Abstract
The National Institutes of Health (NIH) is the largest source of funding for biomedical research in the world. Funding decisions are made largely based on the outcome of a peer review process that is intended to provide a fair, equitable, timely, and unbiased review of the quality, scientific merit, and potential impact of the research. There have been concerns about the criteria reviewers are using, and recent changes in review procedures at the NIH now make it possible to conduct an analysis of how reviewers evaluate applications for funding. This study examined the criteria and overall impact scores recorded by assigned reviewers for R01 grant applications. The results suggest that all the scored review criteria, including innovation, are related to the overall impact score. Further, good scores are necessary on all five scored review criteria, not just the score for research methodology, in order to achieve a good overall impact score.

Keywords
peer review, scored review criteria, National Institutes of Health, R01

Innovations and new knowledge produced from investment in research have tremendous benefits to society. In fact, the economic growth and increase in the quality of life that have occurred in industrialized nations over the last 100 years have been attributed primarily to investments in research and development. For example, government investment of taxpayer money in biomedical research led to discoveries that dramatically reduced disease and suffering, increased the length and quality of life, and produced new products, jobs, businesses, and even whole new industries (for review, see National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2007).

The best mechanisms that governments have developed for evaluating and selecting applications for funding research seem to be those developed and used in the United States that rely on peer review by expert panels using consistent, systematic procedures, evaluating a range of qualitative and quantitative variables, in a process that is transparent to all stakeholders (Coryn, Hattie, Scriven, 1 Center for Scientific Review, National Institutes of Health, Bethesda, MD, USA
2 Office of the Vice-Chancellor for Research and Graduate School of Library and Information Sciences, University of Illinois Urbana-Champaign, Champaign 61820

Corresponding Author:
Mark D. Lindner, Center for Scientific Review, National Institutes of Health, 6701 Rockledge Dr., Bethesda, MD 20892, USA.
Email: mark.lindner@nih.gov
& Hartmann, 2007). However, the predictive validity of peer review has not yet been empirically demonstrated (Demicheli & Di Pietrantonj, 2007) and it is not clear how reviewers actually evaluate applications or if the best research projects are being funded.

The National Institutes of Health (NIH) is the single largest source of funding of biomedical research in the United States. It reviews approximately 70,000 grant applications each year using approximately 16,000 reviewers from the scientific community in a peer review process intended to provide fair, equitable, timely, unbiased reviews, based on a process that is transparent to all applicants and reviewers. The size, well-defined procedures, and transparency makes the peer review process at the NIH a unique and rich case study for evaluation methodology. With an annual budget of US$31 billion, 92% of the NIH budget is allocated to funding research and 53% is allocated for Research Project Grants. Of the several types of Research Project Grants, the R01 is the original mechanism for funding research at the NIH. It typically supports a discrete, specified, and circumscribed project of up to 5 years duration, to be performed by applicants in an area representing their specific interests and competencies, based on the mission of the NIH (see http://grants.nih.gov/grants/funding/r01.htm).

**Background on NIH Review Process**

The procedures the NIH uses to review R01 applications for funding are constantly evolving to optimize the process. For example, it has been recognized since the advent of modern science and the scientific method that scientists and the scientific community have difficulty recognizing and are often resistant to accepting innovative and transformative hypotheses and findings. This may be a natural consequence of the education and training that scientists receive (Bacon, 1620; Bernard, 1865; Kuhn, 1996, 1975). It is so difficult for people to accept new findings that are inconsistent with what they believe that Max Planck famously wrote, “A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it” (Planck, 1949, p. 33).

The NIH has been aware of this issue for decades and has instituted policies and procedures to facilitate the recognition and support of innovative and transformative projects. Since at least as early as the 1970s, NIH reviewers of R01 applications were asked to consider, among a wide range of other issues, the novelty and originality of the approach and whether projects would produce new data and concepts or confirm existing hypotheses. But their reviews resulted in only a single score that reflected the overall quality and scientific merit of each application and it was not clear from this score how much emphasis each of the component considerations were given in the review process (Henley, 1977).

In 1997, innovation and four other core review criteria were specified for reviewers to evaluate and rely on when determining their score for the overall impact of the applications (see Table 1, from NIH policy announcement number NOT-95-10 at http://grants.nih.gov/grants/guide/notice-files/not97-010.html). However, the format of the application still did not require or allow the applicants to state explicitly how their applications were innovative. Concerns continued to persist that reviewers fail to recognize innovation and/or that there may actually be an inverse correlation such that the most innovative projects tend to receive poor overall impact scores (e.g., Azoulay, Zivin, & Manso, 2009; Kaplan, 2011; Kolata, 2009).

Therefore, in 2010, the NIH made additional changes to the peer review process to increase the emphasis on innovation and decrease the focus on methodological detail. The length of the research strategy section for R01 applications was reduced from 25 pages to only 12 pages and instructions were added for reviewers to focus on overall, general issues, not routine methodological detail. The application format was changed to allow applicants to specify how their applications are innovative as well as how they meet the other four review criteria and reviewers were also required to provide integer scores on a scale of 1–9 (Table 2) for each of the five core review criteria as well as the
Table 1. Scores Assigned for Overall Impact and Five Review Criteria.

| Description                                                                 | Overall Impact Score |
|----------------------------------------------------------------------------|----------------------|
| Reviewers will provide an overall impact/priority score to reflect their    |                      |
| assessment of the likelihood for the project to exert a sustained, powerful  |                      |
| influence on the research field(s) involved, in consideration of the        |                      |
| following five core review criteria, and additional review criteria (as     |                      |
| applicable for the project proposed).                                    |                      |

Scored Review Criteria

| Scored Review Criteria | Description                                                                 |
|------------------------|-----------------------------------------------------------------------------|
| Significance           | Does the project address an important problem or a critical barrier to      |
|                        | progress in the field? If the aims of the project are achieved, how will   |
|                        | scientific knowledge, technical capability, and/or clinical practice be    |
|                        | improved? How will successful completion of the aims change the concepts,  |
|                        | methods, technologies, treatments, services, or preventative interventions   |
|                        | that drive this field?                                                     |
| Investigators          | Are the principal investigators, collaborators, and other researchers well  |
|                        | suited to the project? If early stage investigators or new investigators are |
|                        | in the early stages of independent careers, do they have appropriate      |
|                        | experience and training? If established, have they demonstrated an ongoing  |
|                        | record of accomplishments that have advanced their field(s)?               |
| Innovation             | Does the application challenge and seek to shift current research or clinical|
|                        | practice paradigms by utilizing novel theoretical concepts, approaches or   |
|                        | methodologies, instrumentation, or interventions? Are the concepts,         |
|                        | approaches or methodologies, instrumentation, or interventions novel to     |
|                        | one field of research or novel in a broad sense? Is a refinement,         |
|                        | improvement, or new application of theoretical concepts, approaches or      |
|                        | methodologies, instrumentation, or interventions proposed?                  |
| Approach                | Are the overall strategy, methodology, and analyses well reasoned and      |
|                        | appropriate to accomplish the specific aims of the project? Are potential  |
|                        | problems, alternative strategies, and benchmarks for success presented? If  |
|                        | the project is in the early stages of development, will the strategy        |
|                        | establish feasibility and will particularly risky aspects be managed?     |
| Environment            | Will the scientific environment in which the work will be done contribute   |
|                        | to the probability of success? Are the institutional support, equipment,    |
|                        | and other physical resources available to the investigators adequate for    |
|                        | the project proposed? Will the project benefit from unique features of the |
|                        | scientific environment, subject populations, or collaborative arrangements?|

Table 2. Current Scoring System.

| Overall Impact or Criterion Strength | Score | Descriptor |
|-------------------------------------|-------|------------|
| High                                | 1     | Exceptional|
|                                     | 2     | Outstanding|
|                                     | 3     | Excellent  |
| Medium                              | 4     | Very Good  |
|                                     | 5     | Good       |
|                                     | 6     | Satisfactory|
| Low                                 | 7     | Fair       |
|                                     | 8     | Marginal   |
|                                     | 9     | Poor       |
overall impact score (see NOT-OD-09-025 at http://grants.nih.gov/grants/guide/notice-files/not-od-09-025.html).

As for logistics of the review process, reviewers are recruited by the NIH from the scientific community and they serve on a voluntary basis. They are selected for their expertise in the field of science of the applications being reviewed, as established primarily by their standing in the scientific community (e.g., their position in their institutional hierarchies, history of funding, and publications). Grant applications for research are clustered into groups based on the fields of research and reviewed by committees of 20–40 reviewers in chartered study sections (chartered committees must ensure balanced representation according to geographical location, gender, and race/ethnicity as well as other legal requirements).

Each application is reviewed in-depth by three assigned reviewers. Assigned reviewers are usually given about 6 weeks to review the applications based on the criteria defined by the NIH. They record their scores for the five scored review criteria and use those evaluations as the basis for determining the overall impact score. The average of the overall impact scores from the three assigned reviewers is used to determine which applications will be discussed by the full committee in the review meeting. Usually, about half of the applications reviewed by each committee—the applications with the better average preliminary overall impact scores—are discussed in each review meeting. For discussed applications, all members of the committee record their overall impact scores, resulting in an average overall impact score or composite score. The applications are also usually assigned percentile scores based on where each application stands relative to other applications reviewed in the same study section.

Thus, a reviewer assigned to review an application now provides six integer scores, but while reviewers are asked to provide an overall impact score based largely on the five scored review criteria, specific guidance is not provided on translating the criteria scores into an overall impact score. Instead, reviewers use their own discretion in synthesizing an overall impact score from the criteria scores.

The availability of criteria scores now makes it possible to examine how assigned reviewers evaluate the NIH-defined criteria and how those criteria are related to the overall impact scores. Little if any peer-reviewed research has been published with these new scored review criteria. Some analyses have been posted on NIH blogs (e.g., see http://nexus.od.nih.gov/all/2011/03/08/overall-impact-and-criterion-scores/ and https://loop.nigms.nih.gov/2010/07/more-on-criterion-scores/) and those analyses suggest, at least to some, that reviewers may still base their overall impact scores largely on their assessment of the research methodology (i.e., the approach criterion), with little if any consideration for how innovative the projects are. This study was conducted to examine how the criterion scores assigned by reviewers are related to the overall impact score and to reexamine the assumption that (1) reviewers base their overall impact scores largely on their assessments of the approach, with little if any consideration for innovation or any of the other scored review criteria and (2) that reviewers actually give worse overall impact scores to the most innovative projects.

**Method**

Analyses of peer review data usually focus on the composite scores and percentiles produced by the entire committee, but the evaluations are driven by the five review criteria and scores for those criteria are only produced as part of the evaluations by the assigned reviewers. In addition, overall impact scores from individual reviewers for the applications that are not discussed, even the assigned reviewers, are deleted 15 days after review meetings. Therefore, the data analyzed in this study were the scores recorded by the assigned reviewers to new and competing continuation R01 applications reviewed in chartered study sections at the Center for Scientific Review at the NIH and those scores were captured within 15 days of the review meetings that were held between May and July 2013.
This time period allowed 4 years for the reviewers and applicants to develop familiarity with the new scoring procedures introduced in 2009 and captured a large number of records from applications reviewed across all NIH institutes, including applications that were discussed and those that were not discussed in the committee meetings, without filtering or excluding any area of science. Each record consisted of the five scored criteria scores and the overall impact score from one assigned reviewer. A total of 19,719 records were captured with missing scores in 8.5% of them, primarily from records that were captured before the final scores were assigned. Only records with missing scores were excluded from the analyses, which resulted in a total of 18,043 records (91.5% of the initially captured records) included in the analyses. Analyses were conducted to examine the pattern and distribution of each of the six scores. Correlations and regression analyses were also conducted to evaluate the relationships between the scored criteria and the overall impact score.

**Results**

Four of the five scored criteria exhibited moderate to severe restriction of range: 30–40% of the scores for investigators and environment were assigned the best possible score of “1” and 80–90% fell in the high (i.e., 1–3) range; only 10–12% of the scores for significance and innovation were assigned the highest possible score of “1” and 70–72% fell in the 2–4 range (Figure 1). Scores for the approach criterion and overall impact were more broadly distributed: Only 3–3.5% of the scores for approach and overall impact were assigned the highest possible score of “1” and 57–60% were in the 3–5 range (Figures 1 and 2). These different distributions resulted in differences in their overall means: approximately “2 = Outstanding” for environment and investigators, followed by “3 = Excellent” for innovation and significance, with the worst average score of “4 = Very Good” on the approach criterion (Figure 3 and Table 2).

All five scored criteria were related to the overall impact score. In terms of bivariate analyses, as the criteria scores increased, there was a monotonic increase in the overall impact score (Figure 4). The pattern of criteria scores for each overall impact score also showed that for an overall impact score of 1, virtually all the criteria scores were also 1 (Figure 5). In general, all criteria scores increased as the overall impact score increased, and if any of the criterion scores was greater than “3 = Excellent,” 98–99% of the overall impact scores were also greater than 3.

Tests for goodness of fit with the normal distribution (the Kolmogorov–Smirnov, Anderson–Darling, and Cramér–von Mises tests) showed that criteria scores were skewed and non-normally distributed (Table 3). All scores are integer values and their severely restricted range, highly skewed, non-normal, heteroscedastic distributions preclude data transformations and
 violate the assumptions of statistical tests that rely on normal distributions, including correlations and multiple regression (Cohen & Cohen, 1983, pp. 253–255; Nunnally, 1978, pp. 140–141; Thorndike, 1949, pp. 170–171).

**Figure 2.** Percentage of assigned reviewers’ scores at each level of the scoring system for overall impact.

**Figure 3.** Means and standard deviations of scores for each of the scored criteria and overall impact.
Scored review criteria were also all correlated with each other (Table 4), but correlations between the scored criteria and overall impact scores are more or less limited by differences in their restriction of range, so differences in the size of those correlations should not be overinterpreted. In other words, the larger correlation between the approach and the overall impact scores does not necessarily mean that it is more closely related to the overall impact score than any of the other scored review criteria, it may simply be due to the fact that the approach scores are much more widely distributed and have more variance than any of the other criteria scores.

The intercorrelations between all the scored review criteria create additional problems for their use in regression analyses. Regression analyses are robust under departures from many assumptions,
but a critical assumption is that predictor variables should be independent from each other (Cohen & Cohen, 1983; Farrar & Glauber, 1967). Substantial intercorrelations between predictor variables produce the problem of multicollinearity, which complicates the determination of how much unique variance in the overall impact scores should be attributed to each of the scored review criteria (Darlington, 1968).

In regression analyses with substantial multicollinearity among the explanatory variables, the proportion of variance attributed to each predictor variable will differ significantly depending on the order in which the variables are entered in the analysis (Cohen & Cohen, 1983; Darlington, 1968; Farrar & Glauber, 1967). A stepwise regression analysis of the five scored review criteria shows that approach is the best predictor variable; when entered first in the model, it accounts for 72% of the variance in overall impact scores (Table 5). Significance is the second predictor entered and accounts for an additional 4% of variance. Innovation, investigators, and environment then each account for less than 1% of the remaining variance. These results are consistent with previous reports, http://nexus.od.nih.gov/all/2011/03/08/overall-impact-and-criterion-scores/.

### Table 3. Goodness-of-Fit of Scored Criteria to Normal Distribution.

| Measure   | Skewness | Kurtosis | Kolmogorov–Smirnov D | p Value | Cramér–von Mises W² | p Value | Anderson–Darling A² | p Value |
|-----------|----------|----------|-----------------------|---------|----------------------|---------|---------------------|---------|
| Environment | 1.917    | 6.278    | 0.265                 | <.01    | 235.23               | <.005   | 1338.6              | <.005   |
| Investigators | 1.400   | 2.800    | 0.252                 | <.01    | 172.2                | <.005   | 989.0               | <.005   |
| Innovation  | 0.893    | 0.787    | 0.197                 | <.01    | 102.9                | <.005   | 570.2               | <.005   |
| Significance | 0.872   | 0.530    | 0.190                 | <.01    | 102.4                | <.005   | 577.1               | <.005   |
| Approach    | 0.243    | -0.562   | 0.125                 | <.01    | 46.82                | <.005   | 275.1               | <.005   |

### Table 4. Correlations.\(^a\)

| Overall Impact | Environment | Investigators | Significance | Innovation | Approach |
|----------------|-------------|---------------|--------------|------------|----------|
| Overall Impact | 1.00        | 0.43          | 0.55         | 0.67       | 0.61     | 0.85     |
| Environment    | 0.43        | 1.00          | 0.62         | 0.38       | 0.40     | 0.42     |
| Investigators   | 0.55        | 0.62          | 1.00         | 0.46       | 0.46     | 0.53     |
| Significance    | 0.67        | 0.38          | 0.46         | 1.00       | 0.61     | 0.60     |
| Innovation     | 0.61        | 0.40          | 0.47         | 0.61       | 1.00     | 0.56     |
| Approach       | 0.85        | 0.42          | 0.53         | 0.60       | 0.56     | 1.00     |

\(^a\)Correlations between the five scored criteria scores and the overall impact scores. All correlations are significant at \(p < .0001\).

### Table 5. Evidence of Multicollinearity.

| Step | Variable Entered | Partial \(R^2\) | Predictors Entered in Reverse Order | Partial \(R^2\) | Variable Entered |
|------|------------------|-----------------|-------------------------------------|-----------------|-----------------|
| 1    | Approach         | .72295          |                                     | .18853          | Environment     |
| 2    | Significance     | .04180          |                                     | .12882          | Investigators   |
| 3    | Innovation       | .00829          |                                     | .15299          | Innovation     |
| 4    | Investigators     | .00385          |                                     | .09180          | Significance    |
| 5    | Environment      | .00002          |                                     | .21477          | Approach        |
|      | Model \(R^2\)    | .77691          |                                     | .77691          | Model \(R^2\)   |
However, if the variables are entered in reverse order, each of the five scored review criteria account for approximately 10–20% of the variance in overall impact scores (Table 5). Environment accounts for 19% of the variance, investigators account for an additional 13% of the variance, innovation accounts for an additional 15% of variance, and the approach accounts for 21% of additional variance. The five scored review criteria account for 78% of the variance in overall impact scores, regardless of whatever order the predictors are entered, but the amount of unique variance attributed to each predictor varies significantly depending on their restriction of range and the order that they are entered in the model.

Discussion

Peer review of grant applications at the NIH is based on the review criteria defined by the NIH (Table 1), and recent changes in NIH review procedures have made it possible to examine how the assigned reviewers evaluate the review criteria and how those evaluations are related to the overall impact score. This study examined the pattern and distribution of the scored review criteria and overall impact scores recorded by assigned reviewers for new and renewal R01 applications reviewed in chartered study sections at the NIH by the Center for Scientific Review in 2013, using the revised review criteria and integer scoring first instituted 4 years earlier, in October 2009.

While acknowledging the points raised (e.g., Azoulay et al., 2009; Kaplan, 2011; Kolata, 2009), the present results suggest that all the scored review criteria, not just the approach scores, are related to the overall impact scores. As the criteria scores increased, there was a monotonic increase in the overall impact score, and if any of the criterion scores was greater than “3 = Excellent,” the overall impact score was also greater than 3. In particular, the scores for innovation are closely related to the overall impact scores and there was no evidence of any inverse correlation between scores for innovation and overall impact scores. Furthermore, reviewers assigned the best possible score for innovation on 11% of their critiques and 67% of scores for innovation were in the high range (i.e., 1–3), which suggests that reviewers are not overly critical or unwilling to assign good scores for Innovation.

Step-wise regression analyses seem to support the view that the approach score is used almost exclusively by reviewers. However, the present study shows that differences in distributions, the combination of integer values with severe restriction of range, and significant multicollinearity between the criteria scores, all complicate the interpretation of regression analyses, making it impossible with these data to determine how much unique variance in overall impact scores is accounted for by each scored review criteria. It might be possible to determine how reviewers weight each of the scored review criteria if the criteria could be independently manipulated, experimentally, but it is difficult to determine how such a study might be conducted. All that can be said based on the present, retrospective analysis is that all the scored criteria combined account for almost 80% of the variance in the overall impact scores.

It is important to note that there are additional review criteria defined by the NIH that do not receive separate scores and those criteria also contribute to the determination of the overall impact scores. For example, reviewers are asked to consider if the risks and protections for human participants and animal subjects have been adequately discussed and addressed, and if biohazardous materials are being handled and disposed of properly, and if so, those issues are reflected in the overall impact score. Since separate scores are not assigned for those items, they are not included in the present analyses, but those factors undoubtedly account for some of the variance that was unaccounted for by the scored review criteria.

It is not possible in the present study to determine the causes of the differences in the distributions of the scored criteria; it may be that judgments about one criterion naturally condition the judgments of other criteria in complex ways that are hard to tease out in archival data. However, it is important
to recognize that the review criteria shape applications in ways that are not always apparent based on the distinctions made between the applications during peer review. For example, improved communication of the review criteria and a requirement for formatting an application so that all the criteria are explicitly addressed might have improved the quality of the entire pool of submitted applications and resulted in ceiling or floor effects for some of the criteria among submitted applications.

With respect to the approach, it is not clear that the wider range of scores on this criterion is due to reviewers focusing more on the approach than any of the other criteria. As an apical measure, weaknesses in all the other criteria often naturally contribute to the approach score. For example, the approach includes the experimental design, methods, procedures, and analytical plan. Weaknesses in the experimental design are sometimes related to lack of adequate expertise in biostatistics, or expertise with critical technology or specific patient populations, which affects scores for both investigator and approach. Lack of access to appropriate equipment or patient/subject populations also often affect both scores for environment and approach. In addition, if the significance of the project is not clearly articulated or adequately supported, it can be difficult to determine if the approach will achieve those poorly defined objectives, which can affect both the scores for significance and approach. Weaknesses in innovation may also affect approach scores. A study that lacks any innovation is less likely to receive the best scores for the approach. Therefore, weaknesses in each of the other scored criteria often combine to detract from the approach score and result in approach scores which are worse and more widely distributed than any of the other scored criteria.

The wide distribution and relatively large number of worse scores for the approach criterion may also reflect legitimate, significant weaknesses in those areas of the applications covered by the approach criterion. That possibility is supported by numerous publications in the literature. For example, most studies published in the literature have grossly inadequate sample size and statistical power (Bakker, van Dijk, & Wicherts, 2012; Button et al., 2013), fail to adequately identify reagents and their sources (Vasilevsky et al., 2013), and are designed and conducted to confirm, not to provide the most rigorous and challenging test of the investigator’s hypotheses (for review, see Lindner, 2007; Mahoney & Kimper, 2004). Such weaknesses in the experimental design, methods, procedures, and analytical plan are not only very common but are also extremely consequential. They contribute to high rates of false positive results, inflated effect sizes, and poor reproducibility (Ioannidis, 2005a, 2005b, 2008; Young, Ioannidis, & Al-Ubaydli, 2008) and there is increasing evidence that many of the findings published in the literature are not reproducible (Begley & Ellis, 2012; Lohmueller, Pearce, Pike, Lander, & Hirschhorn, 2003; Prinz, Schlange, & Asadullah, 2011; Steward, Popovich, Dietrich, & Kleitman, 2012; Vines et al., 2009).

In conclusion, the availability of criteria scores now makes it possible to examine how NIH reviewers are evaluating the review criteria. While there are always limitations to interpreting the causal influences among variables that are all reported at the same time, the results nonetheless suggest that all the scored review criteria are related to the overall impact score and good scores are necessary on all five scored review criteria in order to achieve a good overall impact score.

Acknowledgment
The views expressed in this article are those of the authors and do not necessarily represent those of CSR, NIH, or the U.S. Dept. of Health and Human Services. The authors thank Jim Onken, at the Office of Extramural Research, for his critical advice, and Amanda Manning (CSR) for technical support. The authors also thank the associate editor, George Julnes, and anonymous reviewers at AJE for their helpful critiques.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

References
Azoulay, P., Zivin, J. S. G., & Manso, G. (2009). Incentives and creativity: Evidence from the academic life sciences. The RAND Journal of Economics, 42, 527–554.

Bacon, F. (1620). Novum Organum: True directions concerning the interpretation of nature (with other parts of The Great Instauration). Chicago, IL: Open Court.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. Perspectives on Psychological Science, 7, 543–554.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. Nature, 483, 531–533.

Bernard, C. (1865). An introduction to the study of experimental medicine. New York, NY: Dover Publications.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14, 365–376.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences. (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Coryn, C. L. S., Hattie, J. A., Scriven, M., & Hartmann, D. J. (2007). Models and mechanisms for evaluating government-funded research: An international comparison. American Journal of Evaluation, 28, 437–457.

Darlington, R. B. (1968). Multiple regressio in psychological research and practice. Psychological Bulletin, 69, 161–182.

Demicheli, V., & Di Pietrantonj, C. (2007). Peer review for improving the quality of grant applications. Cochrane Database of Systematic Reviews, 2, MR000003.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisied. The Review of Economics and Statistics, 49, 92–107.

Henley, C. (1977). Peer review of research grant appliations at the National Institutes of Health 2: Review by an Initial Review Group. Federation Proceedings, 36, 2186–2190.

Ioannidis, J. P. (2005a). Contradicted and initially stronger effects in highly cited clinical research. Journal of the American Medical Association, 294, 218–228.

Ioannidis, J. P. (2005b). Why most published research findings are false. PLoS Medicine, 2, e124.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. Epidemiology, 19, 640–648.

Kaplan, D. (2011). Social choice at NIH: The principle of complementarity. FASEB Journal, 25, 3763–3764.

Kolata, G. (2009, January 29). Grant system leads cancer researchers to play it safe (p. A1). The New York Times.

Kuhn, T. S. (1996). The structure of scientific revolutions. (3rd ed.) Chicago, IL: The University of Chicago Press.

Kuhn, T. S. (1975). The Copernican revolution: Planetary astronomy in the development of western thought. Cambridge, MA: Harvard University Press.

Lindner, M. D. (2007). Clinical attrition due to biased preclinical assessments of potential efficacy. Pharmacology & Therapeutics, 115, 148–175.

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., & Hirschhorn, J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nature Genetics, 33, 177–182.

Mahoney, M. J., & Kimper, T. P. (2004). From ethics to logic: A survey of scientists. In M. J. Mahoney (Ed.), Scientist as subject: The psychological imperative (pp. 187–193). New York, NY: Percheron Press.
National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter future.* Washington, DC: The National Academies.

Nunnally, J. C. (1978). *Psychometric theory.* (2nd ed.) New York, NY: McGraw-Hill Book Company.

Planck, M. (1949). *Scientific autobiography and other papers.* New York, NY: Philosophical Library, translated from German by Frank Gaynor.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10,* 712.

Steward, O., Popovich, P. G., Dietrich, W. D., & Kleitman, N. (2012). Replication and reproducibility in spinal cord injury research. *Experimental Neurology, 233,* 597–605.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques.* New York, NY: John Wiley.

Vasilevsky, N. A., Brush, M. H., Paddock, H., Ponting, L., Tripathy, S. J., Larocca, G. M., & Haendel, M. A. (2013). On the reproducibility of science: Unique identification of research resources in the biomedical literature. *PeerJ, 1,* e148.

Vineis, P., Manuguerra, M., Kavvoura, F. K., Guerrera, S., Allione, A., & Rosa, F., . . . Matullo, G. (2009). A field synopsis on low-penetrance variants in DNA repair genes and cancer susceptibility. *Journal of the National Cancer Institute, 101,* 24–36.

Young, N. S., Ioannidis, J. P., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine, 5,* e201.