Précis of *Social Perception and Social Reality: Why Accuracy Dominates Bias and Self-Fulfilling Prophecy*

Lee Jussim, Rutgers University

**Abstract:** *Social Perception and Social Reality* reviews the evidence in social psychology and related fields and reaches three conclusions: 1. Although errors, biases, and self-fulfilling prophecies in person perception, are real, reliable, and occasionally quite powerful, on average, they tend to be weak, fragile and fleeting; 2. Perceptions of individuals and groups tend to be at least moderately, and often highly accurate; and 3. Conclusions based on the research on error, bias, and self-fulfilling prophecies routinely greatly overstates their power and pervasiveness, and consistently ignores evidence of accuracy, agreement, and rationality in social perception. The weight of the evidence – including some of the most classic research widely interpreted as testifying to the power of biased and self-fulfilling processes – is that interpersonal expectations related to social reality primarily because they reflect rather than cause social reality. This is the case not only of teacher expectations, but also social stereotypes, both as perceptions of groups, and as the bases of expectations regarding individuals. The time is long overdue to replace cherry-picked and unjustified stories emphasizing error, bias, the power of self-fulfilling prophecies and the inaccuracy of stereotypes with conclusions that more closely correspond to the full range of empirical findings, which includes multiple failed replications of classic expectancy studies, meta-analyses consistently demonstrating small or at best moderate expectancy effects, and high accuracy in social perception.

**Keywords:** Social psychology, social cognition, social perception, accuracy, bias, self-fulfilling prophecy, teacher expectations, stereotypes
Is social perception – how people go about understanding other people, both individuals and groups – routinely compromised by a slew of flawed and biased processes, so that it becomes primarily a “reign of error” (Merton’s 1948, oft-repeated phrase)? Much social psychological scholarship would seem to converge on the conclusion that the answer is “yes.” And for many good reasons. Social and cognitive psychologists have clearly and successfully identified and documented a vast array of errors and biases that can and do sometimes undermine the validity, rationality, and reasonableness of lay judgment and social perception. Thus, for over half a century now, leading scholars of social perception have emphasized error and bias:

"Social perception is a process dominated far more by what the judge brings to it than by what he takes in during it" (Gage & Cronbach, 1955, p. 420).

"... the literature has stressed the power of expectancies to shape perceptions and interpretations in their own image” (E. E. Jones, 1986, p. 42).

“It does seem, in fact, that several decades of experimental research in social psychology have been devoted to demonstrating the depths and patterns of inaccuracy in social perception … This applies … to most empirical work in social cognition” (Jost & Kruglanski, 2003, pp. 172).

Such conclusions are the norm, not the exception, in social psychology. Consider next this passage from Clark & Clark-Polner’s (2012) review of Social Perception and Social Reality:

“Without relying on Jussim’s examples (though he presents many), we opened a social psychology textbook that was, simply, the one most accessible to us (Gilovich, Keltner, & Nisbett, 2006). It included references to “striking” demonstrations of stereotypes influencing interpretations of events, to research in which self-fulfilling prophecies has been “powerfully” illustrated (p. 455), and to self-fulfilling prophecies perpetuating a “reign of error” (quoting Merton, 1957, in the last case, pp. 455–456). The same chapter did not include a discussion of accuracy in perceptions or of accuracy captured in stereotypes themselves.”

Thus, social psychology has a longstanding consensus that social perception is dominated by error and bias. Social Perception and Social Reality, however, reviews almost 100 years of research and reaches a very different conclusion: People’s social perceptions (perceptions regarding individuals and groups) are
often reasonable, accurate, and arrived at through approximately rational processes. How can anyone make such a claim, given the overwhelming evidence of error, bias, and self-fulfilling prophecy, and the overwhelming consensus that such effects are powerful and pervasive? Although answering that question required an entire book, this article summarizes some of those arguments.

This précis is organized around reviewing and critically evaluating the empirical literature in social psychology and related fields, on the roles of error, bias, self-fulfilling prophecy, and accuracy in social perception. Very broad and seemingly unrelated literatures converge on three conclusions: 1. Although errors, biases, and self-fulfilling prophecies in person perception, are real, reliable, and occasionally quite powerful, on average, they tend to be weak, fragile and fleeting; 2. Perceptions of individuals and groups tend to be at least moderately, and often highly accurate; and 3. Common scholarly conclusions routinely overstates the power and pervasiveness of expectancy effects, and consistently ignores evidence of accuracy, agreement, and rationality in social perception. This pattern occurs over and over again across a wide variety of research areas within social perception. For short, therefore, I simply refer to it in this précis as “the tripartite pattern.”

Although chronology per se was not the main organizing principle, Social Perception and Social Reality reviews the literatures that bear on these questions in approximately chronological order. This is because it was important to first identify the scientific and scholarly foundations on which the dominant emphasis on error and basis were based. Thus, I begin with some of the earliest evidence on stereotypes, and on the New Look in Perception, both of which emphasized error and distortion in social perception (Section I: The Scientific Roots of Emphasis on the Biasing and Self-Fulfilling Power of Social Expectations). This emphasis received an intellectual “booster shot” with the publication of several articles in the late 1960s and 1970s on self-fulfilling prophecies (Section II: The Once Raging and Still Smoldering Pygmalion Controversy) and yet a second shot when research in the 1970s and 1980s began demonstrating a slew of expectancy-confirming biases (Section III: The Awesome Power of Expectations to Create Reality and Distort Perceptions).

Because of the combination of these diverse literatures, by the 1980s it was clear to many social
psychologists that expectancy-confirmation was a powerful and pervasive phenomena. *Social Perception and Social Reality* reconsiders and critically evaluates this evidence, concluding that such emphases were overstated, even on the basis of the research conducted up to that time (*Section IV: The Less Than Awesome Power of Expectations to Create Reality and Distort Perceptions*). Of course, demonstrating that error and bias are overstated is not equivalent to demonstrating that accuracy was high. However, accuracy itself is controversial in social psychology, and those controversies (*Section V: Accuracy Controversies*) and some key data (*Section VI: The Accuracy of Teacher Expectations*), are reviewed next. Last, I turn to one of the most difficult and controversial topics – the accuracy and inaccuracy of stereotypes, both as perceptions of groups (*Section VII: The Unbearable Accuracy of Stereotypes*), and their role in increasing or reducing the accuracy of person perception (*Section VIII: Stereotypes and Person Perception*).

I. The Scientific Roots of Emphasis on the Biasing and Self-Fulfilling Power of Social Expectations

The Early Research on Stereotypes

One of the first arguments that our perceptions are not necessarily strongly linked to objective reality came from a journalist. In a broad-ranging book called *Public Opinion*, Walter Lippmann (1922) touched on stereotypes -- and defined them in such a way as to color generations of social scientists' views of stereotypes. Lippmann suggested that to understand the world in its full complexity is an impossible task. So people simplify and reduce the overwhelming amount of information they receive. Stereotypes, for Lippmann, arose out of this need for simplicity. He believed that people's beliefs about groups were essentially "pictures in the head."

A “picture in the head” is a static, two-dimensional representation of a four-dimensional stimulus (most real-world stimuli have width, length, and depth, and also change over time). A picture is rigid, fixed, and unchanging. It is over-simplified and can never capture the full complexity of life for even one member of any group. This should sound familiar -- it constitutes the working definition of stereotypes that many people, including many social scientists, still hold today. Thus, it constitutes one of the earliest perspectives suggesting that people's social beliefs may not be fully in touch with social reality.
Social psychologists ran with these ideas. Katz & Braly (1933) concluded that the high levels of agreement they observed regarding national, racial, and ethnic groups could not possibly reflect personal experience and instead most likely reflected the shared expectations and biases of the perceiver. This analysis was flawed because agreement *per se* is not evidence of inaccuracy (often, though not always, it reflects accuracy – e.g., Funder, 1987). In a similarly flawed manner, La Piere (1936) interpreted his empirical results as demonstrating that stereotypes were inaccurate rationalizations of antipathy towards outgroups, even though (except for some anecdotes) he did not assess people’s stereotypes.

G. Allport (1954), in perhaps the most influential social psychological book written about stereotypes and prejudice, distinguished between, on the one hand, rational and flexible beliefs about groups, and on the other, stereotypes. Long ignored in many citations to G. Allport (1954/1979) is the fact that he clearly acknowledged the existence of rational and flexible beliefs about groups. He merely did not consider such beliefs to be stereotypes. For G. Allport, stereotypes are faulty exaggerations. All-or-none beliefs, such as "all Turks are cruel," are stereotypes that are clearly inaccurate, overgeneralized, and irrational, because there are virtually no social groups whose individual members universally share some set of attributes. G. Allport also characterized stereotypes as unjustifiably resistant to change, steeped in prejudice, lead to all sorts of errors and biases in social perception, and concluded they were a major contributor to social injustice. Overall, therefore, the early research on stereotypes helped set the stage for social psychology’s later emphasis on error and bias.

*Early Social Perception Research*

*The New Look in Perception.* The New Look of the 1940s was, in large part, a reaction against the prevailing view at the time that perception reflected the objective aspects of external stimuli. The dominant behaviorist perspective of the period banished fears, needs, and expectations from study, dismissing such internal states as unscientific. Then came the New Look researchers who, en masse, set out to demonstrate ways in which exactly such internal states could influence and distort perception (see F. Allport, 1955 for a review). The main claims of the New Look could be captured by two concepts: Perceptual vigilance and
perceptual defense. Perceptual vigilance referred to the tendency for people to be hypersensitive to perceiving stimuli that met their needs or were consistent with their values, beliefs, or personalities.

_Hastorf & Cantril (1954)._ Towards the end of the New Look era, Hastorf & Cantril (1954) published a paper that, though not formally part of the New Look program of research, is generally cited as an early classic supposedly demonstrating the powerful role of beliefs and motives in social perception. In 1951 Dartmouth and Princeton played a hotly contested, aggressive football game. A Princeton player received a broken nose; a Dartmouth player broke his leg. Accusations flew in both directions: Dartmouth loyalists accused Princeton of playing a dirty game; Princeton loyalists accused Dartmouth of playing a dirty game. Hastorf & Cantril (1954) showed a film of the game to 48 Dartmouth students and 49 Princeton students, and had them rate the total number of infractions by each team. Dartmouth students saw both the Dartmouth and Princeton teams as committing slightly over four (on average) infractions. The Princeton students also saw the Princeton team as committing slightly over four infractions; but they also saw the Dartmouth team as committing nearly ten infractions.

Because the Dartmouth and Princeton students diverged in the number of fractions they claimed were committed by Dartmouth, Hastorf & Cantril (1954) concluded that Princeton and Dartmouth students seemed to be actually seeing different games. The study has long been cited as a demonstration of how motivations and beliefs color social perception (e.g., Ross, Lepper, & Ward, 2010; Schneider, Hastorf, & Ellsworth, 1979; Sedikedes & Skowronski, 1991). As Ross et al, 2010, p. 23) put it: “The early classic study by Hastorf & Cantril (1954) … reflected a radical view of the ‘constructive’ nature of perception that anticipated later discussions of naïve realism.”

_F. Allport’s prescience about overemphasis on error and bias._ The New Look eventually faded away due to intractable difficulties overcoming alternative explanations for its findings (F. Allport, 1955). Nonetheless, it had a profound and lasting influence on social psychology. Despite losing many intellectual battles with those challenging their interpretations at the time, the New Lookers ultimately won the war -- and the victory was nearly absolute. Within social and personality psychology, the idea that motivations, goals, and expectations influence perception is now so well-established that it is largely taken for granted.
Floyd Allport (1955) saw this coming:

"Where the perception is bound so little by the stimulus and is thought to be so pervasively controlled by socially oriented motives, roles, and social norms, the latitude given for individual and group differences, for deviating and hence non-veridical awareness, is very great" (p. 367).

He also warned against overemphasizing bias and inaccuracy:

"What we are urging here is that social psychologists, in building their theories of perception, assume their share of the responsibility for reconciling and integrating their 'social-perceptual' concepts, fraught with all their deviations and special cognitive loadings, with the common and mainly veridical character of the basic human perceptions” (p. 372).

Floyd was right on both counts -- his concern that the New Look could lead to an overemphasis on subjective influences on perception could not have come more true; and he was right to urge social psychologists to develop theories that presented a more balanced vision of the roles of error, bias, and accuracy in social perception.

One can readily see this emerging pattern of overstated emphasis on error and bias in Hastorf & Cantril’s (1954, p. 133) own extraordinary and extreme interpretations of their study:

“There is no such ‘thing’ as a ‘game’ existing ‘out there’ in its own right which people merely ‘observe.’” and

“The ‘thing’ simply is not the same for different people…”

With such interpretations it is, perhaps, understandable why some (e.g., Ross et al, 2010) would cite the study as emphasizing radical constructivism. Unfortunately, however, the study’s results did not support such extreme conclusions. First, there was no difference in the infractions perceived by Dartmouth and Princeton students regarding the Princeton team. Thus, for half the data, the students saw essentially the same game, and there was no evidence of bias or “radical constructivism” at all.

Perceptions of the Dartmouth team did show about a six perceived infraction difference between the Princeton and Dartmouth students. This is indeed bias, and it was statistically significant. However, it is also useful to consider how much of a bias this was. Most college football games have about 100 plays, or
more. If one conservatively estimates that this particular game only had 60 plays (a low estimate biases conclusions in favor of bias), then bias of six means that 54 judgments, or 90%, were unbiased. So, half the judgments (for the Princeton team) were completely unbiased; half the judgments were 90% unbiased. At least 95% of the time, judgments were unbiased.

This study, then, is indeed foundational for modern social psychology, but not for the reasons it is usually cited. Instead, it should be foundational because:

1. It demonstrated that bias was real but quite modest.
2. It demonstrated that unbiased responding overwhelmingly dominated social perception.
3. Conclusions regarding the extent to which the data supported strong claims about the power of bias were greatly overstated by the original authors and by many of those subsequently citing the study.

This tripartite pattern does indeed anticipate much of the next 60 years of research on social perception.

II. The Once Raging and Still Smoldering Pygmalion Controversy

Although Merton (1948) first developed the self-fulfilling prophecy concept, it was Rosenthal & Jacobson’s (1968) book, *Pygmalion in the Classroom*, that launched self-fulfilling prophecies as a major area of inquiry in the social sciences and education. Rosenthal & Jacobson (1968) performed a study in which elementary school teachers were led to believe that certain of their students (who were actually randomly selected) would show dramatic IQ increases over the course of the year. Confirming the self-fulfilling prophecy hypothesis, on average, those late bloomers did indeed show greater IQ increases than their classmates. The study has frequently been cited in support of arguments claiming that self-fulfilling prophecies are pervasive, and potentially a powerful force in the creation of social inequalities and injustices. (e.g., Gilbert, 1995; Jones, 1990; Weinstein, Gregory, & Strambler, 2004; see Wineburg, 1987 for a critical review).

Are such claims justified? The combination of uncritical social psychological acceptance of the study and scathing methodological and statistical criticisms (Elashoff & Snow, 1971; Snow, 1995) complicates answering this question. Nonetheless, even if one takes its results entirely at face value, the justified conclusions are considerably more narrow than claims of powerful and pervasive self-fulfilling prophecies.
suggest, as can be shown by the answers to six simple questions about the study:

1. Were teacher expectations typically inaccurate? This was not assessed.

2. Did stereotypes bias expectations? This was not assessed.

3. Were self-fulfilling prophecies powerful and pervasive? They were not typically powerful. The overall effect size equaled a correlation of .15. The mean difference in IQ gain scores between late bloomers and controls was four IQ points.

   Nor were they pervasive. Significant teacher expectation effects only occurred in two of six grades (in year one) and in one of five grades in year 2. Self-fulfilling prophecies did not occur in eight of eleven grades examined.

4. Were powerful expectancy effects ever found? Yes. The results in first and second grade in year one (15 and 10 point bloomer-control differences) were quite large.

5. Were self-fulfilling prophecies harmful? No. Rosenthal & Jacobson (1968) only manipulated positive expectations. They showed that false positive expectations could be self-fulfilling. They did not assess whether false negative expectations undermine student IQ or achievement.

6. Did self-fulfilling prophecies accumulate over time? No. The mean IQ difference between bloomers and controls in year one was about 4 points; in year two it was under 3 points.

The finding that teacher expectations might sometimes produce self-fulfilling prophecies was interesting and important on its merits. Nonetheless, these results provided little terra firma for theoretical testaments to the power of beliefs to create reality, or practical concerns about the role of self-fulfilling stereotypes in oppression and inequality.

That is all true if the study is taken at face value. However, it is not clear that the study’s results should be taken at face value. Snow’s (1995; Elashoff & Snow, 1971) critiques raised questions about the ability of the study to reach any conclusions about self-fulfilling prophecies. For example, there were five “bloomers” with wild IQ score gains: 17-110, 18-122, 133-202, 111-208, and 113-211. If one excluded these five pairs of bizarre scores, the difference between the bloomers and the controls evaporated.

Such controversies sparked attempts at replication. Nearly 2/3 failed, providing fodder for the critics
But over 1/3 succeeded, when only 5% should succeed if there was really no effect. One of the earliest meta-analyses showed that there was an overall statistically significant effect of experimentally manipulated expectations (Rosenthal & Rubin, 1978).

It might seem this should end the controversies, but it did not. A paper titled, “The self-fulfillment of the self-fulfilling prophecy” contested the central and most controversial aspect of the original Pygmalion study – the effect on IQ (Wineburg, 1987; the Rosenthal & Rubin, 1978, meta-analysis included many self-fulfilling outcomes and did not focus on IQ, so did not resolve this issue).

Several reviews and meta-analyses have addressed the IQ controversy, with some authors emphasizing the existence of the effect on IQ (Raudenbush, 1984, 1994) and others remaining deeply skeptical (e.g., Snow, 1995; Spitz, 1998; Wineburg, 1987). Nonetheless, one conclusion does clearly emerge from this ongoing controversy: If there is an effect on IQ, it is not very large. Even the meta-analyses reporting the strongest effects showed that the mean and median effect sizes, overall, were r<.10 (Raudenbush, 1984, 1994). The strongest effects on IQ occurred in a handful of experiments in which teacher expectations were manipulated within the first two weeks of the school year, and even those were merely r=.21 (Raudenbush, 1984, 1994). Others have concluded that the average IQ effect was actually closer to r=0 (Snow, 1995; Wineburg, 1987).

What, then, are justifiable take-home messages from Pygmalion and the subsequent controversies and follow-up research? Self-fulfilling prophecies in the classroom are real, but far from inevitable. Although such effects are occasionally powerful, they are generally weak, fragile, and fleeting. Self-fulfilling outcomes can occur on a wide variety of variables, including grades and standardized tests. However, if there is any effect on IQ, it is typically small.

For all its limitations, Pygmalion also became a seminal study, at least in part, because it provided a simple and elegant methodology for examining self-fulfilling prophecies – experimentally manipulate expectations and then assess effects on targets. Thus, many social psychologists were about to fall in love with expectancy effects. I review this material here twice: Once in the unabashedly enthusiastic manner typically used to describe this research in the social psychology literature (III. The Awesome Power of
Expectations to Create Reality and Distort Perceptions); and then again, in a separate section that critically examines this research (IV. The Less Than Awesome Power of Expectations to Create Reality and Distort Perceptions). By conveying a sense of this initial enthusiasm, I hope to provide some insight into the good reasons why so much writing about expectancy effects has emphasized their power and pervasiveness (indeed, I could not think of a better way to explain why this research is still commonly discussed or cited in a similarly uncritical and enthusiastic manner to this day – e.g., Jost & Kruglanski, 2002; Ross et al., 2010; Weinstein, Gregory & Strambler, 2004 -- than to present this research in an enthusiastic and uncritical manner).

III. The Awesome Power of Expectations to Create Reality and Distort Perceptions

Despite the many limitations to Pygmalion in particular, and to teacher expectation research more generally, social psychological reviews generally accepted its conclusions and ran with its implications enthusiastically (e.g., Darley & Fazio, 1980; Jones, 1986; Miller & Turnbull, 1986). Pygmalion hit a sensitive social and political nerve. It was published in the late 1960's, when liberalism was at a political peak. The consciousness of much of the country had been raised regarding the extent to which racism and discrimination contributed to the massive inequalities between Whites and minorities. So when the Rosenthal and Jacobson (1968) study came along, and to this day, it has frequently been interpreted as demonstrating a widely generalizable mechanism of racial and social oppression.

Social Psychology Falls In Love with Self-Fulfilling Prophecies

Many social psychologists were able to tell compelling stories about the results of Pygmalion in particular, and the power of self-fulfilling prophecies more generally (e.g., Darley & Fazio, 1980; Gilbert, 1995; Jones, 1986; Jost & Kruglanski, 2002). Many studies yielded results seeming to support this perspective. Self-fulfilling prophecies occur, in part, because expectations lead perceivers to treat high expectancy targets differently than they treat low expectancy targets, and this differential treatment evokes expectancy-confirming target behavior. One classic pair of studies demonstrated this process: White interviewers’ nonverbal behavior discriminated against Black interviewees, and when white interviewees were subjected to the same behavior, their interview performance declined (Word, Zanna, & Cooper, 1974).
Similarly, teachers were at least sometimes more supportive of White students than of Black students (Rubovitz & Maehr, 1973; Taylor, 1979). When women believed an attractive male interviewer was sexist, they presented themselves as more traditional, scored lower on an anagrams test, wore more makeup and accessories, and talked less (von Baeyer, Sherk, & Zanna, 1981; Zanna & Pack, 1975). An observational study of children in kindergarten through second grade concluded that teachers social class-based expectations created a “caste system” advantaging middle class students over lower class students (Rist, 1970).

One of the most influential and highly-cited classics of this era demonstrated the self-fulfilling effects of the physical attractiveness stereotype (Snyder, Tanke, & Berscheid, 1977). Men were misled (through photographs) to believe a woman in another room was either attractive or unattractive. Not only did they behavior in a friendlier and warmer manner to the women believed to be attractive, those women reciprocated with warmer and friendlier behavior themselves. Thus, originally false beliefs about the social skill of the attractive became (self-)fulfilled.

Self-fulfilling prophecies were not restricted to stereotypes. Competitive people saw the world as competitive and evoked competitive behavior even from people predisposed to be cooperative (Kelley & Stahelski, 1970). People who falsely believed others are hostile evoke hostile behavior (Snyder & Swann, 1978a). Israeli military instructors evoked expectancy-confirming performance from military trainees (Eden & Shani, 1982). Self-fulfilling prophecies seemed to be everywhere psychologists turned.

Expectancy-Confirming Biases

Self-fulfilling prophecies are not the only effect of expectations. Interpersonal expectancies also bias judgments of social reality. The extraordinary power of stereotypes regarding demographic categories, occupation, roles, mental diagnoses and many other social categories to bias judgments is a common theme in social psychological scholarship. For example, in one classic study, after viewing a fourth grade girl take a test, perceivers judged her to have performed more highly and to be smarter if they believed she was from a higher rather than lower social class background (Darley & Gross, 1983). Yet another concluded that mental illness labels (e.g., “schizophrenia”) led to such powerful expectancy biases that it became impossible to
distinguish the sane from the insane (Rosenhan, 1971). People constructed false “memories” about the supposed facts of a woman’s life based on their stereotypes of whether she was lesbian or heterosexual (Snyder & Uranowitz, 1978). Similar findings obtained for stereotypes based on race, gender, and many other categories. In this context, it is perhaps unsurprising that one major review declared stereotypes to be the “default” basis of person perception (Fiske & Neuberg, 1990).

Such biases were not restricted to stereotypes, and occurred for expectations regarding intro/extraversion, friendliness, and intelligence (e.g., Kulik, 1983; Rothbart, Evans, & Fulero, 1979; Williams, 1976). Furthermore, such biases also infected social information-seeking. In an influential series of studies, Snyder & Swann (1978b) found that not only do people systematically seek information that confirms their hypotheses, they constrain targets’ ability to do much other than confirm the initially erroneous expectation.

The extent to which expectations influence, change, and color (or, for stereotypes, taint) our interactions with and perceptions of other people seemed to be nothing short of stunning. The social psychological enthusiasm for expectancy-induced biases was at least comparable to that expressed for self-fulfilling prophecies. Here are some quotes representative of a widespread consensus in social psychology:

“Owing to a variety of cognitive biases, a perceiver’s initial expectancies for a target are apt to be maintained, regardless of whether the target’s behavior confirms, disconfirms, or is ambiguous with respect to the perceiver’s expectancy (Darley & Fazio, 1980).” (Deaux & Major, 1987, p. 381).

“Specifically, all of these processes are biased in the direction of maintaining the preexisting belief system, that is, the very stereotype that initiated these biasing mechanisms.” (Hamilton, Sherman, & Ruvolo, 1990, p. 39).

The thrust of dozens of experiments on the self-fulfilling prophecy and expectancy-confirmation processes, for example, is that erroneous impressions tend to be perpetuated rather than supplanted, because of the impressive extent to which people see what they want to see and act as others want them lifto act …” (Jost & Kruglanski, 2003, pp. 172-173).

“A particularly pernicious example of self-fulfilling beliefs and expectations, and the one most studied by social psychologists, is that of stereotypes and other negative beliefs about particular groups of
people. Some of these effects are obvious, although no less important for their obviousness. If it is widely believed that the members of some group disproportionately possess some virtue or vice relevant to academic or on-the-job performance, one is likely (in the absence of specific legal or social sanctions) to make school admission or hiring decisions accordingly – and in so doing to deprive or privilege group members in terms of opportunities to nurture their talents, acquire credentials, or otherwise succeed or fail in accord with the beliefs and expectations that dictated their life chances.” (Ross et al, 2010, p. 30, emphasis mine).

IV. The Less Than Awesome Power of Expectations to Create Reality and Distort Perceptions

In fact, however, this emphasis on the power of interpersonal expectancies was unjustified. It was not justified by the classic early studies that remain highly cited today; it was not justified by other, less well-known research on expectancy effects from the same era; and it was not justified by the subsequent research.

This can be readily seen from Table 1, which presents the average effect size for both self-fulfilling prophecies and biases, as obtained in every relevant meta-analysis I could find. Except for the .52 effect among military personnel, all range from about 0 to about .3 and do not show powerful or pervasive expectancy effects. In light of the conclusions emphasizing their power, how can the effects be as modest as shown in Table 1?

That answer is complex, because it involves a scientific tradition that once emphasized telling compelling theoretical/political stories over attention to effect sizes and replication. It involves some blatant cherry-picking (highlighting studies that make for great stories, and systematically ignoring studies inconsistent with the preferred story). And it involved an apparent suspension of the skepticism that often justifiably characterizes scientific scholarship.

Many of the most influential and highly-cited classics of the expectancy-confirmation literature either suffered from serious methodological or interpretive problems, or have proven difficult to replicate. I review only two examples here, and the book presents many more.
Rist (1970)

Rist (1970) conducted an observational study of kindergarten through second grade, and concluded that teachers’ social-class based expectations were so powerfully self-fulfilling that they created a “caste system” serving to maintain the advantages of middle class students. According to Google Scholar, this study has been cited over 1600 times. It is quite striking, therefore, to discover that it actually provided no evidence of self-fulfilling prophecies whatsoever. Rist (1970) reported only a single piece of evidence regarding student achievement, and that was in a footnote (footnote 5, p. 443). That footnote reported that, at the end of the year, there were no differences in the IQ scores among the kindergarten students who were targets of high or low social class teacher expectations. In other words, his only quantitative assessment of achievement provided no evidence that teacher expectations produced changes in student achievement.

Rist (1970) did provide a wealth of information about teacher treatment of students. In short, the teacher assigned the students to tables based on their social class, and proceeded to direct most of her attention to the middle class students. Rist’s (1970) “caste system” conclusion was based on his observation that this table assignment pattern continued partially intact through second grade. However, it was only partially intact, and, indeed, there was actually considerable movement among students from kindergarten to first grade and again from first grade to second grade. If there was a “caste system,” it was a strikingly fluid one that produced no observed impact on students’ achievement by the only measure of such impact reported.

Rosenhan (1973)

Rosenhan (1973, cited over 2000 times) tested – and claimed to confirm -- one of the most audacious hypotheses in all of psychology: that the insane are indistinguishable from the sane. This is so extreme that readers might naturally wonder if I am setting up some sort of straw argument by overstating Rosenhan’s claims. Here is what Rosenhan (1973) himself wrote in his paper:

“If sanity and insanity exist, how shall we know them? The question is neither capricious nor itself insane. However much we may be personally convinced that we can tell the normal from the abnormal, the evidence is simply not compelling.” (opening sentences, p. 250).
“Based in part on theoretical and anthropological considerations, but also on philosophical, legal, and therapeutic ones, the view has grown that psychological categorization of mental illness is useless at best and downright harmful, misleading, and pejorative at worst.” (p. 251).

“Psychiatric diagnoses, in this view, are in the minds of the observers and are not valid summaries of characteristics displayed by the observed.” (p. 251).

“We now know we cannot distinguish insanity from sanity.” (p. 257).

I have not overstated his claims; instead, his claims themselves are vast overstatements. To understand how and why, it is necessary to first summarize his report. He had eight people (“pseudopatients”) with no prior histories of mental illness admitted to psychiatric hospitals in order to see if the professional staff could identify them as sane. To get admitted, all eight complained that they had been hearing voices. Upon admission, they ceased displaying all intentionally false expressions of disturbed behavior and they did not intentionally alter any other aspect of their life history.

They were kept institutionalized for an average of 19 days. When they were released, none were identified as sane; all were released with a diagnosis of “schizophrenia in remission.” Rosenhan (1973) also provided qualitative examples of staff interpreting normal behavior as evidence of pathology (e.g. pacing halls out of boredom was interpreted as nervousness). Thus, Rosenhan (1973) concluded that the sane were indistinguishable from the insane because diagnosis pervasively colored the institutional staff members' interpretations of the pseudo-patients' behavior and life histories.

However, there is actually far more evidence of reasonable, rational, and valid judgment on the part of the doctors and staff than first appears. How the pseudopatients initially got themselves admitted should give some reason for pause. *They were admitted complaining of auditory hallucinations.* Regularly hearing voices saying things like "thud," "empty," and "hollow"(what they claimed to be hearing) is not remotely normal. Therefore, an initial diagnosis of some form of psychosis does not seem to reflect gross distortion on the part of the psychiatric staff.

How rigidly resistant to change were the doctors' and staffs' expectations? Rosenhan's (1973) interpretation was that they were highly rigid. After all, none were diagnosed as sane. But let's focus on
Rosenhan's actual results, rather than his interpretations. First, the average hospital stay was 19 days, and most were kept under two weeks. How this reflects rigidity was never articulated.

How about the diagnosis of "schizophrenia in remission"? Rosenhan argued that it showed that there was nothing these completely sane pseudopatients could do to convince the doctors that they were really sane. However, “schizophrenia in remission,” at that time, meant “the patient is showing no current signs of schizophrenia” (Spitzer, 1975; Spitzer, Andreasen, & Endicott, 1978). Thus, in Rosenhan’s (1973) own data, and in contrast to his conclusions, the staff did indeed recognize that the pseudopatients were behaving in a manner devoid of evidence of psychosis.

Rosenhan (1973) also reported a follow-up study in which staff at institutions were informed to be on the lookout for pseudopatients. Because none were actually sent, any identification of a person as a pseudopatient is an error, and all such errors were interpreted by Rosenhan (1973) as supporting his extraordinary “the sane are indistinguishable from the insane” hypothesis. How many such errors did the psychiatrists make? Although Rosenhan (1973) did not report the data necessary to compute this figure exactly, it can be plausibly estimated as no higher than 6%, and probably considerably lower.

To keep the math simple, let’s assume there were only two psychiatrists and we interpret “at least one” to mean “half” (the result is the same if we take half of two, or half of 100). If it was more than half, Rosenhan (1973) probably would have stated so. Two psychiatrists by 193 patients is 386 judgments. 21 (judged fakers)/386 = 6%. 6% errors is the same as 94% accuracy.

Given the possibility that 6% of those admitted were, in fact, not suffering from psychopathology, even 6% may overstate the actual error rate. Any error is, well, an error – but these results are not exactly a testament to the extraordinary biasing power of psychiatric diagnoses and expectations. Indeed, the entire study – its results demonstrating high accuracy and small but real bias, and the manner in which its evidence of bias was so greatly overstated – is consistent with the tripartite pattern I first used to describe Hastorf & Cantril (1954): 1. Bias is real but small; 2. accuracy is very high; 3. The conclusions greatly overstated the power and pervasiveness of bias.
The Replication Failures

Many classic studies in the expectancy-confirmation literature have proven difficult to replicate. Attempts to replicate Snyder et al.’s (1977) self-fulfilling physical attractiveness stereotype study, Darley & Gross’s (1983) social class stereotype bias study, and Snyder & Uranowitz’s (1978) stereotype-based reconstructive memory studies all failed (Andersen & Bem, 1981; Baron, Albright, & Malloy, 1995; Belezza & Bower, 1981). In contrast to Rist’s (1970) conclusions, social class biases found in large-scale, quantitative studies of teacher expectations have consistently been nonexistent (Jussim, Eccles, & Madon, 1996; Madon et al. 1998; Williams, 1976).

Several lines of research followed up on the Snyder & Swann (1978b) study finding that people seek to confirm their social expectations by asking people leading questions that essentially remove from targets the opportunity to do anything except provide confirmatory answers. These have generally focused, not on attempts at exact replication, but on the validity of Snyder & Swann’s (1978) conclusion that people are heavily biased towards confirming their social expectations. Snyder & Swann (1978) only gave people the opportunity to ask leading questions. Numerous follow-up studies, however, recognized this limitation and addressed it either by allowing people to make up their own questions or to select from both leading and diagnostic questions (e.g., Devine, Hirt, & Gehrke, 1990; Trope & Bassok, 1982, 1983). When left to their own devices, or given adequate choice, people overwhelmingly ask diagnostic questions, and they almost never ask the type of leading questions found in Snyder & Swann (1978). There does appear to be a slight tendency to ask questions to which a “yes” answer will confirm perceivers’ expectations, and combined with a slight tendency on the part of targets to acquiesce, social hypothesis-testing may indeed be slightly biased in favor of confirming perceivers’ hypotheses (Zuckerman, Knee, Hodgins, & Miyake, 1996).

Nonetheless, Snyder & Swann (1978b) is cited more than all these other studies put together, and the most common pattern is to cite it as demonstrating biased social hypothesis testing, without citing any of the research showing that people generally ask diagnostic questions (e.g. Deaux & Major, 1987; Miller & Turnbull, 1986). Similar citation patterns characterize much of the expectancy literature. Dramatic
demonstrations of bias or self-fulfilling prophecy typically receive abundant attention whereas the failures to replicate that finding, and demonstrations of accuracy and rationality are largely overlooked.

This, then, is another route demonstrating the tripartite conclusion – bias is real but generally small; people are mostly accurate and rational; results demonstrating bias are overstated. In these cases, however, it is not necessarily the original researchers who overstate the result; instead, the overstatement occurs because attention (citations) primarily focus on, and conclusions primarily emphasize, results of one dramatic (though flawed) demonstration of bias, and the more abundant and generally higher quality research demonstrating small (or irreplicable) bias and high accuracy/rationality is typically overlooked or ignored.

*Quest for the Powerful Self-Fulfilling Prophecy*

Having discovered this tripartite pattern repeated over and over, it seemed important to try to discover if there were *any* conditions under which truly powerful self-fulfilling prophecies in the classroom occurred. Thus, we embarked on a quest to systematically search for conditions under which large expectancy effects occurred (Jussim et al., 1996; Madon, Jussim, & Eccles, 1997). Using a data set including over 100 teachers and over 1000 students, we found a slew of powerful self-fulfilling prophecies, with effect sizes (standardized regression coefficients) ranging from about .40 to about .60. Powerful self-fulfilling prophecies occurred among:

1. African-American students
2. Students from lower SES backgrounds (regardless of ethnicity)
3. Students with histories of low prior achievement who were from lower SES backgrounds (these .6 effects are among the most powerful ever found in social psychology)
4. Students with histories of low achievement who were the target of high expectations. High expectations uplifted such students more than they uplifted high achievers, and more than low expectations harmed achievement.

Although powerful self-fulfilling prophecies are the exception rather than the rule, they systematically occurred among students from stigmatized social backgrounds. Interestingly, in our data, they seemed to *ameliorate more than cause* social inequalities (uplifting students with histories of low achievement).
Do Self-Fulfilling Prophecies Accumulate or Dissipate?

In light of findings that expectancy-based biases and self-fulfilling prophecies are occasionally large but generally quite modest, researchers seeking to maintain a view of self-fulfilling prophecies as powerful and pervasive contributors to social problems needed to generate new arguments for doing so. The seemingly most compelling of these was that self-fulfilling prophecies may accumulate over time and/or over multiple perceivers (e.g., Claire & Fiske, 1998; Fiske, 1998). The logic of accumulation is straightforward:

1. Small effects are typically obtained in both short-term laboratory studies of self-fulfilling prophecies and teacher expectation studies conducted over a school year;
2. Although small in such contexts, many targets may be subjected to the same or similar erroneous expectations over and over again. Social stereotypes, widely assumed to be widely shared and erroneous, are often presented as an obvious reason to predict that targets from stigmatized groups will be subjected to repeated self-fulfilling prophecies from multiple perceivers over long periods of time. Thus, effects of expectancies on any particular target are likely to be much higher than demonstrated in any particular study.

There are, however, also compelling reasons to predict that, rather than accumulating, self-fulfilling prophecies will dissipate, including regression to the mean, self-verification (Swann & Ely, 1984), and accuracy (see the book for a full discussion of each). Thus, regardless of how “compelling” the accumulation argument may seem at first glance, the issue is an empirical one. Do self-fulfilling prophecies accumulate?

Every teacher expectation study that has assessed whether self-fulfilling effects that occurred in one year accumulate over time has found the exact opposite: They dissipate over time. Self-fulfilling prophecies dissipated in the original Rosenthal & Jacobson (1968) study, where the IQ difference between bloomers and controls was about four points in the first year, and under three points in the second year. Rist (1970) is often cited as evidence of accumulation, but he found neither accumulation across years nor self-fulfilling prophecy. West & Anderson (1976) followed 3000 students through high school, and found that teacher expectation effects declined from .12 the first year to .06 in the final year (standardized regression coefficients). We also tested accumulation over five to six years in math (from sixth or seventh grade...
through twelfth grade), and, instead, found dissipation (Smith, Jussim, & Eccles, 1999). The typically modest self-fulfilling prophecies found in sixth and seventh grade (.10, .16, respectively) declined to 0 and .09, respectively, by twelfth grade. Dissipation has also been found when research has followed students from first through fifth grade, in both reading and math (Hinnant, O’Brien, & Ghazarian, 2009).

Compelling stories can and have been told about how the accumulation of self-fulfilling prophecy upon self-fulfilling prophecy constitutes a major mechanism by which social stereotypes confirm themselves and maintain unjustified systems of oppression and status (e.g., Claire & Fiske, 1998; Darley & Fazio, 1980; Snyder, 1984; Weinstein et al, 2004) – typically without consideration or review of the considerable evidence indicating that self-fulfilling prophecies dissipate. Nonetheless, there is currently no clear evidence supporting such an analysis, and a great deal of evidence disconfirming it.

Conclusion: The Less Than Awesome Power of Expectations to Create Self-Fulfilling Prophecies, and Bias Perception, Judgment, and Memory

Do expectations lead to self-fulfilling prophecies and biases in judgment, perception, and memory? Yes, at least sometimes. But even the early blush of research on expectancy effects – the era filled with “classics” in the study of self-fulfilling prophecies and bias – never showed that such effects are, on average, inevitable, powerful or as pervasive as often claimed. Such effects are not only relatively small, on average, but they tend to be quite fragile, in the sense that seemingly small changes in experimental procedure, geography, type of dependent variable, or researcher often seem to lead such biases to mostly or completely evaporate, and sometimes, to completely reverse.

Just because bias tends to be small, however, does not necessarily mean that accuracy tends to be high. Evaluating the accuracy question is simultaneously very simple and dauntingly complex. Therefore, the complexities of studying accuracy are summarized next.

V. Accuracy Controversies

What could be a more basic or obvious purpose of social perception research than assessment of the accuracy of people’s perceptions of one another? And what could be simpler? Although both questions are phrased rhetorically, it turned out that, not only was the study of accuracy less simple than it seemed, it is, in
fact, a theoretical, methodological and political minefield. This section reviews, critically evaluates, and contests many of the reasons why social scientists have claimed that social perceptual accuracy is an unimportant, dangerous, or intractable topic.

Political Objections

Some have criticized accuracy research because it can be used to justify inequality. For example, Stangor (1995, p. 288-9) explains why stereotype accuracy is not worthwhile to study, in part this way: "As scientists concerned with improving the social condition, we must be wary of arguments that can be used to justify the use of stereotypes." (Stangor, p. 288-9). And then later in the same paragraph: “…we cannot allow a bigot to use his or her stereotypes, even if those beliefs seem to them to be accurate.” This is an explicitly political criticism of accuracy research. It refers quite bluntly to political power rather than science (“cannot allow a bigot”). People in power make decisions about what is allowable, whereas, presumably, scientific research does not.

Nonetheless, such a criticism fails to threaten accuracy research. First, scientific conclusions should be based on empirical evidence, and not be subject to political litmus tests. Second, it cannot be logically possible to reach conclusions about inaccuracy – and the four decade-long emphasis on error and bias in social cognition provides ample evidence that social psychologists do indeed often wish to reach conclusions about inaccuracy -- unless we can also reach conclusions about accuracy. Third, if we think we are curing a social problem (e.g., inequality) by treating the wrong disease (the supposedly inaccurate expectations whose accuracy social psychologists rarely assess and which, therefore, may be far more accurate than many seem to assume) we may not get very far.

Furthermore, there will be no way to assess our success at leading people to adopt more accurate beliefs, unless we have techniques for assessing accuracy. By understanding what leads people astray, and what leads them to accurate judgments, we will be much more capable of harnessing those factors that lead to accurate judgments, and therefore, reduce social problems resulting from inaccurate beliefs. Thus, even on the political grounds of aspiring to reduce inequality, political objections fail to provide a serious scientific threat to the study of accuracy.
Regardless, opposition to accuracy research on political grounds has a kernel of truth. Accuracy cannot explain social problems. Demonstrating that people's sex stereotypes are accurate (Swim, 1994) or that people's racial stereotypes are accurate (McCauley & Stitt, 1978) does nothing to alleviate or explain injustices associated with sexism or racism. Worse, demonstrating social perceptual accuracy can be viewed as not merely documenting high acumen in perceiving individual and group differences, but as implicitly reifying and justifying those differences. To characterize a belief that some kid is not too bright, or is a klutz on the basketball court, or is socially inept as "accurate" has a feel of "blaming the victim." Blaming the victim is a bad thing to do -- it means we have callously joined the perpetrators of injustice.

Nonetheless, this argument fails to threaten accuracy research. First, scientific conclusions should be based on empirical evidence, and not be subject to political litmus tests. Second, it cannot be logically possible to reach conclusions about inaccuracy – and the four decade-long emphasis on error and bias in social cognition provides ample evidence that social psychologists do indeed often wish to reach conclusions about inaccuracy -- unless we can also reach conclusions about accuracy. Third, if we think we are curing a social problem (e.g., inequality) by treating the wrong disease (the supposedly inaccurate expectations whose accuracy social psychologists rarely assess and which, therefore, may be far more accurate than many seem to assume) we may not get very far.

Furthermore, there will be no way to assess our success at leading people to adopt more accurate beliefs, unless we have techniques for assessing accuracy. By understanding what leads people astray, and what leads them to accurate judgments, we will be much more capable of harnessing those factors that lead to accurate judgments, and therefore, reduce social problems resulting from inaccurate beliefs. Thus, even on the political grounds of aspiring to reduce inequality, political objections fail to provide a serious scientific threat to the study of accuracy.

**Theoretical Objections**

Not all objections to accuracy research are political. Next, therefore, I consider some of the most common substantive and theoretical objections to accuracy research.
"Cognitive processes are important, error and bias is important, but accuracy is not." This strong argument has been explicitly articulated by various social psychologists (Jones, 1986, 1990; Schneider, et al., 1979; Stangor, 1995). Furthermore, it is implicit in the topics studied by most social psychologists -- with vastly more research on process, error, and bias than on accuracy.

Psychological research articles are filled with excellent experimental studies of cognitive processes that researchers interpret as suggesting that bias, error, and self-fulfilling prophecy is likely to be common in daily life. But such generalizations are only justifiable by research that examines the accuracy of people's judgments in real-world contexts, not in artificial or even realistic laboratory contexts. No matter how much researchers think the processes discovered in the lab should lead to bias and error, the only way to find out for sure would be by assessing the accuracy of real social perceptions. A social perceiver whose beliefs closely correspond to social reality is accurate, regardless of the processes by which that perceiver arrived at those beliefs. Thus, although there are many good arguments to study process, none constitute good arguments not to study accuracy.

Accuracy of explanations: "Just because it can be shown that some belief about some person or group is correct does not tell us why or how the person or group got that way." The dismissal of accuracy as something uninteresting or unimportant is often implicit in perspectives arguing that social processes and phenomena (e.g., discrimination, poverty) create the differences that are perceived (e.g., Fiske, 1998; Jost & Banaji, 1994). Social processes undoubtedly create many group and individual differences. Nonetheless, this sort of analysis, which emphasizes the explanations for the origins of group and individual differences fails to threaten or undermine the viability of accuracy research. Both points are next illustrated with a hypothetical example.

Let's say that Ben believes Joe is hostile. This "objection" focusing on the accuracy of explanations leads to at least four different questions: 1) Is Ben right? 2) What is Ben's explanation for Joe's hostility? 3) If Joe is hostile, how did he get that way? and 4) Why does Ben believe Joe is hostile?

Providing an answer to one question provides no information about the others. For example, establishing that Ben is correct (Joe really is hostile) tells us nothing about how Ben explains Joe's hostility. Nor does it
provide any information on how Joe actually became hostile. Ben’s belief in Joe’s hostility can be accurate and his explanation inaccurate. Of course the lack of information about answers to other questions constitutes no fatal flaw, indeed, no limitation at all, to the assessment of the accuracy of Ben’s belief in Joe’s hostility. Indeed the latter two questions (how did Joe get that way, and how did Ben come to believe Joe is hostile) are not even accuracy questions; they are process questions. Thus, failure to explain how a person or group develops some characteristic constitutes no threat to accuracy research.

Accuracy vs. self-fulfilling prophecy: "Prior self-fulfilling prophecies may influence that which is accurately perceived.” The logic underlying this objection seems to be the following: 1. Self-fulfilling prophecies occur. 2. Sometimes differences between targets reflect self-fulfilling prophecies. 3. If so, attributing "accuracy" to those perceptions is, at best, meaningless, and, at worst, reifies differences produced through social processes (Claire & Fiske, 1998; Fiske, 1998).

The first two premises are true. Self-fulfilling prophecies do indeed occur sometimes; and, at any point in time, the differences between targets may indeed reflect self-fulfilling prophecies to some extent. Thus, differences that are accurately perceived at some point in time may reflect effects of prior self-fulfilling prophecies.

Nonetheless, the conclusion that this renders accuracy research meaningless is unjustified for several reasons. First, if a perceiver cannot have caused differences among targets, self-fulfilling effects of that perceiver's expectations cannot account for those differences. If, by the time Johnny gets to fourth grade, his performance in school is stellar, should his teachers reduce his grades from A’s to B’s because part of his performance resulted from self-fulfilling prophecies in prior years? That would be silly. When a perceiver's judgments closely correspond to targets' attributes, and when that same perceiver's expectations cannot have caused those attributes, how shall we refer to this correspondence? There is only one viable answer: accuracy.

But the argument that accuracy is meaningless because self-fulfilling prophecies may cause that which is “accurately” perceived fails even if, through self-fulfilling prophecies, the same perceiver did cause the target’s behavior or accomplishment. The key issue here is time. If a perceiver’s expectations trigger a
social interaction sequence that causes target to become a very pleasant person, those expectations (which came prior to the interaction) are self-fulfilling. But, once the interaction is over, how should the target be perceived? Would it be most accurate to perceive the target as nasty, neither nasty nor pleasant, or as pleasant? Again, the answer is obvious. A "problem" arises only when we fail to account for the difference between predictions (which may be either self-fulfilling or accurate) and impressions of past behavior (which can only be accurate or inaccurate, and, by virtue of referring to behavior that has already occurred, cannot be self-fulfilling). Of course, today's impressions can become tomorrow's (self-fulfilling) predictions.

It is completely true that prior self-fulfilling prophecies may influence that which is subsequently accurately perceived. This is interesting and important, but fails to constitute a threat or obstacle of any kind to assessing the accuracy of those perceptions.

The criterion “problem.” The criterion "problem" has been one of the most common objections appearing in the literature criticizing accuracy research (e.g., Fiske, 1998; Jones, 1990; Schneider, et al., 1979; Stangor, 1995). Many prominent researchers have declared or strongly implied that it is difficult or impossible to identify criteria to assess the accuracy of social beliefs:

"The naiveté of this early assessment research was ultimately exposed by Cronbach's elegant critique in 1955. Cronbach showed that accuracy criteria are elusive and that the determinants of rating responses are psychometrically complex” (Jones, 1985, p. 87).

“Even if I thought it were desirable or important to catalog the accuracy of social stereotypes, I would be pessimistic about our ability to make definitive statements in this regard. This is because I believe the prognosis for developing unambiguous criteria on which to make such statements is small.” (Stangor, 1995 p. 282).

“In any event, what does it mean to say that, “actually,” women are dependent, men are aggressive, Jews are stingy, the elderly are conservative, blacks are criminal, or whites are conceited? The problem of the actual criterion is complex, especially for traits (Judd & Park, 1993). The target group’s self-report is a common criteria, but this is plagued by various self-report biases and sample selection biases. Also, the validity of self-reports is affected by group identity issues (Judd, et al. 1995). Another plausible criterion
would be “objective” measures, but their validity, too, is unclear. What measure would objectively indicate whether a group is ambitious, lazy, or efficient? And how ambitious is ambitious? And for what proportion of the group, compared to what other group, does the trait have to hold? Expert judgments are possible, but they themselves are not immune to stereotypes.” (Fiske, 1998, p. 382).

I address criteria later in this précis. For now, however, several aspects of these perspectives are worth noting. Jones’s (1985) citation of Cronbach (1955) in support of the argument that “accuracy criteria are elusive” is particularly odd, because Cronbach (1955) did not address the issue of criteria. The passage from Fiske (1998) is also revealing. Why are both “actually” and “objective” in quotes? The implication seems to be that there is little or no “actually” or “objectivity” out there. The quote is largely a series of rhetorical questions that are plausibly interpreted as implying, without quite stating, that “it is impossible to answer these questions because there are no good criteria.”

Furthermore, none of these articles identify a single criterion that the authors do consider appropriate to use to study accuracy. This leaves the reader with either blanket dismissals of criteria (Jones, Stangor), or a long list of unacceptable criteria, and no identified acceptable criteria (Fiske). Indeed, it is not clear how to avoid the interpretation that this scholarship means that there are no good criteria for assessing accuracy. If this is not what these and other authors mean when they provide blanket dismissals of accuracy criteria, it would be invaluable for them to describe what criteria they do consider to be appropriate. Next, therefore, I consider the scientific justifiability for such blanket dismissals of criteria for accuracy.

Psychologists – including all three quoted here -- routinely engage in the scientific study of one or more of the following attributes: aggression, political attitudes, generosity, intelligence, achievement, morality, motivation, and even conceit (aka “self-serving bias”). Who would study political attitudes or achievement (etc.) without believing such constructs “really exist”? I have not found any scholarship from these same authors generally arguing that motivation, generosity, attitudes, etc., cannot be assessed in other, non-accuracy-related, contexts. It is hard to avoid the implication from this line of argument dismissing accuracy criteria that these constructs cannot be assessed when studying accuracy, but they can be assessed in other types of psychological research. At minimum, the logical bases for such an argument have never previously
been articulated. Furthermore, if psychological constructs such as motivation, attitudes, generosity, etc., can be studied in other contexts, then it would seem there are good criteria for establishing the accuracy of social beliefs, because they would be the very same criteria that psychological scientists use to establish the reality of the constructs they study. Attempts to dismiss the appropriateness of criteria for studying the accuracy, say, of lay beliefs about individuals’ or groups’ motivation (laziness), attitudes (conservatism), charitable giving (stinginess), etc., would appear to be logically compelled to similarly dismiss the appropriateness of using the same criteria to study, say, the accuracy of psychologists’ hypothesis about motivation, attitudes, charitable giving, etc.

Logical issues with the dismissal of criteria for assessing accuracy are highlighted even more starkly when raised by psychologists who emphasize the power and importance of self-fulfilling prophecies, including some by the very same authors raising the criteria issue for accuracy (e.g., Fiske, 1998; Jones, 1986). Although the processes by which perceivers' beliefs become valid are different for self-fulfilling prophecies and accuracy, the criteria for establishing their validity must be identical. When assessing both self-fulfilling prophecies and accuracy, the question is "to what extent does the expectation correspond to the outcome?" How it can be impossible to identify criteria for establishing accuracy and unproblematic to identify criteria for establishing self-fulfilling prophecy, when both require establishing correspondence between social perceptions and social realities, has never been articulated.

Criteria and Construct Validity

Accuracy’s inherent kinship with construct validity. Understanding what criteria exist to assess accuracy requires first defining accuracy. The approach taken here is probabilistic realism. Probabilistic realism assumes that there is an objective reality, and that, flawed and imperfect though we may be, we can eventually come to know or understand it, at least much of the time (in the book, this perspective is contrasted with functional and social constructivist perspectives on accuracy).

Social perceptual accuracy is correspondence between perceivers' beliefs (expectations, perceptions, judgments, etc.) about one or more target people and what those target people are actually like, independent of perceivers' influence on them. More correspondence without influence, more accuracy.
Identifying criteria for accuracy can be approached much as establishing construct validity of criteria against which researchers test testing psychological hypotheses, which then addresses many of the doubts and criticisms (Jones, 1985; Fiske, 1998; Stangor, 1995). Finding criteria for assessing the accuracy of social beliefs is virtually identical to finding criteria for assessing the accuracy of social psychological hypotheses. Indeed, as shall be shown next, the construct validity of the criteria used in accuracy research has often been far more strongly established than that used in much social psychological research, which often involves measures made up on the fly for particular studies.

Criteria. Types of criteria that have been productively used in accuracy research are, therefore, essentially the same as used in other research to test psychological hypotheses (objective criteria, behavior, agreement with experts, agreement with other perceivers, agreement with targets self-reports and self-perceptions). Criteria are objective when that which is being judged is assessed in a standardized manner that is independent of the perceiver's judgment. Examples of objective criteria that have been used in accuracy research are Census data, most sports outcomes, cognitive ability tests, and meta-analyses of group differences. Objective criteria may indeed have imperfections, but they are evidence assessed in standardized manners independent of perceivers’ judgments. For example, consider Ali, who predicts that Derek Jeter will hit a home run in his last at bat at Yankee stadium. He will be either right or wrong about this. There is nothing the least bit difficult or "problematic" about this. Although the rules of baseball can only be established through agreement, once established, the criteria for hits, home runs, etc. are mostly independent of human judgment. The role of umpires is primarily to exercise subjective judgment for (the relatively few) close calls, to prevent unruly or aggressive behavior, and to enforce the more esoteric rules of the game.

Similarly objective criteria – such as Census data about the proportions of people with high school degrees or on welfare, and meta-analyses of group differences are also useful as criteria precisely because, whatever their imperfections, they are standardized and independent of the judgments of perceivers in any particular study. Not all people may agree that certain objective criteria are good ones. Such agreement might be irrelevant regarding, say, guessing targets’ number of children, but they become much more
relevant when estimating, say, extraversion or intelligence via a personality questionnaire or standardized IQ test. Is the personality questionnaire a good one? Is it reliable? Valid? IQ tests, in particular, have a long and controversial history (e.g., Gould, 1981; Herrnstein & Murray, 1994; Neisser, et al., 1996).

To the extent that some people do not find such tests credible, they are likely to discredit or dismiss research on accuracy using such criteria. Thus, use of objective but controversial criteria can be viewed as boiling down to agreement (if you agree with the criteria, the study assesses accuracy; if you do not agree with the criteria, it does not -- see Kruglanski, 1989). And socially and politically, this is probably how things work. People who do not accept one’s criteria most likely will not accept one’s conclusions (whether on accuracy or any other social science topic).

Often, however, what may happen is the reverse: People who do not like scientific conclusions will come up with arguments against the appropriateness of using criteria involved in those conclusions. This may help explain why social psychologists were much more critical of the criteria used in accuracy research than in self-fulfilling prophecy research, even when the criteria were identical. A similar analysis could be presented for cognitive ability tests. Indeed, cognitive ability tests are among the most highly validated measures in all of psychology, predicting important life outcomes such as educational attainment, income, and criminality (e.g., Neisser, et al., 1996; Schmidt & Hunter, 1998). The grounds for arguing that such tests are somehow invalid on the part of any psychologists who have used measures developed on the fly (i.e., subject to little or no validity assessment) for a particular research purpose, but at the same time, believes the on-the-fly measures constitute appropriate criteria for assessing the validity of scientific hypotheses, has never been articulated.

VI. The Accuracy of Teacher Expectations

Having established the scientific appropriateness and viability of studying social perceptual accuracy, it was then possible to revisit some of the clearest evidence that bore on the accuracy question – which, ironically (given that it kicked off social psychology’s infatuation with expectancy effects), was teacher expectation research. First, teachers’ expectations are generally heavily based on students prior grades and standardized test scores, with multiple correlations often in the .6 to .8 range (Jussim et al., 1996). In
contrast, demographic variables, such as race, gender, and social class often have no predictive value (after controlling for prior achievement), and rarely have effects exceeding standardized coefficients of .15 (Jussim et al., 1996; Madon et al., 1998; Williams, 1976).

Furthermore, the main reason teacher expectations predict student achievement is because they are accurate, not because they are self-fulfilling or biasing. Correlations of teacher expectations with student achievement typically range from about .4 to .8, whereas bias and self-fulfilling prophecy effects are typically no larger than .10 to .20 each. The difference between the correlation and the teacher expectation effect can be used as an estimate of accuracy because it constitutes predictive validity without (self-fulfilling) influence. This means that accuracy consistently accounts for about 60-70% of the relationship between teacher expectations and student achievement with the remaining 30%-40 divided among bias and self-fulfilling prophecy (see Jussim & Eccles, 1995; Jussim et al, 1996; Jussim & Harber, 2005 for reviews).

VII. The Unbearable Accuracy of Stereotypes

Are stereotypes inaccurate? The assumption or definition of stereotypes as inaccurate has long and deep roots in psychology (see reviews by G. Allport, 1954/1979; Ashmore & Del Boca, 1981; Brigham, 1971; Jussim, 2012). Because some have argued that assessing stereotype accuracy may be impossible or undesirable (Fiske, 1998; Stangor, 1995), the first order of business is to address when assessment of stereotype accuracy is scientifically possible.

First, only descriptive or predictive beliefs can be evaluated for accuracy. “Jews are richer than other Americans” can be evaluated for accuracy; the accuracy of “I like (dislike) Jews,” however psychologically important, cannot be evaluated for accuracy. Stereotypes as prescriptive beliefs, too, cannot be evaluated for their accuracy. Accuracy is irrelevant to notions such as “children should be seen and not heard” or “men should not wear dresses.” Therefore, to the extent that stereotypes are defined as something other than descriptive or predictive beliefs, one is precluded from making any claim about inaccuracy.

The assumption that stereotypes are inaccurate is only relevant to descriptive or predictive beliefs and, therefore, can mean only one of two things:

1. All such beliefs about groups are stereotypes and all are inaccurate
or

2. Not all beliefs about groups are inaccurate, but stereotypes are inaccurate beliefs about groups.

Why each is logically incoherent is discussed next.

The Logical Incoherence of Defining Stereotypes as Inaccurate

A claim that all beliefs about all groups are inaccurate is logically incoherent. It would mean that:

1. Believing that two groups differ is inaccurate; and 2. Believing two groups do not differ is inaccurate.

Both 1 and 2 are not simultaneously possible, so we can reject any claim that all beliefs about groups are inaccurate.

If stereotypes are the subset of beliefs about groups that are inaccurate, then only inaccurate beliefs about groups can be considered stereotypes. Accurate beliefs about groups have been defined away as not stereotypes. This has the (probably unintended) effect of defining away nearly all existing research on stereotypes. Why? Because vanishingly few studies of stereotypes have actually first demonstrated that the beliefs about groups under study are inaccurate. Holding social psychology to this interpretation of “stereotypes are inaccurate” means concluding that decades of research framed as addressing stereotypes really has not. There would be no studies of the role of stereotypes in expectancy effects, self-fulfilling prophecies, person perception, subtyping, memory, etc.

There are additional logical problems with defining stereotypes as inaccurate. No scholarship that has done so has also identified the point at which a belief crosses over from being an “accurate” belief about a group, to being a “stereotype.” Absent a standard for (in)accuracy, this means that we cannot know whether any belief is a (defined as inaccurate) stereotype. Similarly, if one claims that accuracy cannot or should not be assessed, or that existing research fails to validly assess accuracy (Fiske, 1998, 2004; Stangor, 1995), one has dismissed all evidence that bears on accuracy and therefore precluded one’s self from making any statements about stereotypes’ (in)accuracy. In summary, defining stereotypes as inaccurate is severely problematic no matter what the definer means. Any scientist who wishes to maintain such a definition needs to precisely articulate how each of these forms of logical incoherence have been overcome.

A Viable, Logically Coherent Definition
I concur with the minority of scientists who have left inaccuracy out of the definition of stereotype (e.g., Ashmore & Del Boca, 1981; Judd & Park, 1993; Ryan, 2002), and who have generally defined stereotypes as beliefs about the attributes of social groups. This allows for many possibilities not explicitly stated.

Stereotypes may or may not:

- be accurate and rational
- be widely shared
- be conscious be rigid
- exaggerate group differences
- assume group differences are essential or biological
- cause or reflect prejudice and discrimination
- cause biases and self-fulfilling prophecies
- play a major role in some social problems.

This definition retrieves accuracy from premature foreclosure by definition and turns it into a scientific empirical question. How well do people’s beliefs about groups correspond to what those groups are actually like?
The Rigorous Assessments of Stereotype (In)Accuracy

To be included here, empirical studies assessing the accuracy of stereotypes needed to meet two major criteria. First, they had to relate perceivers’ beliefs about a target group with some measure of what that group was actually like. This may seem obvious, but the social psychological discourse on stereotypes has often drawn conclusions about the inaccurate or unjustified nature of stereotypes based entirely on evidence addressing social cognitive processes – illusory correlations, priming, expectancy effects, attributional patterns, etc. Such research, although important on its merits, does not directly address accuracy, which can only done by comparing beliefs about groups to criteria regarding those group’s characteristics.

Second, studies needed to use an appropriate target group. If the stereotype is of “American women,” the target group should be a representative sample of American women; it cannot be a convenience sample (Judd & Park, 1993). Studies that met both of these criteria were included; those that did not were excluded.

Four Types of Stereotype (In)Accuracy

Accuracy is often a multidimensional construct (e.g., Judd & Park, 1993; Kenny, 1994), as can be readily illustrated with a simple example. Consider Fred, judging the average height of male Americans, Columbians, and Dutch. Fred estimates the average heights, respectively, as 5’8, 5’5”, and 5’10”. Let’s say the real average heights are, respectively, 5’10”, 5’7”, and 6’0”. In absolute terms, Fred is inaccurate – he consistently underestimates height by two inches. However, in relative terms, Fred is perfectly accurate – his estimates correlate 1.0 with the actual heights. Although Fred has a downward bias in perceiving the absolute heights among men in the different countries, he is superb at perceiving the relative height differences.

Discrepancy from perfection refers to how close people’s beliefs about groups are to those groups’ actual mean characteristics on criteria. These are assessed with discrepancy scores. Correspondence with differences refers to how well people detect either variations between or within groups on some set of attributes. These are assessed with correlations between beliefs and criteria. Personal stereotypes are the beliefs about groups held by a particular individual. Consensual stereotypes are the overall, or average, beliefs about a group held by some group of perceivers. This creates four types of stereotype (in)accuracy:
**Personal discrepancies** (how discrepant a single person’s stereotypes are from a criterion); **consensual discrepancies** (how discrepant a sample’s or group’s mean stereotypes are from a criterion); **personal correspondence** (the correlation of a single person’s stereotypes with criteria); and **consensual correspondence** (the correlation of a sample’s or group’s stereotypes with criteria).

**What is a Reasonable Standard for Characterizing a Stereotypic Belief as “Accurate”?**

Discrepancies from perfection (discrepancy scores) and correspondence with real differences (correlations) capture different but important aspects of accuracy. Because both can and have been used to assess accuracy, there needs to be two separate standards for characterizing a belief as accurate – one for discrepancy scores, and another for correlations.

**Discrepancies.** In this review of the empirical evidence assessing the accuracy of stereotypes, beliefs that are within 10% of the criterion are characterized as **accurate**; beliefs that are more than 10% off, but 20% or less off, as “near misses,” and beliefs that are more than 20% off as **inaccurate**. For studies that do not report their results as percentages, effect sizes of \( d = .25 \) are used as the cutoff for accuracy because it corresponds to an approximately 10% difference. These cutoffs are appropriate, perhaps even stringent, because for most practical and even scientific purposes, predictions no more than 10% off correspond to many pre-existing high standards (e.g., 90% or more on a test is usually an A; a researcher who predicts an effect of .30, but obtains one of .27, will generally see this result as supporting the hypothesis). Nonetheless, 10% is somewhat arbitrary and, for certain purposes, different criteria for accuracy might be appropriate.

**Correspondence.** Effect sizes of \( d = .8 \) have, by longstanding convention (J. Cohen, 1988), been characterized as “large.” This corresponds approximately to a correlation of .40 between belief and criteria. I therefore use \( r = .40 \) as the cutoff for considering a stereotype to be accurate. Similarly, \( d \)'s of .5 are considered “moderate” so that \( r = .25 \) as the cutoff for “moderate accuracy.”

**Pervasive Stereotype Accuracy**

I use the term “pervasive stereotype accuracy” here to refer to the widespread evidence of at least some accuracy, and sometimes quite high accuracy, found in nearly every study that has assessed stereotype accuracy. Nonetheless, there is no evidence that stereotypes are **perfectly** accurate. Furthermore, even
within a single study, accuracy levels may vary, not just across judgments or perceivers, but, because accuracy is a multi-faceted construct, across the four types of accuracy described previously. The few studies of highly inaccurate stereotypes are also reviewed. Nonetheless, the evidence of pervasive stereotype accuracy is inconsistent with virtually all perspectives defining stereotypes as inaccurate, or emphasizing their inaccuracy.

**Ethnic and racial stereotypes.** Table 2 summarizes the results of all studies assessing the accuracy of racial/ethnic stereotypes that met the criteria for inclusion. Consensual discrepancies are mostly accurate or near misses. For example, in McCauley & Stitt (1978), out of 70 judgments about Americans (in general) and African-Americans, 34 were accurate, and another 30 were near-misses. Ashton & Esses (1999) found a similar pattern in the consensual ethnic stereotypes held by Canadian college students’ judgments about the academic achievement of eight ethnic groups were accurate and one was a near-miss. Ryan (1996) found evidence of both accuracy and inaccuracy in African-American and White college students’ consensual racial stereotypes (over all six perceiver group-target group combinations (African-American and White perceivers making judgments about African-Americans, Whites, and their differences) 34 judgments were accurate, 20 were near misses, and 48 were inaccurate.

Furthermore, the results from these studies provide little consistent support for the idea that stereotypes exaggerate real differences. Exaggeration of real differences occurred more often than underestimations in some studies (Ashton & Esses, 1999; Ryan, 1996), but underestimation occurred more often in others (McCauley & Stitt, 1978; Wolsko, Park, Judd, & Wittenbrink, 2000). Even the evidence of exaggeration, however, was more mixed than this summary suggests. For example, the only study to assess the accuracy of personal discrepancies (Ashton & Esses, 1999) found that a plurality of people were generally accurate (n=36), and that more exaggerated (n=34) than underestimated (n=25). If stereotypes are defined as exaggerations of group differences, a definition I reject but which has deep roots in social psychology (e.g., Allport, 1954/1979; Campbell, 1967), then in the Ashton & Esses (1999) study, 61 of 95 people did not hold stereotypes.

Stereotype accuracy as correspondence between belief and criteria was generally very strong. Consensual
stereotype accuracy correlations ranged from .53 to .93. Personal stereotype accuracy correlations ranged from .36 to .69. Although the participants in these studies were only fairly good at identifying the precise level of some characteristic of racial and ethnic groups, their perceptions of differences both within and between groups across the different attributes were quite high.
Gender stereotypes. Table 3 summarizes the results of studies of gender stereotypes accuracy. In most cases, at least a plurality, and often a majority, of consensual stereotype judgments were accurate, and accurate plus near miss judgments predominate in every study. For example, in the Swim (1994) study, of 33 judgments, 18 were accurate and 7 were near misses. There was no support for the hypothesis that stereotypes generally lead people to exaggerate real differences. As with race, underestimations counterbalanced exaggerations.

Across the studies summarized in Table 3, consensual stereotype accuracy correlations were quite high, ranging from .34 to .98, with most falling between .66 and .80. The results for personal stereotypes were more variable. Once they were inaccurate, with a near zero correlation with criteria (Beyer, 1999, perceptions of female targets). In general, though, they were at least moderately, and sometimes highly accurate (most correlations ranged from .40 to .60 -- see Table 3). A recent multi-national study found a fundamentally similar pattern, with consensual gender stereotype accuracy correlations ranging from .36 to .70, and showing no evidence of exaggeration (Löckenhoff et al, in press).

Other stereotypes. Empirical research on stereotype accuracy has also addressed a wide variety of other stereotypes (e.g., occupations, college majors, etc., see Table 4), and found essentially the same broad and general patterns as obtained for race, ethnicity, and gender: high levels of accuracy and, little or no general tendency to exaggerate real differences.

Inaccurate Stereotypes

Despite the impressive and surprising evidence of the accuracy of stereotypes, there is some evidence of inaccurate stereotypes. In the U.S., an early study found little accuracy in political stereotypes (Judd & Park, 1993). More recent research on the stereotypes of the moral beliefs held by liberals and conservatives, found a more mixed picture, of both accuracy and exaggeration (Graham, Nosek, & Haidt, 2013).

A large scale study conducted in scores of countries found that there is also little evidence of accuracy in national stereotypes regarding personality (Terracciano, et al, 2005). However, Heine, Buchtel, & Norenzayan (2008) found that, when behavioral rather than self-report data were used as the criteria, far more evidence of stereotype accuracy emerged for the conscientiousness factor (correlations between
consensual stereotypes and behavior averaged about .60). A recent replication (McCrae et al, in press) addressing many of the issues raised by Heine et al (2008), but still using self-reports as criteria, again showed almost no accuracy in national character stereotypes. The fairest conclusion, therefore, seems that the (in)accuracy of national character stereotype remains contested and unresolved.

Strengths and Weaknesses of Research on the Accuracy of Racial, Ethnic, and Gender Stereotypes

Stereotype accuracy correlations are among the largest and most replicable effects in all of social psychology (see Table 5), and are typically far larger than the effect sizes routinely interpreted as support for more famous social psychological hypotheses. Several methodological aspects of these studies are worth noting because they bear on the generalizability of these results. Only one (Judd & Park, 1993) was based on a nationally representative sample, so that, although evidence of stereotype accuracy is common in the data, the generalizability of those findings is currently unknown. Although most studies assessed the accuracy of undergraduates’ stereotypes, several assessed the accuracy of samples of adults (see Tables 2 and 3). Some of the highest levels of accuracy occurred with these adult samples, suggesting that the levels of accuracy obtained do not represent some artifact resulting from the study of undergraduate samples.

Second, the studies used a wide variety of criteria: U.S. Census data, self-reports, Board of Education data, nationally representative surveys, locally representative surveys, U.S. government reports, etc. The consistency of the results across studies, therefore, does not reflect some artifact resulting from use of any particular criteria.

Third, the studies examined a wide range of stereotype content: beliefs about demographic characteristics, academic achievement, personality and behavior. The consistency of the results across studies, therefore, does not reflect some artifact resulting from a particular type of stereotype content.

Fourth, personal discrepancies were the least studied of the four types of accuracy. Thus, the studies do not provide much information about the extent to which individual people’s stereotypes deviate from perfection. Because of wisdom of crowd effects (Surowiecki, 2004), it is likely that consensual discrepancies are more accurate than individual discrepancies, though identifying and understanding sources of individual discrepancies remains an important area for future research.
VIII. Stereotypes and Person Perception

People should primarily use relevant individuating information, when it is available, rather than stereotypes when judging others, because usually, relevant individuating information will be more diagnostic than stereotypes (though not always, see Crawford, Jussim, Madon, Cain, & Stevens, 2011). This area of research has been highly controversial, with many researchers emphasizing the power of stereotypes to bias judgments (Devine, 1995; Fiske & Neuberg, 1990; Fiske & Taylor, 1991; Jones, 1986; Jost & Kruglanski, 2002) and others emphasizing the relatively modest influence of stereotypes and the relatively large role of individuating information (Jussim et al, 1996; Kunda & Thagard, 1996).

Fortunately, multiple meta-analyses have been performed addressing these issues (see Table 1). The effects of stereotypes on person judgments, averaged over hundreds of experiments, range from 0 to .25. The simple arithmetic mean of the effect sizes in Table 1 is .10.

Furthermore, people do generally rely heavily on individuating information. The one meta-analysis that has addressed this issue found that the effect of individuating information on person perception was among the largest effects found in social psychology, $r = .71$ (Kunda & Thagard, 1996). People seem to be generally doing what most social psychologists say they should do – they rely on individuating information far more than stereotypes.

But what about the .10 effect of stereotypes? Doesn’t that demonstrate inaccuracy? It generally does, at least when the stereotype itself is clearly inaccurate.

But as has just been shown, the empirical research demonstrates considerable accuracy in many stereotypes that have been studied. Therefore, one cannot assume that stereotypes are inaccurate, absent data demonstrating inaccuracy. This means that the .10 bias effect does not necessarily demonstrate inaccuracy for two reasons:

1) Most of the studies examining these issues have examined experimentally created fictitious targets who had no “real” attributes, so that there was no criteria with which to assess accuracy; and 2) None first demonstrated that the stereotype under study was inaccurate.

All of which raises the question: Can the “biasing” effects of stereotypes increase the accuracy of person
perception?

Stereotype “Biases” That Increase Accuracy

When individuating information is unavailable or ambiguous, reliance on an accurate stereotype to make a guess or inference about a target will lead one to be as accurate as possible (e.g., Jussim, 1991) and more accurate than will ignoring the stereotype. Whether this is morally desirable and whether acting on such judgments is always legal is a complex issue addressed in the book.

When stereotypes have substantial degrees of accuracy, then, in the absence of completely diagnostic individuating information, people will generally arrive at more accurate judgments when using than when ignoring their stereotypes. This is exactly what has been found in studies of occupational stereotypes (C. Cohen, 1981), college dorm stereotypes (Brodt & Ross, 1996), role stereotypes (Macrae, Milne, & Bodenhausen, 1994), and gender stereotypes (Gosling, Ko, Mannarelli, & Morris, 2002; Jussim, et al, 1996; Madon, et al, 1998). Research published after Social Perception and Reality similarly shows that reliance on stereotypes regarding mothers increases empathic accuracy (Lewis, Hodges, Laurent, Srivastava, & Biancarosa, 2012), and that reliance on gay stereotypes increased perceiver’s ability to identify whether a target was gay from a photograph of the face alone, if the gay target appeared feminine, but not if the gay target appeared masculine (Stern, West, Jost, & Rule, 2013).

Stereotype “biases” sometimes increase rather than decrease the accuracy of person perception. Because only a handful of studies have addressed this issue, psychologists are not yet in a position to reach broad conclusions about the generality or pervasiveness of this pattern. Nonetheless, the data that has actually assessed person perception accuracy contrast sharply with the common interpretation in social psychology that any influence of a stereotype on person perception is an unjustified distortion (Brown, 2010; Fiske, 1998; Stangor, 1995).

Why Is The Evidence Of Stereotype Accuracy And Rationality Important And Useful?

1. Claims, such as “stereotypes are inaccurate” and “it is hard to get people to individuate” are unjustified by existing data. Unqualified claims that stereotypes “exaggerate real differences” are not justified. The consensual stereotypes that have been studied demonstrated extraordinarily high levels of accuracy, so
that unqualified suggestions that stereotypes are false cultural myths are not justified. Stereotype accuracy has been obtained by multiple independent research teams. It is also one of the largest effects in all of social psychology.

2. Stereotypes are a central component of how people think about other people. This is part of many core definitions of social psychology and social cognition.

3. Allowing for the possibility that some stereotypes may have some degree of accuracy leads to a coherent understanding of past and future research. Absent a recognition that on both logical and empirical grounds stereotypes may be accurate, past research will be haunted by a scientifically incoherent definitional tautology, i.e., that people who believe in stereotypes are in error because stereotypes are erroneous beliefs.

4. Recognizing the existence of stereotype accuracy raises interesting and important theoretical and empirical questions. When do stereotypes flexibly change in response to changes in social reality and when do they remain rigidly resistant to change? When do people’s person perception judgments correspond to, or deviate from, Bayesian rationality? Are the deviations predictable from implicit or explicit prejudice? What characteristics of perceivers and targets (both individuals and groups) moderate degree of (in)accuracy?

5. Acknowledging accuracy and rationality in stereotypes and stereotyping neither contests nor diminishes the importance of scientific research on stereotype biases or sources of oppression or inequality. Accuracy and bias can and often do co-exist (Jussim, 1991, 2012). Accuracy, construed here as a question of degree, rather than something absolute, leaves open ample room for inaccuracy and bias. Demonstrating accuracy rarely precludes the possibility of bias, even socially important ones; demonstrating bias does not preclude high levels of accuracy. And, perhaps even more important, if stereotypes are often reasonably accurate, it highlights the question: what other phenomena create or maintain inequality?
IX. Conclusions

Unfortunately, space considerations precluded addressing important areas of research, such as detailed evaluations of componential (e.g., Cronbach, 1995; Judd & Park, 1993; Kenny, 1994) versus noncomponential (e.g., Brunswik, 1952; Dawes, 1979; Funder, 1995) approaches to the assessment of accuracy, and a fuller consideration of confirmation, disconfirmation, and diagnosticity in lay hypothesis testing (e.g., Klayman & Ha, 1987). Similarly, this précis has not addressed accumulation of self-fulfilling prophecies across perceivers (Jussim et al, 1996), or the role of parents (e.g., Madon, Guyl, Spoth, & Willard, 2004), rejection sensitivity (e.g., Downey, Freitas, Michaelis, & Khouri (1998), stereotype threat (Steele & Aronson, 1995), and unconscious/automatic processes in producing self-fulfilling prophecies (Chen & Bargh, 1997). The role of expectations in attributional (e.g., Kulik, 1983) and memory biases (e.g., Stangor & McMillan, 1992), too, have been largely omitted here, as has been a broader discussion and critical evaluation of the role of labeling in person perception bias (e.g., Harris, Milich, Corbitt, Hoover, & Brady, 1992). These studies, issues and phenomena are all quite important on their merits, and are addressed in Social perception and social reality. The empirical evidence generally provides further support for the tripartite conclusion reached here.

The “Story”

Social psychologists have long emphasized expectancy effects – both biases and self-fulfilling prophecies – as playing a major role in social, educational, and economic inequality. Teacher expectations supposedly disadvantage students from already disadvantaged backgrounds and advantage students from advantaged backgrounds. Because stereotypes are, the story goes, so widely shared and so widely inaccurate, their powerfully self-fulfilling effects will accumulate over time and across perceivers. Because self-fulfilling prophecies are so consistently harmful, damaging self-fulfilling prophecy on top of damaging self-fulfilling prophecy will be heaped upon the backs of those already most heavily burdened by disadvantage and oppression.
The Inadequacy of the “Story”

The most benevolent interpretation is that this story is woefully incomplete. Cognitive biases do sometimes lead to expectancy confirmation and expectancies do sometimes lead to self-fulfilling prophecies. But the power of expectations to distort social beliefs through biases, and to create actual social reality through self-fulfilling prophecies is, in general, so small, fragile, and fleeting that it is quite difficult to make an empirical case that such effects constitute a major source of inequality. Hundreds of studies show that biasing effects of expectations and stereotypes on person perception hover barely above zero (see Table 1), that self-fulfilling prophecy effects are often modest and fleeting, and that some of the largest self-fulfilling prophecy effects ever obtained increased rather than decreased the performance of low achieving students. “The story” can be maintained primarily by selectively overlooking this abundant evidence of weak effects, and by speculative arguments about the implications of existing data.

Concern for combating oppression has inspired a great deal of important research that has yielded profound insights into stereotypes, prejudice, and discrimination. This includes the abundant research on biases and self-fulfilling prophecies. Nonetheless, the evidence overwhelmingly supports the tripartite pattern: 1. Although errors, biases, and self-fulfilling prophecies in person perception, are occasionally powerful, on average, they tend to be weak, fragile and fleeting; 2. Perceptions of individuals and groups tend to be at least moderately, and often highly accurate; and 3. Conclusions based on the research on error, bias, and self-fulfilling prophecies routinely overstate their power and pervasiveness, and consistently ignore evidence of accuracy, agreement, and rationality in social perception.

Accuracy Dominates Bias and Self-Fulfilling Prophecy

Social perception and social reality did not focus on many topics that strongly makes the case for reasonableness and accuracy in social perception – such as empathic accuracy, accuracy based on thin slices of behavior, and demonstrations that perceptions and judgments are often approximately Bayesian (Ambady, Hallahan, & Conner, 1999; Griffiths & Tenenbaum, 2006; Ickes, 1995). Instead, it focused on self-fulfilling prophecies, interpersonal expectancies, and stereotypes precisely because even those areas – which have
long held a central place in emphases on error and bias -- typically provide far more evidence of reasonableness, rationality, and accuracy than they do of error, bias, and social constructionism.

Social perceptions can construct social realities. People are indeed subject to all sorts of imperfections, errors, and biases. Occasionally, such effects are quite large. Sometimes, such effects can have important effects on targets’ lives. In general, however, the evidence to date shows that they are generally weak, fragile, and fleeting, and that many social perceptions, including social stereotypes, are often more heavily based on social reality than they distort or create such realities.

References


doi:10.1371/journal.pone.0050092.


Table 1
AVERAGE EXPECTANCY EFFECT SIZES TYPICALLY RANGE FROM SMALL TO MODERATE

<table>
<thead>
<tr>
<th>Meta-Analysis</th>
<th>Topic/Research Question</th>
<th>Number of Studies</th>
<th>Average Expectancy Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SELF-FULFILLING PROPHECY:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rosenthal &amp; Rubin (1978)</td>
<td>Do interpersonal expectations create self-fulfilling prophecies?</td>
<td>330</td>
<td>.29&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>Raudenbush (1984)</td>
<td>Do teacher expectations have self-fulfilling effects on student IQ?</td>
<td>18</td>
<td>.06</td>
</tr>
<tr>
<td>McNatt (2000)</td>
<td>Do manager's expectations have self-fulfilling effects on employees' performance?</td>
<td>6</td>
<td>.23</td>
</tr>
<tr>
<td>McNatt (2000)</td>
<td>Do military officers' expectations have self-fulfilling effects on trainees?</td>
<td>11</td>
<td>.52</td>
</tr>
<tr>
<td><strong>BIAS IN JUDGMENT, MEMORY AND PERCEPTION:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swim, et al. (1989)</td>
<td>Do sex stereotypes bias evaluations of men's and women's work?</td>
<td>119</td>
<td>-.04&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Mazella &amp; Feingold (1994)</td>
<td>Does defendant social category affect mock juror's verdicts?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defenders:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attractiveness</td>
<td></td>
<td>25</td>
<td>.10</td>
</tr>
<tr>
<td>Race (African-American or White)</td>
<td></td>
<td>29</td>
<td>.01</td>
</tr>
<tr>
<td>Social class</td>
<td></td>
<td>4</td>
<td>.08</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td>21</td>
<td>.04&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Kunda &amp; Thagard (1996)</td>
<td>Do stereotypes bias judgments of targets in the absence of any individuating information?</td>
<td>7</td>
<td>.25</td>
</tr>
<tr>
<td>Kunda &amp; Thagard (1996)</td>
<td>Do stereotypes bias judgments of targets in the presence of individuating information?</td>
<td>40</td>
<td>.19</td>
</tr>
</tbody>
</table>

Effect sizes are r’s.

<sup>1</sup> This excludes the results of 15 studies on animal learning included in Rosenthal & Rubin's (1978) meta-analysis. Expectations for animals are not “interpersonal” expectations.

<sup>2</sup> A negative coefficient indicates favoring men; a positive coefficient indicates favoring women.
Table 2
The Accuracy of Racial and Ethnic Stereotypes

<table>
<thead>
<tr>
<th>Study &amp; Stereotype</th>
<th>Perceivers</th>
<th>Criterion</th>
<th>Predominant Pattern of Discrepancies</th>
<th>Individual Correlations (Personal Stereotype Accuracy)</th>
<th>Aggregate Correlations (Consensual Stereotype Accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>McCauley &amp; Stitt (1978); beliefs about demographic differences between African-Americans and other Americans</td>
<td>Five haphazard samples (church choir, union members, students, etc.), total N=62</td>
<td>U.S. Census data</td>
<td>Accuracy</td>
<td>Not Available</td>
<td>Beliefs about:</td>
</tr>
<tr>
<td>Ashton &amp; Esses (1999); beliefs about the achievement of nine Canadian ethnic groups</td>
<td>94 U. of Western Ontario students</td>
<td>Board of Education achievement data</td>
<td>Accuracy: 36 of 94 Exaggeration: 33 of 94 Underestimation: 25 of 94³</td>
<td>.69</td>
<td>Not Available</td>
</tr>
<tr>
<td>Wolsko et al (2000); beliefs about differences between African-Americans and White Americans</td>
<td>83 White U. of Colorado undergraduates</td>
<td>Objective data from govt. (e.g., Census) and other (e.g., National Basketball Association) sources</td>
<td>Underestimation</td>
<td>Not Available</td>
<td>Not Available</td>
</tr>
</tbody>
</table>

Ryan’s (1996) results refer to her stereotypicality results, not her dispersion results. Exaggeration means the perceived differences between groups exceeded the group differences on the criteria. Underestimation means the perceived differences between groups was smaller than the group differences on the criteria. Individual correlations involve computing, for each individual perceiver, the correlation between their judgments (stereotypes) and the criterion. Studies performing this analysis typically report the average of those correlations. Aggregate correlations refer to the correlation between the overall average perceived difference between the groups (for the whole sample) and the group difference on the criteria.
Except where otherwise stated, all discrepancy results occur at the consensual level. Accuracy means within 10% of the real percentage or within .25 of a standard deviation. Exaggeration means the perceived differences between groups exceeded the group differences on the criteria. Underestimation means the perceived differences between groups was smaller than the group differences on the criteria. Except where otherwise noted, only one word is entered in this column when one pattern (e.g., “accuracy”) occurred for a majority of results reported. When there was no majority, the top two results (most frequent first) are reported here except where otherwise noted.

2 If the study reported more than one individual level (average) correlation, their correlations were averaged to give an overall sense of the degree of accuracy.

3 These correlations do not appear in the original article, but are computable from data that was reported.

4 For each group of perceivers, the first correlation is the correspondence between their judgments and the self-reports of their own groups; the second correlation is the correspondence between their judgments and the self-reports of the other group; and the third correlation is the correspondence between the perceived difference between the groups and the difference in the self-reports of the two groups.

5 These are personal discrepancies. Ashton & Esses (1999) computed a personal discrepancy score for each perceiver, and then reported the number of perceivers who were within .2 standard deviations of the criteria, the number that exaggerated real differences (saw a difference greater than .2 sd larger than the real difference) or underestimated real differences (saw a difference more than .2 sd smaller than the real difference).
<table>
<thead>
<tr>
<th>Study &amp; Stereotype</th>
<th>Perceivers</th>
<th>Criterion</th>
<th>Predominant Pattern of Discrepancies</th>
<th>Individual Correlations (Personal Stereotype Accuracy)</th>
<th>Aggregate Correlations (Consensual Stereotype Accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>McCauley, Thangavelu &amp; Rozin (1988); McCauley &amp; Thangavelu (1991); beliefs about the sex distribution into different occupations</td>
<td>College students, high school students, rail commuters (N=521 over the 5 studies)</td>
<td>Census data on proportion of women employed in various occupations</td>
<td>Accuracy</td>
<td>Not Available</td>
<td>.94-.98 across 5 studies</td>
</tr>
<tr>
<td>Swim (1994); beliefs about sex differences on 17 characteristics</td>
<td>Introductory psychology students (N=293 over two studies)</td>
<td>Meta-analyses of sex differences on 17 characteristics</td>
<td>Accuracy</td>
<td>Not Available</td>
<td>Study One: .78 Study Two: .79</td>
</tr>
<tr>
<td>Briton &amp; Hall (1995); beliefs about sex differences in nonverbal behavior</td>
<td>441 introductory psychology students</td>
<td>Meta-analysis of nonverbal sex differences</td>
<td>Accuracy</td>
<td>Not Available</td>
<td>Female perceivers: .74 Male perceivers: .68</td>
</tr>
<tr>
<td>Cejka &amp; Eagly (1999); beliefs about the sex distribution into different occupations</td>
<td>189 introductory psychology students</td>
<td>Census data on proportion of women employed in 80 occupations</td>
<td>Accuracy and Underestimation</td>
<td>Not Available</td>
<td>.91</td>
</tr>
<tr>
<td>Beyer (1999); beliefs about the sex distribution into different majors and mean GPA of men and women in those majors</td>
<td>265 college students</td>
<td>College data on proportion of men and women in different majors, and their GPAs</td>
<td>Accuracy and Underestimation</td>
<td>Proportion: Male perceivers: .48 Female perceivers: .52 GPA Male targets: .22 Female targets: -.04</td>
<td>Proportion: Male perceivers: .80 Female perceivers: .79 GPA Male perceivers: .35 Female perceivers: .34</td>
</tr>
<tr>
<td>Hall &amp; Carter (1999); beliefs about sex differences on 77 characteristics</td>
<td>708 introductory psychology students</td>
<td>Meta-analyses of sex differences on 77 characteristics</td>
<td>Not Available</td>
<td>.43</td>
<td>.79</td>
</tr>
</tbody>
</table>
Diekman, Eagly, & Kulesa (2002); beliefs about the attitudes of men and women 617 college students over three studies Attitude positions endorsed by men and women on the General Social Survey (random sample of American adults) Accuracy for consensual discrepancies; near miss for personal discrepancies Male targets: .45\(^5\) Female targets: .54\(^5\) When judging sex differences: .60 Male targets: .66\(^5\) Female targets: .77\(^5\) When judging sex differences: .80

Individual correlations involve computing, for each individual perceiver, the correlation between their judgments (stereotypes) and the criterion. Studies performing this analysis typically report the average of those correlations. Aggregate correlations refer to the correlation between the overall average perceived difference between the groups (for the whole sample) and the group difference on the criteria. Only one word is entered in this column when one pattern (e.g., “accuracy”) occurred for a majority of results reported. When there was no majority (or the majority could not be determined from their data), the top two results, in order of frequency (most frequent first) are reported here.

1 Except where otherwise stated, all discrepancy results occur at the consensual level. Accuracy means within 10% of the real percentage or within .25 of a standard deviation. Exaggeration means the perceived differences between groups exceeded the group differences on the criteria. Underestimation means the perceived differences between groups was smaller than the group differences on the criteria. “Near miss” means perceivers were more than 10% wrong, but no more than 20% wrong. except where otherwise noted, only one word is entered in this column when one pattern (e.g., “accuracy”) occurred for a majority of results reported. When there was no majority, the top two results (most frequent first) are reported here except where otherwise noted.

2 These correlations do not appear in the original article, but are computable from data that was reported.

3 Swim (1994) sometimes reported more than one meta-analysis as a criterion for a perceived difference. In that case, I simply averaged together the real differences indicated by the meta-analyses in order to have a single criterion against which to evaluate the accuracy of the perceived difference.

4 For Beyer (1999), all results are reported separately for men and women perceivers, except the individual correlations for GPA. Because there was no significant sex of perceiver difference in these correlations, Beyer reported the results separately for male and female targets.

5 For simplicity, if the study reported more than two correlations, I have simply averaged all their correlations together to give an overall sense of the degree of accuracy.
<table>
<thead>
<tr>
<th>Study &amp; Stereotype</th>
<th>Perceivers</th>
<th>Criterion</th>
<th>Predominant Pattern of Discrepancies[^1]</th>
<th>Individual Correlations (Personal Stereotype Accuracy)</th>
<th>Aggregate Correlations (Consensual Stereotype Accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judd, Ryan, and Park (1991); Beliefs about engineering and business majors at U. of Colorado</td>
<td>116 U. of Colorado business and engineering majors (58 each) randomly selected</td>
<td>Self-reports of those randomly selected 116 business and engineering majors</td>
<td>Accuracy and exaggeration</td>
<td>.63[^2]</td>
<td>Not Available</td>
</tr>
<tr>
<td>Judd &amp; Park (1993); Democrats’ and Republicans’ beliefs about one another’s political attitudes</td>
<td>An unspecified number of people randomly surveyed as part of the 1976 National Election Study</td>
<td>Self-reported attitudes of self-identified Democrats and Republicans</td>
<td>Accuracy[^3] and “liberalism bias”: overestimating the liberalism of both parties</td>
<td>.25[^2]</td>
<td>Not Available</td>
</tr>
<tr>
<td>Cejka &amp; Eagly (1999); Wages in different occupations</td>
<td>189 introductory psychology students</td>
<td>Census data on 80 occupations</td>
<td>“Contraction bias” – overestimating wages in low wage jobs and underestimating wages in high wage jobs</td>
<td>Not Available</td>
<td>.94</td>
</tr>
<tr>
<td>Ryan &amp; Bogart (2001); beliefs that sorority members hold about their own and other sororities</td>
<td>136-181 sorority members (attrition over time due to graduation and dropping out of school or sorority)</td>
<td>Self-reports of 85%-100% of the full members (not including pledges and new initiates) of each sorority</td>
<td>Accuracy perceiving their own sorority; exaggeration of stereotypicality when perceiving other sororities</td>
<td>When perceiving their own sorority: .52[^2] When perceiving other sororities: .39[^2]</td>
<td>Not Available</td>
</tr>
</tbody>
</table>
Clabaugh & Morling (2004): beliefs about the psychological characteristics of ballet dancers and modern dancers

| 175 ballet dancers, modern dancers, and psychology students | Self-reports of the ballet dancers and modern dancers | Accuracy | Perceiving differences between groups: 
| Ballet perceivers: .67 | Modern dance perceivers: .71 | Psych student perceivers: .62 |
| Perceiving differences across traits within target groups: 
| Ballet perceivers: .59<sup>2</sup> | Modern dance perceivers: .67<sup>2</sup> | Psych student perceivers: .45<sup>2</sup> |

<sup>1</sup>Except where otherwise stated, all discrepancy results occur at the consensual level. Except where otherwise noted: 1. accuracy means within 10% of the real percentage or within .25 of a standard deviation; 2. exaggeration means the perceived differences between groups exceeded the group differences on the criteria; and 3. underestimation means the perceived differences between groups was smaller than the group differences on the criteria. Only one word is entered in this column when one pattern (e.g., “accuracy”) occurred for a majority of results reported. When there was no majority (or the majority could not be determined from their data), the top two results, in order of frequency (most frequent first) are reported here.

<sup>2</sup> If the study reported more than one individual level (average) correlation, their correlations were averaged to give an overall sense of the degree of accuracy.

<sup>3</sup> Neither percentages nor standard deviations were reported. I characterize the main results of their discrepancy analyses as “accurate” because seven of eight mean discrepancies are all less than 1 scale point (on a seven point scale).

<sup>4</sup> These correlations do not appear in the original article, but are computable from data that was reported.
Table 5
Social Stereotypes are More Valid than Most Social Psychological Hypotheses

<table>
<thead>
<tr>
<th></th>
<th>Proportion of Social Psychological Effects</th>
<th>Proportion of Consensual Stereotype Accuracy Correlations</th>
<th>Proportion of Personal Stereotype Accuracy Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exceeding .30</td>
<td>24%</td>
<td>100% (31/31)</td>
<td>86% (18/21)</td>
</tr>
<tr>
<td>Exceeding .50</td>
<td>5%</td>
<td>94% (29/31)</td>
<td>52% (11/21)</td>
</tr>
</tbody>
</table>

1 Date obtained from Richard et al’s (2003) review of meta-analyses of all of social psychology, including thousands of studies. Effects are in terms of the correlation coefficient, r.

2 From Tables 2 through 4. Within parentheses, the numerator is the number of stereotype accuracy correlations meeting the criteria for that row (exceeding .30 or .50) and the denominator is the total number of stereotype accuracy correlations. Because Table 2 summarizes the results for five studies for McCauley et al (1988), the .94-.98 figure is counted five times. These numbers probably underestimate the degree of stereotype accuracy, because all single entries in Tables 2 through 4 only count once, even though they often constitute averages of several correlations found in the original articles, and because I did not use the r-to-z transformation (this table describes the data, it does not report tests of statistical significance). Furthermore, the simple average of correlations is conservative, tending to underestimate the true correlation (Silver & Dunlap, 1987).