author's attitude. Complicating this judgment is the fact that the author was assigned what position to express, which pits two competing explanations—holding the attitude, or complying with the request—against each other. Participants tend to make attitude attributions in line with the essay even when the position was assigned, which is called the correspondence bias (Gilbert & Jones, 1986).

It is not straightforward to say whether people's responses in the attitude attribution paradigm are in fact biased. On the one hand, when participants rate essays that other study participants wrote under constraint, the attributed attitudes are more in line with the essay than with the authors' self-reported attitudes (e.g., Reeder, Fletcher, & Furman, 1989). On the other hand, if people's attributions are internally consistent with their own perceptions and assumptions, it is hard to call their attributions completely biased (Jones, Worchel, Goethals, & Grumet, 1971; Morris & Larrick, 1995; Forsyth, 2004). These two views involve two different standards for correctness, known as correspondence and coherence (Hammond, 1996), which concern the external validity and internal consistency of the judgment, respectively. To avoid confusion between correspondence *criteria* and the correspondence *bias*, the terms external validity and internal consistency will be used. Though people's judgments probably lack external validity, it is not clear whether they are at least internally consistent.

Checking internal consistency requires knowing what information is relevant to a judgment, and how that information determines the correct answer. This paper develops a Bayesian model relating assumptions and perceptions to attitude attributions. Since the model is grounded in mathematics, the steps between premises and conclusions can be more readily verified than with verbally justified standards (Morris & Larrick, 1995). Additionally, the model is agnostic to what process people might use to make judgments, helping researchers advocating different mechanisms at least agree on the correct outcome.

Normative Model

In the attitude attribution paradigm, observers know what essay was written, and the circumstances under which it was written. Their judgment of whether the essay author holds the expressed attitude is (Morris & Larrick, 1995):

$$P(\text{attitude} \mid \text{essay, circumstances})$$

Letting A , E , and C stand for the attitude, essay, and circumstances, and strategically applying Bayes' rule,¹ this equals:

$$P(A) \cdot \frac{P(C \mid A)}{P(C)} \cdot \frac{P(E \mid A, C)}{P(E \mid C)}$$

Intuitively, these terms express the prior probability of the attitude, the co-occurrence of the circumstances and the attitude, and the relative likelihood of a person writing the essay, comparing someone with the attitude to the average person.

The model can be applied in two ways. First, it can be interpreted schematically in order to draw conclusions about the general direction of normative inferences. For instance, the standard shows that the conventional wisdom that the essay communicates no information about the author's attitude in light of the circumstances is correct only if two conditions are met. First, the co-occurrence term must be one, meaning that positions must be assigned without respect to the author's attitude. Studies that merely say that the position to advocate was assigned leave open the possibility that the author's attitude was considered when making the assignment, in which case a correspondent inference may be justifiable. Second, the likelihood term must also be one, meaning that the constraint must be seen as equally compelling regardless of the author's attitude (with completely compelling being a special case of this). As other researchers have argued (e.g., Jones

¹ $P(A \mid E, C) = P(A, E, C) / P(E, C) = P(E \mid A, C)P(A, C) / P(E, C) = P(E \mid A, C)P(C \mid A)P(A) / [P(E \mid C)P(C)]$. See also Jennings (2010).

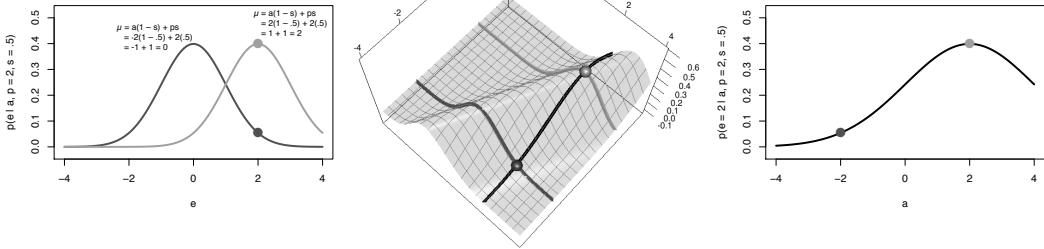


Figure 1: Illustration of $p(e | a, p = 2, s = .5)$. Left panel shows two essay distributions, for $a = -2$ and $a = 2$. Right panel shows likelihood distribution over attitudes for $e = 2$. As shown in the middle panel, these two distributions are really slices of the same three-dimensional function.

et al., 1971) or found (e.g., Forsyth, 2004), observers do not seem to hold this belief.

Though these conclusions are powerful, it is possible to do better. The second way to apply a model such as this one is to use it to quantitatively assess the internal consistency of people's judgments, which is done by measuring quantities on both sides of the equation. Previous authors have done this using alternative Bayesian standards (Trope, 1974; Morris & Larrick, 1995; Forsyth, 2004). However, every previous model has required participants to think in terms of discrete probabilities (e.g., the probability that the author holds the attitude expressed in the essay), while nearly all other studies of attitude attribution ask participants to estimate the author's attitude on a Likert-type scale. Achieving a match to what participants customarily estimate requires switching from the probabilities of dichotomous events to probability *densities* over continuous variables, as follows:

- “Pro” and “con” attitudes are generalized to real-valued attitudes along a “con” (negative) to “pro” (positive) continuum, with attitudes further from zero being more extreme. The variable a will refer to the author's attitude, while e will refer to the position expressed in the essay.
- The circumstances (C) are decomposed into two things: p , the position that the author was asked to express, and s , the strength of that request. The variable p can vary as discussed above, while s can vary between zero (no inducement) and one (a completely compelling inducement).

Using the above variables, $P(A | E, C)$ becomes $p(a | e, p, s)$. Converting the prior, co-occurrence, and relative likelihood terms into probability distributions and multiplying the three over the range of a gives the probability of each possible attitude. The expected value of this distribution will be the attitude attribution, and the confidence in this attribution will be proportional to the distribution's standard deviation.

Completing the normative model requires specifying the forms of the three terms. The prior distribution, $p(a)$, is just the assumed attitude distribution in the population. The co-

occurrence term expresses how the circumstances vary with the author's attitude. Assuming a random assignment process, then this term equals one. This leaves the likelihood term, $p(e | a, p, s)/p(e | p, s)$. Since the denominator does not involve a , the expression can be written:

$$p(a | e, p, s) \propto p(a) \cdot p(e | a, p, s)$$

These terms will be called the posterior, the prior, and the (essay) likelihood, respectively.

The final task is to specify a form for the likelihood, $p(e | a, p, s)$. This can be done by determining the distribution of essay positions that an author with attitude a would write when asked to express position p , facing an inducement of strength s . Instead of requiring participants to estimate this themselves, the form of the function will be specified mathematically. Past research has found that observers expect constrained authors to express an attitude somewhere in between their own attitude and the position that was assigned (Miller & Rorer, 1982), which Reeder et al. (1989) refer to as the central tendency assumption. Thus, an author with (say) a strong con attitude who was asked to express a strong pro position would attempt to write a neutral essay. This expectation can be modeled by saying that when a , p , and s are known, $p(e | a, p, s)$ is a normal distribution, with:

$$\mu = a \cdot (1 - s) + p \cdot s$$

With no inducement ($s = 0$), $\mu = a$, the author's own attitude. With a completely compelling inducement ($s = 1$), $\mu = p$, the requested position. For other values of s , μ is a weighted compromise between a and p . While one could imagine ways that the distribution's standard deviation might depend on a , p , and s , for parsimony it will be assumed to be constant.

The above model of authors' responses specifies the distribution of e , given that the other variables are known. However, when applied, e is known but a is unknown. This does not present a problem, as illustrated in Figure 1. The left graph shows the essay distributions for two values of a (-2 and 2), where $s = .5$ and $p = 2$. The right graph shows the

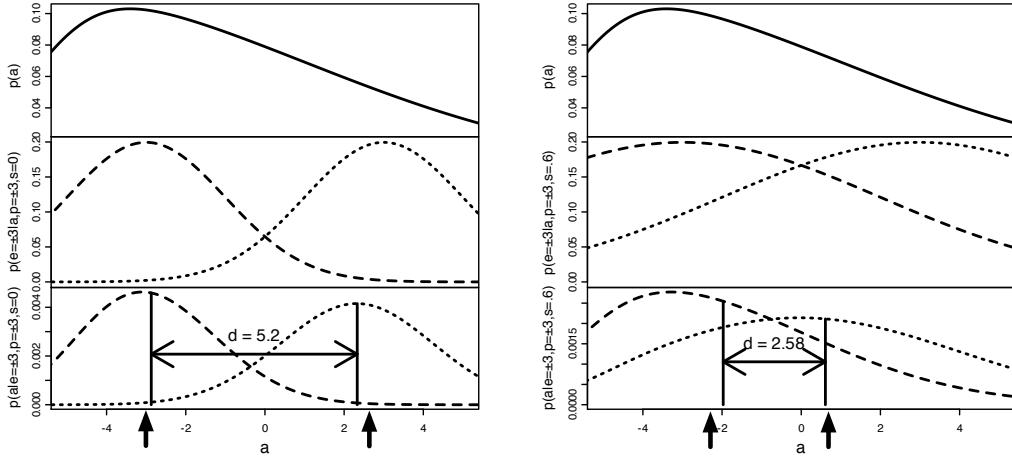


Figure 2: Model of the choice (left) and no choice (right) conditions for Jones and Harris (1967), Study 1. In the bottom panel, black arrows show original results, and black lines show the model’s results.

likelihood distribution over the range of a , where $e = 2$. The middle image shows how the two are related, with two points shown on all three graphs ($e = 2, a = \pm 2$) for reference.

As already mentioned, this model improves upon previous Bayesian models of attitude attribution in that its output is the same kind of variable as participants usually estimate. In addition, the model’s inputs correspond to perceptions that are relatively straightforward for participants to reason about (the position expressed in the essay, e , whether the essay is weaker or stronger than was expected of the author, $e - p$, and how constraining the situation was, s). When testing scenarios schematically, this makes it possible to continuously vary the model parameters, or to test specific combinations of parameters, rather than having to make verbal arguments about, say, the relative sizes of $P(\text{Essay} | \text{Attitude})$ and $P(\text{Essay} | \bar{\text{Attitude}})$. When testing the internal consistency of participants’ actual judgments, it becomes possible to directly ask for the relevant quantities. The tradeoff is that the model assumes people believe that constrained authors will express a position between their own attitude and the request. However, other models of how authors respond could be translated into likelihood functions, and the results compared.

Illustrations

Choice and Prior Probabilities

Correspondent inference theory (Jones & Davis, 1965) was intended to be a normative standard for how people should make attributions, and aims to specify which behaviors justify the inference of information about a person that would not have been assumed previously (Jones & McGillis, 1976). Jones and Harris (1967) was an attempt to show that while both constrained behavior and expected behavior do not contribute new information, an expected behavior performed under constraint will still lead to a corresponding attribution, simply because the underlying disposition would be expected

anyway. It is for this reason that they used advocacy for “Castro’s Cuba”—a highly unexpected behavior in 1960’s America—as the critical test. As predicted, they found that people made attributions corresponding to the constrained behavior when the behavior was expected (arguing against Castro). What they were surprised to find was that though people did not make completely corresponding attributions when the behavior was unexpected (arguing for Castro), their attributions did not revert to the level that would be obtained had the behavior been completely disregarded. This is the result that triggered the volumes of research on the correspondence bias that continues to this day.

This paper’s model can reproduce the pattern of results that Jones and Harris obtained, using two reasonable assumptions. First, assume that the prior attitude distribution was strongly right skewed (i.e., very few people supporting Castro).² Second, for parsimony, assume that the pro and con essays were equivalently strong, and no weaker or stronger than requested.

To reproduce the “choice” condition, the model is run with strength set to zero ($s = 0$), which is illustrated in the left half of Figure 2. The top panel of this graph shows the prior distribution, $p(a)$, while the middle panel shows the likelihood functions for the con (dashed line) and pro (dotted line) essays. These lines show $p(e | a, p, s)$, where $e = \pm 3$,³ $s = 0$ since authors could choose what to express (making the requested position, p , irrelevant), and a varies across the x -axis to encompass the range of attitudes shown. The bottom panel shows the posterior distributions, $p(a | e, p, s)$, which are the result of multiplying the prior distribution by either likelihood distribution. In this case, the prior distribution has only a small effect on the posterior distributions, and the expected

²See Jones and Harris (1967), p. 5.

³Note that attitude values are always rescaled to a -4 [con] to 4 [pro] for consistency of comparison across studies.

	Con				Pro			
	Strong	Weak	Weak	Strong	Strong	Weak	Weak	Strong
Requested position (p)	-2	-2	2	2				
Essay position (e)	-3	-1	1	3				
Attribution (a), weak constraint ($s = .25$)	-2.78	-0.56	0.56	2.78				
Attribution (a), strong constraint ($s = .75$)	-2.16	0.72	-0.72	2.16				
Essay position (e)	-3	-2	2	3				
Attribution (a), strong constraint ($s = .75$)	-2.16	-0.72	0.72	2.16				

Table 1: Model-predicted attributions for strong and weak essays under weak and strong situational constraint. Weak situation shows no reversal for the weak essays, but strong situation does. The same pattern can be obtained by keeping situation strength constant but making the weak essays less weak.

values of the distributions (shown by the black, vertical lines) are a very close match to the results that Jones and Harris originally obtained (shown by the black arrows).

Jones and Harris found nothing counterintuitive about their results for the choice condition, but were surprised by the results in the no choice condition, which can be replicated by choosing an appropriate value for s . Not shown in the figure is the case where the situation is seen as completely constraining ($s = 1$). Under conditions with no behavioral freedom, everyone is equally likely to have written the requested essay, and so the two likelihood functions are flat lines. As such, both posterior distributions are equal to the prior distribution, making the normative attribution for both essays equal to the mean attitude in the population. This result is what Jones and Harris were expecting to find. Since this is not what they obtained, values of s less than one must be tried.

A good fit to the original results was obtained with $s = .6$. The right half of Figure 2 shows this case, where it can be seen that though the prior distribution is the same and the likelihood functions have the same locations, the likelihood functions are also more spread out (since constrained behavior is less informative than freely chosen behavior). Even though the pro and con likelihood functions are equal and opposite, the posteriors are not, which is a result of multiplying by the asymmetric prior.

As the bottom panel shows, the expected values of either posterior (black, vertical lines) are quite close to the results that Jones and Harris obtained (black arrows). In particular, for the “con” essay, the model-derived and actual attributions are still in the direction of the essay. For the “pro” essay, however, multiplying by the prior probability has brought the model-derived results closer to the midpoint, and like the actual results, still somewhat correspondent with the essay itself. It is also worth noting that the “pro” posterior is more spread out than the “con” posterior, just as Jones and Harris found greater variance in this condition than in the other conditions of their study.

As the above shows, the model can reproduce the important features of the original demonstration of the correspondence bias, with only one parameter varying between the choice and no choice conditions. As such, it establishes that the results

in Study 1 of Jones and Harris (1967) could be the result of an internally consistent reasoning process, given the assumption that the participants did not believe that the author’s situation in the no choice condition was completely constraining. In fact, according to the model, the *only* internally consistent way for perceivers to make attributions other than to the mean attitude in the population is if they believe that the situation leaves room for choice, and that this choice depends on the compatibility of the author’s attitude and the assigned position. Though these attributions are probably externally invalid, the possibility that they are internally consistent suggests that defects in observers’ reasoning processes are not necessary to explain the results. Likewise, people may make perfectly reasonable assumptions about how situational constraints in general would influence essay authors. The source of “bias” may simply be that people applied those assumptions using an insufficiently strong appraisal of the power of the author’s particular situation.

Degree of Compliance

Thus far, it has been assumed that perceivers believe that the essay written was no weaker or stronger than was requested. However, compliance needn’t be all-or-nothing. Jones et al. (1971) manipulate the strength of the essay in order to understand how behavioral extremity affects attributions. One of their key results is that when people read an essay written under constraint and expressing a weak position, they attribute the *opposite* attitude to the author as was assigned. When the essay position was strongly argued, they attribute a corresponding attitude. In an attempt to replicate this result, Miller (1974) found that people made less extreme attributions when reading a weak essay than when reading a strong essay, but did not find any reversal. In both cases, however, the degree of compliance affected the attributions.

The model is able to reproduce these result patterns by varying the situation strength parameter, s , and leaving everything else constant. Model-predicted attitude attributions for weak and strong levels of constraint are shown in the top and middle of Table 1. Reversal occurs for the strong constraint, but not for weak constraint. Intuitively, this is because stronger constraints make it less likely that a person

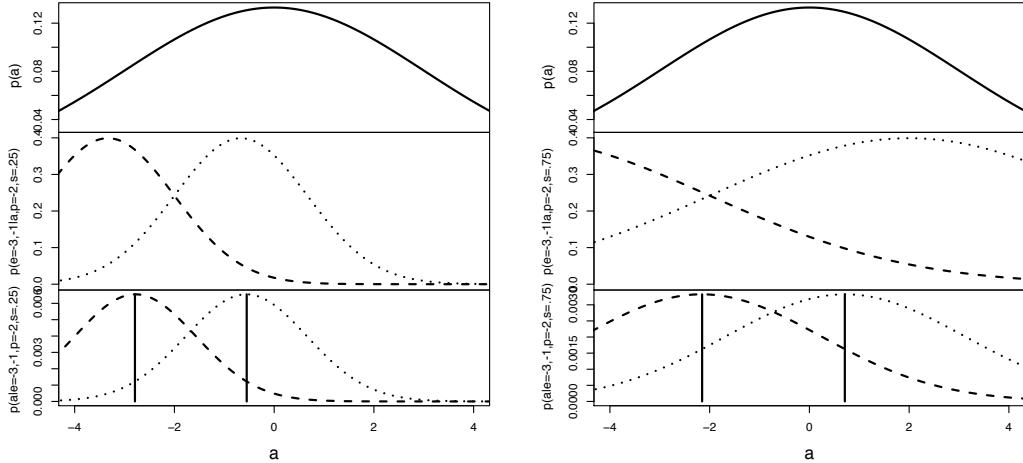


Figure 3: Illustration of strong and weak “con” essays for weak constraint (left, $s = .25$) and strong constraint (right, $s = .75$). In both cases, the requested position, p , is -2 and the essay positions, e , are -3 and -1 for the strong and weak essays.

would deviate from the requested position. Therefore, when someone does deviate from the requested position by writing a weaker-than-expected essay, it is reasonable to conclude that this person must hold an attitude very different than what was requested. This is illustrated in Figure 3. As can be seen, the likelihood functions are both further from the requested positions and more spread out at higher constraint.

Varying situation strength is not the only way to replicate the different patterns of results. The bottom two rows of Table 1 show what happens when the weak essays are made less ambivalent, but the strong level of constraint kept consistent. This change removes the reversal that had been obtained with the weak essays. In speculating on the failure to replicate the Jones et al. (1971) weak essay reversal, Miller (1974) does in fact note that his weak essays were not as weak as those in Jones et al. While both Jones et al. and Miller speculate that strong essays “engulf the field” whereas the weak essays allow the perceiver more latitude to notice the situation, the model suggests that no such perceptual metaphors are necessary. Instead, both outcomes are reasonable conclusions of an internally consistent logic that does not depend on any distortions in perception, failure to notice the situation, or alteration in the underlying behavioral model being used to make the attribution. This example also makes clear that there are often multiple internally-consistent ways to obtain the same pattern of results. The model makes it possible to explore many sources of a result, thereby suggesting hypotheses for behavioral research.

Empirical Results

In addition to fitting previous research results, the model fits new data (collected for a different purpose).⁴ Participants ($N = 246$) read essays for and against the death penalty, and then learned that the author was randomly assigned the position to take. Participants then rated what they thought the author’s attitude was, how confident they were in their answer, and other perceptions (detailed next). Replicating past results, there was a significant difference between the pro and con essay attributions ($M = -0.90$ vs. $M = 0.98$, $t(244) = -8.44$, $p < .0001$). Model-based predictions were then tested, after reversing all of the relevant quantities for participants in the “con” essay condition.

As the model of the Jones and Harris (1967) results showed, a skewed prior attitude distribution should result in asymmetric attitude attributions. In particular, attributions for essays expressing rare opinions should be closer to the midpoint than attributions for essays expressing common positions. Additionally, as judged by the variance of the posterior distributions, people should be less confident in their attributions when the expressed position is rare. This was tested by looking at participants’ self-reported prior attitude distributions, which were elicited by having people apportion 100 percentage points to three equal-sized intervals encompassing the measurement scale. A “skew” was calculated for each participant by taking the log ratio of the lower and upper intervals of their priors. Negative ratios imply more probability mass near the “pro” end of the scale, and positive ratios imply more probability mass near the “con” end of the scale. Supporting the model’s predictions, the correlation of attribution and skew was $r = -.14$ ($p < .05$), and the correlation of confidence and skew was $r = -.20$ ($p < .01$).

Next, the co-occurrence between the situation and attitudes was examined. As mentioned at the outset, if assignment is non-random, people might believe that the essay author’s own attitude and the assigned position are related. To test

⁴These data are part of an in-progress replication of Miyamoto and Kitayama (2002), which uses their essays as stimuli. In addition to having “pro” and “con” essays, the study varies essay length.

Since the effects listed next are not qualified by length, the length manipulation is not discussed further.

this, people were compared by whether they indicated (as intended) that the author had no control over assignment. There was a significant difference ($M = 0.73$ vs. $M = 1.21$, for no control vs. control, respectively, $t(244) = 2.16$, $p < .05$).

Finally, the likelihood model predictions were examined. As shown with the modeling of the Jones et al. (1971) and Miller (1974) result patterns, the model predicts that overcompliance and attribution extremity should be positively related, and that perceived situation strength and attribution extremity should be negatively related. Participants estimated overcompliance via a question asking how much weaker (or stronger) the essay was than what they believed was expected, and strength was measured via a question about how much overall choice the author had (reversed). After partialing out the effects of skew and strength, attribution and overcompliance were positively related, $pr = .14$ ($p < .05$). After partialing out skew and overcompliance, attribution and strength were negatively related $pr = -.16$ ($p < .05$). Because higher strengths lead to more spread out likelihood functions, the model also predicts that confidence and strength should be negatively related, which was supported $r = -.24$ ($p < .001$).

Though these results do not prove that people's attributions are internally consistent, they do demonstrate promise. Future work will systematically test the match between model predictions and empirical results in greater detail.

Conclusions

Using a simple yet plausible model of how people respond to instructions to advocate a particular opinion, this work derives a model that can predict prior attitude attribution results, and that fits newly-collected data. Though the correspondence bias can be seen when people's attributions are compared to the ground truth, this work suggests that these attributions could be internally consistent with other beliefs and perceptions that people have. Future work should investigate why these beliefs (e.g., about how people respond to requests) and perceptions (e.g., of the request strength or the essay extremity) don't match reality. The likelihood model could also be extended to encompass other essay features, such as argument quality (cf. Miller & Rorer, 1982; Gawronski, 2003).

Early in the history of correspondence bias research, Jones et al. (1971) conceded that correspondent inferences for constrained behavior are only wrong if every person in that situation would comply. Short of this extreme, they say that "it would be very difficult if not impossible to determine whether [a correspondent inference] should be judged as attributional distortion" (p. 77). The model presented here helps answer this question by encoding a set of assumptions mathematically, and then using the logic of Bayes' rule to understand the implications of those assumptions. It is likely that many attitude attribution findings can fruitfully be reexamined in light of the added precision that this model provides.

Acknowledgments

Tom Griffiths, Rob MacCoun, and Kaiping Peng provided valuable comments on and assistance with this work.

References

Forsyth, D. R. (2004). Inferences about actions performed in constraining contexts: Correspondence bias or correspondent inference? *Current Psychology*, 23(1), 41–51.

Gawronski, B. (2003). Implicational schemata and the correspondence bias: On the diagnostic value of situationally constrained behavior. *Journal of Personality and Social Psychology*, 84(6), 1154–1171.

Gilbert, D. T., & Jones, E. E. (1986). Perceiver-induced constraints: Interpretation of self-generated reality. *Journal of Personality and Social Psychology*, 50, 269–280.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, un-available injustice*. New York: Oxford University Press.

Jennings, K. E. (2010). *Coherent attributions with co-occurring and interacting causes*. Unpublished doctoral dissertation, University of California, Berkeley.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In *Advances in experimental social psychology* (Vol. 2, pp. 219–266). New York: Academic Press.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, 3, 1–24.

Jones, E. E., & McGillis, D. (1976). Correspondent inferences and the attribution cube: A comparative reappraisal. In J. H. Harvey, W. J. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 1, pp. 389–420). Hillsdale, NJ: Erlbaum.

Jones, E. E., Worchel, S., Goethals, G. R., & Grumet, J. F. (1971). Prior expectancy and behavioral extremity as determinants of attitude attribution. *Journal of Experimental Social Psychology*, 7, 59–80.

Miller, A. G. (1974). Perceived freedom and the attribution of attitudes. *Representative Research in Social Psychology*, 5, 61–80.

Miller, A. G., & Rorer, L. G. (1982). Toward an understanding of the fundamental attribution error: Essay diagnosticity in the attitude attribution paradigm. *Journal of Research in Personality*, 16, 41–59.

Miyamoto, Y., & Kitayama, S. (2002). Cultural variation in correspondence bias: The critical role of attitude diagnosticity of socially constrained behavior. *Journal of Personality and Social Psychology*, 83(5), 1239–1248.

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, 102(2), 331–355.

Reeder, G. D., Fletcher, G. J. Q., & Furman, K. (1989). The role of observers' expectations in attitude attribution. *Journal of Experimental Social Psychology*, 25, 168–188.

Trope, Y. (1974). Inferential processes in the forced compliance situation: A Bayesian analysis. *Journal of Experimental Social Psychology*, 10, 1–16.