Biases in research evaluation: Inflated assessment, oversight, or error-type weighting? ☆

Darcy A. Reich a,*, Melanie C. Green b, Timothy C. Brock c, Philip E. Tetlock d

a Department of Psychology, MS 2051, Texas Tech University, Lubbock, TX 79409, USA
b Department of Psychology, University of North Carolina at Chapel Hill, USA
c Department of Psychology, The Ohio State University, USA
d Haas School of Business, University of California, Berkeley, USA

Received 11 July 2005; revised 17 February 2006

Abstract

Reviewers of research are more lenient when evaluating studies on important topics [Wilson, T. D., Depaulo, B. M., Mook, D. G., & Klaaren, K. J. (1993). Scientists’ evaluations of research: the biasing effects of the importance of the topic. Psychological Science, 4(5), 323–325]. Three experiments (N = 145, 36, and 91 psychologists) investigated different explanations of leniency, including inflation of assessments (applying a heuristic associating importance with quality), oversight (failing to detect flaws), and error-weighting (prioritizing Type II error avoidance). In Experiment 1, psychologists evaluated the publishability and rigor of studies in a 2 (topic importance) × 2 (accuracy motivation) × 2 (research domain) design. Experiment 2 featured an exact replication of Wilson et al. and suggested that report length moderated the effects of importance on perceived rigor, but not on publishability. In Experiment 3, a manipulation of error-weighting replaced the manipulation of domain (Experiment 1). Results favored error-weighting, rather than inflation or oversight. Perceived seriousness of Type II error (in Experiments 1 and 3) and the error-weighting manipulation (in Experiment 3) predicted study evaluations.

© 2006 Published by Elsevier Inc.

Introduction

Scientists ought to transcend the psychological and political factors that bias ordinary mortals’ judgments of research. Yet, scientists often fall short of complete adherence to this ideal MacCoun (1998). Paradoxically, previous research suggests that scientists relax their standards for rigor precisely when critical assessment is most essential. Wilson and his colleagues (Wilson et al., 1993) posited that topic importance may influence evaluators’ judgments in two ways. First, evaluators may be “more forgiving of the methodological weaknesses that they notice in studies of important problems” (p. 322)—which they characterized as a leniency bias. Second, the importance of a research topic may make it more difficult to detect methodological flaws, leading evaluators to overlook these flaws—which they labeled as an oversight bias.

Wilson et al. (1993) elicited evaluations from medical faculty and research psychologists of the methodological...
rigor and publishability of six flawed studies, using a
within-participants design. For each study, one version
involved an important topic (e.g., cardiovascular disease)
and one involved an unimportant topic (e.g., heartburn).
Each participant evaluated studies on three important and
three unimportant topics. Studies of important topics elic-
ited higher ratings of rigor and publishability than did ident-
tical studies of unimportant topics.

60 Processes underlying lenient evaluations

Although studies have shown that belief-consistency and
topic importance can affect research evaluations, much less
research has explored how these “biases” occur (Tetlock &
Levi, 1982). Clarification of the processes underlying these
evaluative tendencies would allow us to predict their occur-
rence and, if appropriate, to seek corrections for their
effects (e.g., through debiasing interventions).

We propose three processes that may contribute to more
lenient evaluations of rigor and publishability for impor-
tant research. First, leniency may stem from a heuristic pro-
cess of inflated assessment, whereby evaluators assume that
if the topic is important, the study also must be important
and of high quality. Thus, they minimize the seriousness of
flaws they notice, and systematically inflate their assess-
ments (at least when results bolster preferred conclusions).

Second, as Wilson et al. (1993) suggested, oversight may
be involved. Evaluators of important research may engage
in less careful scrutiny, without necessarily being aware of
doing so. Due to their enthusiasm about the potential sig-
nificance of the research, they may fail to notice methodo-
logical flaws that they otherwise would notice.

Third, we propose that error-type weighting may contrib-
ute to the tendency to make less critical judgments. Given the
uncertainty inherent in scientific findings, the probability of
“accepting” a false finding (Type I error) must be balanced
against the probability of “rejecting” a true finding (Type II
error). In basic research, Type I errors are seen as most prob-
lematic. Scientists are taught to be vigilant against false find-
ings; indeed, procedures for significance testing clearly
privilege Type I error protection over Type II error protec-
tion. However, many stakeholders (e.g., applied researchers,
policy makers), have a strong desire to seek solutions to
pressing practical problems. For some problems (e.g., AIDS),
the prospect of a potential solution (i.e., a life-extending
drug), no matter how uncertain, may outweigh the benefits of
further testing. In such research, the consequences of a Type
II error—forsaking a true cure—are seen as more disastrous
than those of a Type I error—a false cure.

We propose that evaluators use topic importance as a cue
that signifies the appropriate balance between avoiding Type I
vs. Type II errors. Although consensus exists regarding the
seriousness of Type I errors, evaluators’ emphasis on Type II

1 We use the terms Type I and Type II errors in a conceptual rather than
a strict statistical sense.

Current research

The current experiments attempted to examine the rela-
tive impact of these three processes. We used Wilson et al.’s
(1993) paradigm, in which sophisticated evaluators rated
reports of flawed studies involving either important or
unimportant topics. However, we used longer reports and a
between-participants design. We measured sensitivity to
Type II and Type I errors and manipulated accuracy moti-
vation to encourage participants to overcome bias through
careful scrutiny (see MacCoun, 1998). We reasoned that if
oversight contributes to less critical evaluations, an injunc-
tion to carefully scrutinize the report may prompt evalua-
tors to overcome the tendency to miss flaws in important
research (unless the oversight process is entirely automatic).
If so, an interaction between topic importance and accu-
\[\text{racy motivation on evaluations of rigor and publishability.\]
first received instructions encouraging them to be as accurate as possible when evaluating the research (accuracy motivation condition), while others did not. Thus, participants were randomly assigned to conditions in a 2 (topic importance: important vs. unimportant) × 2 (accuracy motivation: present vs. absent) × 2 (domain: abuse prevention vs. therapy effectiveness) design. Each report deliberately included a flaw: (1) nonrandom assignment to conditions in the abuse prevention studies (i.e., “classes whose teachers did not volunteer...were used as the control group.”) and (2) primary measures requiring retrospective self-report for the therapy effectiveness studies.2

Following Wilson et al. (1993), our dependent measures assessed how likely participants would be to recommend publication (1 = very unlikely, 7 = very likely), how methodologically sound the design was (1 = very flawed, 7 = very sound), and how well the researchers’ conclusions were supported by the data (1 = not at all well-supported, 7 = very well-supported). A 7-point manipulation check assessed participants’ perceptions of topic importance, and two measures assessed sensitivity to Type II and Type I errors (“Given the nature of the problem studied, how serious a mistake (to society) would it be to reject [accept] this report and subsequently find that its conclusions are true [false]?”; 1 = not at all serious, 7 = very serious).

Results

Importance manipulation check

A 2 (importance) × 2 (accuracy) × 2 (domain) ANOVA revealed only that the important versions of the studies were rated as more important (M = 5.59) than the unimportant versions (M = 3.99), F(1,135) = 50.14, p < .001. However, several participants rated the topic of the ‘unimportant’ studies as extremely important (6 or 7). This result may have occurred due to distractions inherent in the conference setting, participants’ tendency to justify their participation by assuming that the study they evaluated must be reasonably important, and genuine interpretations of the unimportant topics as important. However, the integrity of our manipulation depended on participants acknowledging that the unimportant topics were relatively unimportant. Therefore, data for participants who did not perceive topic importance as intended—those who rated importance above the midpoint for unimportant topics (n = 26) or below the midpoint for important topics (n = 6)—were omitted for the primary ANOVAs conducted to examine the effects of the importance manipulation.

Publishability and perceived rigor

As predicted, a 3-way ANOVA revealed that the important studies (M = 3.33) were rated more publishable than the unimportant ones (M = 2.34), F(1,103) = 8.05, p < .01. An effect of domain also emerged, F(1,103) = 4.54, p < .05. The abuse prevention studies (M = 3.36) were rated more publishable than the therapy effectiveness studies (M = 2.52).

Responses on the measures of methodological soundness and support for conclusions (r = .78) were averaged to create an index of perceived rigor, as done by Wilson et al. (1993). A 3-way ANOVA on perceived rigor revealed no effect of importance, F(1,103) = 0.10. However, the abuse prevention studies (M = 3.29) were seen as more rigorous than the therapy effectiveness studies (M = 2.51), F(1,103) = 8.80, p < .01.

No interactions and no effects involving the accuracy manipulation emerged for publishability or rigor.3

Error sensitivity

Did topic importance influence the perceived seriousness of Type II error (rejecting the report when the conclusions were true) and Type I error (accepting the report when the conclusions were false)? Two 3-way ANOVAs revealed that importance affected the perceived seriousness of a Type II error, F(1,102) = 10.74, p < .01, but not of a Type I error, F(1,102) = 0.70,4.5

Participants who read about important topics perceived Type II error as more serious (M = 4.22) than did those with unimportant topics (M = 3.22).

Next, path analyses were conducted on the full sample. All included paths are shown in Fig. 1. As expected, manipulated importance affected perceived importance, which, in turn, enhanced participants’ sensitivity to Type II error. Both sensitivity to Type II error and perceived importance independently predicted publishability ratings, but sensitivity to Type I error did not. Thus, the more participants believed that a Type II error would be serious, the more likely they were to recommend publication. Sensitivity to both Type II and Type I errors predicted perceived rigor, but perceived importance did not. The more participants believed that a Type II error would be serious and the less they believed that a Type I error would be serious, the more likely they were to see the study as rigorous.

Discussion

These findings offered little direct support for oversight: the accuracy manipulation had no effect, topic importance did not enhance perceived rigor, and low evaluations of rigor (M = 2.93) indicated that flaws did not go unrecognized. The findings also did not favor inflated assessment, since importance did not directly affect perceived rigor. Yet, some inflation of publishability ratings may have occurred,

1 Results were identical using the full sample, except that the importance effect on publishability was weaker and nonsignificant.

2 For Type I error sensitivity, there also was a domain effect, F(1,102) = 9.38, qualified by a domain by importance interaction, F(1,102) = 12.01. For important topics, no difference across domains emerged (both Ms = 4.67). For unimportant topics, participants saw Type I error as more serious for the psychic advice (M = 5.33) than the caffeine abuse (M = 3.35) study, F = 4.19, p < .05.

3 Results were identical using the full sample.
given that perceived importance independently predicted publishability in the path model.

In contrast, strong support emerged for the influence of error-type weighting on leniency. When evaluating studies on more important topics, evaluators perceived the costs of missing a true finding (making a Type II error) as more serious. Furthermore, Type II error sensitivity predicted publishability ratings, controlling for both manipulated and perceived importance. The more sensitive participants were to Type II error, the more highly they rated the rigor of the studies (perhaps to justify publication recommendations). Thus, topic importance had an indirect, rather than a direct, influence on perceived rigor by enhancing sensitivity to Type II error.

Consistent with Wilson et al. (1993), manipulated topic importance influenced ratings of publishability. However, contrary to their findings, topic importance did not directly affect perceived rigor. This notable discrepancy indicated that there may be conditions under which evaluators’ perceptions of a study’s rigor are not distorted by topic importance. Clarification of the role of error-type weighting appeared to require a reconciliation of differences between our methodology and that used by Wilson et al., such as the report length and the between- vs. within-participants design.

**Experiment 2**

Experiment 2 included Wilson et al.’s original materials, in addition to one of our longer reports, in a within-participants design. Given that Wilson et al.’s shorter reports contained less information, participants may have relied more heavily on a simple association of importance with rigor when evaluating these studies. If so, importance should predict ratings of rigor more strongly as report length decreases.

**Method**

**Participants**

Participants were 36 volunteers who attended poster sessions at an Association for Behavior Therapy convention. About 14% had PhDs, 39% had master’s degrees, and 44% had bachelor’s degrees. Most had taken statistics (89%) and graduate-level research methods (83%) courses.

**Procedure**

For Experiment 2a, participants evaluated the six flawed research reports used in Wilson et al.’s (1993) study using the same measures of publishability, rigor, and importance used in Experiment 1. Each participant was randomly assigned to read the important version of three reports and the unimportant version of the other three reports; order was counterbalanced between participants. For Experiment 2b, participants read either the drug or caffeine abuse report from Experiment 1, and completed the same measures. The order in which participants received materials for Experiments 2a and 2b was counterbalanced. Thus, each participant evaluated seven reports, and received either the important or unimportant version of each report.

**Results**

**Experiment 2a**

Each participant’s responses were averaged for the three important versions and the three unimportant versions of the six original reports.

**Importance manipulation check.** As expected, important topics were rated as more important ($M = 5.70$) than unimportant ones ($M = 3.99$), $F(1,35) = 117.04, p < .001$. 

---

![Path model predicting publishability ratings and perceived rigor from manipulated importance, perceived importance, and Type I and II error sensitivity for Experiment 1. All paths significant at the $p < .05$ level are indicated in bold, underlined font. Zero-order effects of exogenous variables are presented in parentheses.](image-url)
**Publishability and perceived rigor.** Studies on important topics ($M = 3.77$) were rated as more publishable than studies on unimportant topics ($M = 2.80$), $F(1,35) = 23.16$, $p < .001$. Ratings of methodological soundness and support for conclusions were highly correlated for both important ($r = .85$) and unimportant ($r = .88$) topics, and were averaged to create perceived rigor indexes. Although studies on important topics ($M = 3.25$) overall tended to be rated as more rigorous than studies on unimportant topics ($M = 2.82$), $F(1,35) = 4.49$, $p < .05$, the effect was limited (see Further analyses).

**Experiment 2b**

**Importance manipulation check.** The important topic study was rated more important ($M = 6.06$) than the unimportant one ($M = 4.47$), $t(1,34) = -4.40$, $p < .001$. As in Experiment 1, data for participants who rated importance above the midpoint for unimportant topics ($n = 10$) or below the midpoint for important topics ($n = 0$) were omitted from further analysis.

**Publishability and perceived rigor.** Ratings of methodological soundness and support for conclusions ($r = .87$) were averaged. The important topic study ($M = 3.94$) was rated more publishable than the unimportant topic study ($M = 2.78$), $t(1,24) = 2.12$, $p < .05$. No effect if importance emerged on perceived rigor, $t(1,24) = -.57$.

**Experiments 2a and 2b: Further analyses**

The effects of importance on perceived rigor in Experiment 2a were notably weaker than the effects on publishability, as was also evident in Wilson et al. (1993). Indeed, when independent $t$-tests were conducted on all seven reports from Experiments 2a and 2b, six reports showed effects of importance on publishability (at $p < .08$), but only one—the shortest of the seven—showed an effect approaching significance for perceived rigor ($p < .06$). These results suggest that the effects of importance on perceived rigor can be attenuated when the amount of information in a report is increased. To explore this possibility further, mean differences and standardized effect sizes (Cohen’s $d$) were computed for the between-subjects effect of manipulated importance on perceived rigor and publishability for each report (see Table 1); effect sizes can be interpreted as small at .2, medium at .5, and large at .8 (Cohen, 1988). Simple correlations were computed between these effect sizes and report length (number of words). As report length increased, the effect of importance on perceived rigor became weaker ($r = - .62$), even as the effect on publishability became stronger ($r = .40$). Using Wilson et al.’s original reports only, correlations were $- .59$ for rigor and $.18$ for publishability.

**Discussion**

The effects of topic importance on publishability ratings were replicated for both Wilson et al.’s (1993) brief research reports using a within-subjects design (Experiment 2a) and our longer research reports using a between-subjects design (Experiment 2b). However, as predicted, the effects of importance on perceived rigor were different for Experiments 2a and 2b. For the brief reports as a group, studies on important topics were rated higher in rigor than identical studies on unimportant topics, replicating Wilson et al. However, consistent with Experiment 1, no effect of topic importance on perceived rigor was found for the longer research reports. Further analyses provided evidence that the effects of importance on perceived rigor generally became weaker as report length increased. With more information, evaluators may be less prone to rely solely on heuristics associating importance with rigor. Furthermore, the effect of importance on perceived rigor in within-subjects designs may be an artifact of averaging across ratings. This procedure may enhance power enough to detect an indirect influence of importance on perceived rigor; that is, as shown in Experiment 1, importance may affect the weighting of Type II error, which, in turn, affects rigor, and this impact may be (mis)identified as a direct effect of importance on perceived rigor.

**Table 1**

Effect sizes for the effect of topic importance on ratings of publishability and perceived rigor as a function of report length (Experiment 2)

<table>
<thead>
<tr>
<th>Report length (in # words)</th>
<th>Publishability</th>
<th>Effect size</th>
<th>Perceived Rigor</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean diff (Pooled SD)</td>
<td></td>
<td>Mean diff (Pooled SD)</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>.67 (1.41)</td>
<td>.47</td>
<td>.81 (1.23)</td>
<td>.65</td>
</tr>
<tr>
<td>52</td>
<td>1.11 (1.48)</td>
<td>.75</td>
<td>1.13 (1.13)</td>
<td>.52</td>
</tr>
<tr>
<td>59</td>
<td>1.17 (1.58)</td>
<td>.74</td>
<td>.89 (1.38)</td>
<td>.28</td>
</tr>
<tr>
<td>100</td>
<td>1.22 (1.40)</td>
<td>.87</td>
<td>.11 (1.34)</td>
<td>.08</td>
</tr>
<tr>
<td>106</td>
<td>.98 (1.57)</td>
<td>.62</td>
<td>.31 (1.41)</td>
<td>.22</td>
</tr>
<tr>
<td>121</td>
<td>.86 (1.40)</td>
<td>.62</td>
<td>.53 (1.28)</td>
<td>.42</td>
</tr>
<tr>
<td>129</td>
<td>1.16 (1.33)</td>
<td>.87</td>
<td>.33 (1.40)</td>
<td>.23</td>
</tr>
</tbody>
</table>

Average effect size: .71 .34

Note: the six original reports used by Wilson et al. (1993) are presented in normal type–face; the report we added is presented in italics. Mean Diff refers to the difference between the average ratings of participants who received the important vs. unimportant versions of each report.

* $p < .05$.

+ $p < .1$. 
Experiment 3

Our arguments for error-type weighting rely, thus far, on internal analyses in Experiment 1 showing that participants’ differential sensitivity to Type II error predicted the degree of leniency in their evaluations of flawed studies. To provide stronger evidence that an emphasis on the seriousness of Type II error leads to less critical evaluations, we manipulated participants’ focus on Type II and Type I errors in Experiment 3.

Method

Participants
Participants were 91 volunteers (35 males, 55 females, 1 unreported) who attended poster sessions at another APA convention. About 32% had PhDs, 37% had master’s degrees, and 26% had bachelor’s degrees. Most had taken statistics (87%) and graduate-level research methods (81%) courses.

Procedure
The procedure was identical to Experiment 1, with the following exceptions. Research domain was not varied; only the abuse prevention reports were used. The instructions for the accuracy induction were strengthened by adding two sentences emphasizing participants’ expertise and capability to be accurate. To more fully assess participants’ sensitivity to various costs of making Type I and Type II errors, we added two measures identical to those used in Study 1, except that “mistake to society” was replaced with “mistake to the scientific reputation of psychology.”

We also sought to manipulate participants’ focus on Type II and Type I errors. Before reading the report, participants were told that the study was currently under review at a journal of program evaluation read by academics and school administrators. Participants in the Type II error-focus condition were told,

“...if these results are accurate, overlooking them could mean that schools might miss out on the chance to implement an effective program. ...[and] research progress in this area might be delayed and valuable time, effort, and resources might be lost by following less fruitful lines of inquiry.”

Type I error-focus participants were told,

“...if these results are inaccurate, accepting them could mean that schools and taxpayers might waste thousands of dollars on ineffective programs. ...[and] countless other researchers might waste valuable time, effort, and resources trying to replicate and extend these findings.”

Results

Importance manipulation check
A 2 (importance) × 2 (accuracy) × 2 (error-focus) ANOVA revealed that the important topic was rated as more important (M = 6.12) than the unimportant topic (M = 3.85), F(1, 82) = 83.28, p < .001. In addition, Type II error-focus participants perceived the research topic to be more important (M = 5.43) than did Type I error-focus participants (M = 4.73), F(1, 82) = 11.02, p < .01. Accuracy-motivated participants also perceived the topic to be more important (M = 5.37) than did those who were not accuracy-motivated (M = 4.80), F(1, 82) = 4.72, p < .05. The injunction to scrutinize the report may have served as a cue signaling high importance. Because all three independent variables influenced perceived importance, these ratings could not be used unequivocally as an indicator of the effectiveness of the importance manipulation. Therefore, the procedure for omitting participants based on perceived importance ratings was not used. Although retaining participants who saw the unimportant topic as highly important reduced the likelihood of detecting a main effect of importance on publishability, it allowed a more thorough examination of the effects of manipulated error-focus and accuracy motivation.

Publishability and perceived rigor
A 2 (importance) × 2 (accuracy) × 2 (error-focus) ANOVA revealed only that participants induced to focus on the seriousness of Type II error rated the research as more publishable (M=4.13) than did those focusing on Type I error (M=3.30), F(1, 83) = 7.78, p < .01. Again, ratings of methodological soundness and support for conclusions were averaged to measure perceived rigor (r=.72). A 3-way ANOVA showed only that Type II error-focus participants rated the research as more rigorous (M = 3.98) than did Type I error-focus participants (M = 3.41), F(1, 82) = 4.67, p < .05. Neither manipulated importance nor accuracy significantly affected publishability or perceived rigor, nor were there other significant effects.

When participants who did not perceive importance as intended were omitted, manipulated importance did not significantly affect publishability (F(1, 68) = 2.02, p = .16). However, the mean difference (important M = 3.81; unimportant M = 3.29) was in the predicted direction.

6 Although we predicted no effect of manipulated importance on rigor, we sought indirect evidence of oversight by examining whether the strengthened accuracy manipulation would attenuate the importance effect on publishability.

7 When participants who did not perceive importance as intended were omitted, manipulated importance did not significantly affect publishability (F(1, 68) = 2.02, p = .16). However, the mean difference (important M = 3.81; unimportant M = 3.29) was in the predicted direction.
Error sensitivity

Next, we conducted two 3-way ANOVAs to examine whether the manipulated variables influenced reported error sensitivity. To create more reliable measures, the two measures of Type II error sensitivity ($r = .56$) were averaged, as were the two Type I error measures ($r = .54$). Only topic importance influenced the perceived seriousness of Type II errors, $F(1,89) = 6.13, p < .05$, and of Type I errors, $F(1,83) = 9.25, p < .01$. Participants who read about studies on important topics perceived both Type II ($M = 4.36$) and Type I ($M = 4.78$) errors to be more serious than did those reading about unimportant topics ($Ms = 3.68$ and 3.92, respectively). No other effects were significant. To ensure that the error-focus manipulation did influence Type II error weighting for unimportant topics, as expected, we conducted focused comparisons. For the unimportant topic, participants induced to focus on Type II error saw Type II error as a more serious mistake ($M = 4.07$) than did those induced to focus on Type I error ($M = 3.21$), $t(40) = -2.47, p < .05$. For the important topic, participants reported a relatively high sensitivity to Type II error (overall $M = 4.36$) that did not differ based on manipulated error-focus, $t(47) = 0.20$. As expected, the error-focus manipulation did not affect Type I error sensitivity for either important or unimportant topics, suggesting that it is less malleable.

Given the relatively small sample size, and given that all three manipulated variables influenced perceived importance, a full path model like that conducted for Experiment 1 was deemed unsupportable. Instead, we tested more focused predictions within each error-focus condition. Specifically, in the Type I error-focus condition, perceived importance should influence publishability ratings, and should do so indirectly, by influencing sensitivity to Type II error. However, when participants are directly induced to focus on Type II error, the effects of perceived importance on Type II error sensitivity and on publishability should be diminished. All included paths are shown in Fig. 2. In the Type I error-focus condition, the zero-order effect of perceived importance on publishability was significant; furthermore, perceived importance strongly predicted participants’ sensitivity to Type II error, which, in turn, increased ratings of publishability, as well as perceived rigor. Controlling for Type II error sensitivity, perceived importance did not independently predict publishability. In the Type II error-focus condition, the effect of perceived importance on publishability was not significant. Perceived importance was a more modest predictor of sensitivity to Type II error, which strongly predicted ratings of publishability, as well as perceived rigor. Type I error sensitivity did not predict publishability or perceived rigor in either condition.

Discussion

These findings again supported the role of error-type weighting. Participants induced to focus on the seriousness of Type II error rated the research reports as more publishable and more rigorous than did those induced to focus on Type I error. In the Type I error-focus condition, perceived importance influenced publishability ratings, and did so indirectly, by influencing sensitivity to Type II error. However, when participants were directly induced to focus on Type II error, perceived importance no longer influenced publishability ratings, its effects on Type II error sensitivity were diminished, and Type II error sensitivity strongly predicted publishability. Once again, the null results for the strengthened accuracy manipulation failed to support oversight, and the fact that importance did not predict perceived rigor did not support either inflated assessment or oversight. Furthermore, in neither path model did perceived importance independently predict publishability when controlling for Type II error sensitivity, as it did in Experiment 1. Thus, the more reliable measurement of error sensitivity in Experiment 3 appeared to rule out simple inflation of assessments.

General discussion

To our knowledge, variation among scientists in sensitivity to Type I and Type II error has not been investigated previously. Our findings show, however, that this sensitivity is measurable within the context of specific studies, manipulable as an experimental construct, and predictive of consequential judgments.
The current studies suggested that leniency in the evaluation of research on important topics may stem from a heightened emphasis on avoiding Type II errors. Sensitivity to Type I error was a markedly less consistent predictor of leniency. Because research training emphasizes the minimization of Type I error, personal sensitivity to Type I error may be relatively well-learned and unlikely to fluctuate from topic to topic (e.g., based on importance). When judging research on pressing social problems, scientists may experience heightened concerns about failing to recognize a solution. The Type I/Type II trade-off can thus become critical when scientists evaluate research of a more applied nature.

In contrast to the results of Experiment 2a and Wilson et al. (1993), in which several brief reports were used, topic importance did not affect ratings of perceived rigor across Experiments 1, 2b, and 3 using longer reports. Together, Experiments 2a and 2b indicated that importance may influence perceived rigor to a greater degree when evaluators have minimal information. This finding has implications for the situations in which leniency is most likely to occur. Evaluators of research abstracts (conferences, electronic databases), as well as consumers of scientific news in the mass media, receive what often amounts to sound bites; the lack of detailed information may lead such recipients to rely heavily on heuristics associating importance with rigor. In contrast, journal reviewers dealing with lengthy manuscripts should be less vulnerable to such biases.

Our findings suggest that leniency need not be a product of oversight or inflated assessment. The low evaluations of rigor and the lack of effects for the accuracy manipulations in Experiments 1 and 3 suggest that lack of scrutiny is not the primary process underlying leniency, at least within the context of relatively brief research reports. Furthermore, across three experiments, importance did not directly increase perceptions of rigor for the longer reports, arguing against both oversight and inflated assessment. However, oversight cannot be ruled out completely because the accuracy manipulations may have been ineffective or environmental distractions may have limited participants’ ability to carefully scrutinize the materials.

The current findings indicate that evaluators may sometimes err on the side of avoiding Type II error because they fear the potential mistake of missing important effects. That is, Type II error may be seen as more serious for important topics, and heightened sensitivity to Type II error may, in turn, engender more lenient evaluations. For studies on controversial topics (e.g., abortion), evaluators also may adjust their emphasis on Type II error, but these adjustments may be based on factors other than importance, such as personally-preferred conclusions.

The proposed processes underlying leniency are not necessarily independent, however. Consistent with a motivated reasoning account (Kunda, 1990), concerns about avoiding Type II errors could elicit selective attention (causing oversight) and biased assimilation (causing inflated assessments). Future research should explore further how error-type weighting works (e.g., by shifting evaluators’ cut-off thresholds or shaping their judgments about alternative explanations).

References


