Special Issue Article

The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology

Social Psychological and Personality Science
2017, Vol. 8(4) 363-369

DOI: 10.1177/1948550616673876
journals.sagepub.com/home/spp
SAGE

Nicholas J. L. Brown[1] and James A. J. Heathers[2]

Abstract

We present a simple mathematical technique that we call granularity-related inconsistency of means (GRIM) for verifying the summary statistics of research reports in psychology. This technique evaluates whether the reported means of integer data such as Likert-type scales are consistent with the given sample size and number of items. We tested this technique with a sample of 260 recent empirical articles in leading journals. Of the articles that we could test with the GRIM technique ($N = 71$), around half ($N = 36$) appeared to contain at least one inconsistent mean, and more than 20% ($N = 16$) contained multiple such inconsistencies. We requested the data sets corresponding to 21 of these articles, receiving positive responses in 9 cases. We confirmed the presence of at least one reporting error in all cases, with three articles requiring extensive corrections. The implications for the reliability and replicability of empirical psychology are discussed.

Keywords

research methods, philosophy of science, advanced quantitative methods

Consider the following (fictional) extract from a recent article in the *Journal of Porcine Aviation Potential*:

> Participants ($N = 55$) were randomly assigned to drink 200 ml of water that either contained (experimental condition, $N = 28$) or did not contain (control condition, $N = 27$) 17 g of cherry flavor Kool-Aid® powder. Fifteen minutes after consuming the beverage, participants responded to the question, "To what extent do you believe that pigs can fly?" on a seven-point scale from 1 (*Not at all*) to 7 (*Definitely*). Participants in the "drank the Kool-Aid" condition reported a significantly stronger belief in the ability of pigs to fly ($M = 5.19$, $SD = 1.34$) than those in the control condition ($M = 3.86$, $SD = 1.41$), $t(53) = 3.59$, $p < .001$.

These results seem superficially reasonable but are actually mathematically impossible. The reported means represent errors of transcription, some version of misreporting, or the deliberate manipulation of results. Specifically, the mean of the 28 participants in the experimental condition, reported as 5.19, cannot be correct. Since all responses were integers between 1 and 7, the total of the response scores across all participants must fall in the range 28–196. The two integers that give a result closest to the reported mean of 5.19 are 145 and 146. However, 145 divided by 28 is $5.17\overline{857142}$, which conventional rounding returns as 5.18. Likewise, 146 divided by 28 is $5.2\overline{1428571}$, which rounds to 5.21. That is, there is no combination of responses that can give a mean of 5.19 when correctly rounded. Similar considerations apply to the reported mean of 3.86 in the control condition: Multiplying this value by the sample size (27) gives 104.22, suggesting that the total score across participants must have been either 104 or 105. But 104 divided by 27 is $3.\overline{851}$, which rounds to 3.85, and 105 divided by 27 is $3.88\overline{8}$, which rounds to 3.89.

In this article, we first introduce the general background to and calculation of what we term the granularity-related inconsistency means (GRIM) test. Next, we report on the results of an analysis using the GRIM test of a number of published articles from leading psychological journals. Finally, we discuss the implications of these results for the published literature in empirical psychology.

General Description of the GRIM Technique for Reanalyzing Published Data

Participant response data collected in psychology are typically ordinal in nature—that is, the recorded values have meaning in terms of their rank order, but the numbers representing them

[1] University Medical Center, University of Groningen, The Netherlands
[2] Division of Cardiology and Intensive Therapy, Poznań University of Medical Sciences, University of Sydney, Poland

Corresponding Author:
Nicholas J. L. Brown, University Medical Center, University of Groningen, Antonius Deusinglaan 1, 9713 AV Groningen, The Netherlands.
Email: nick.brown@free.fr

are arbitrary, such that the value corresponding to any item has no significance beyond its ability to establish a position on a continuum relative to the other numbers. For example, the 7-point scale cited in our opening example, running from 1 to 7, could equally well have been coded from 0 to 6, or from 6 to 0, or from 10 to 70 in steps of 10. However, while the limits of ordinal data in measurement have been extensively discussed for many years (e.g., Carifio & Perla, 2007; Coombs, 1960; Jamieson, 2004; Thurstone, 1927), it remains common practice to treat ordinal data composed of small integers as if they were measured on an interval scale, calculate their means and standard deviations (*SD*s), and apply inferential statistics to those values. Other common measures used in psychological research produce genuine interval-level data in the form of integers; for example, one might count the number of anagrams unscrambled, or the number of errors made on the Stroop test, within a given time interval. Thus, psychological data often consist of integer totals, which are then typically divided by the sample size to give the mean.

One often overlooked property of data derived from such noncontinuous measures, whether ordinal or interval, is their *granularity*—that is, the numerical separation between possible values of the summary statistics. Here, we consider the example of the mean. With typical Likert-type data, the smallest amount by which two means can differ is the reciprocal of the product of the number of participants and the number of items (questions) that make up the scale. For example, if we administer a 3-item Likert-type measure to 10 people, the smallest amount by which two mean scores can differ (the granularity of the mean) is $(1/(10 \times 3)) = 0.03\bar{3}$. If means are reported to two decimal places, then—although there are 100 possible numbers with two decimal places in the range $1 \leq X < 2$ (1.00, 1.01, 1.02, etc., up to 1.99)—the possible values of the (rounded) mean are considerably fewer (1.00, 1.03, 1.07, 1.10, etc., up to 1.97). If the number of participants (N) is less than 100 and the measured quantity is an integer, then not all of the possible sequences of two digits can occur after the decimal point in correctly rounded fractions. We use the term *inconsistent* to refer to reported means of integer data whose value, appropriately rounded, cannot be reconciled with the stated sample size. (More generally, if the number of decimal places reported is D, then some combinations of digits will not be consistent if N is less than 10^D.)

This relation is always true for integer data that are recorded as single items, such as participants' ages in whole years, or a 1-item Likert-type measure, as is frequently used as a manipulation check. In particular, the number of possible responses to each item is irrelevant; that is, it makes no difference whether responses can range from 0 to 3 or from 1 to 100. When a composite measure is used, such as one with 3 Likert-type items where the mean of the item scores is taken as the value of the measure, this mean value will not necessarily be an integer; instead, it will be some multiple of $(1/L)$, where L is the number of items in the measure. Similar considerations would apply to a hypothetical 1-item measure where the possible responses are simple fractions instead of integers. For example, a scale with possible responses of 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, and 3.0 would be equivalent to a 2-item measure with integer responses in the range 0–3. Alternatively, in a money game where participants play with quarters, and the final amount won or lost is expressed in dollars, only values ending in 0.00, 0.25, 0.50, or 0.75 are possible. However, the range of possible values that such means can take is still constrained (e.g., in a 3-item Likert-type scale, assuming item scores starting at 1, this range will be 1.00, $1.3\bar{3}$, $1.6\bar{6}$, 2.00, $2.3\bar{3}$, etc.) and so for any given sample size, the range of possible values for the mean of all participants is also constrained. For example, with a sample size of 20 and $L = 3$, possible values for the mean are 1.00, 1.02 (rounded from $1.01\bar{6}$), 1.03 (rounded from $1.03\bar{3}$), 1.05, 1.07, and so on. More generally, the range of means for a measure with L items (or an interval scale with an implicit granularity of $[1/L]$, where L is a small integer, such as 4 in the example of the game played with quarters) and a sample size of N is identical to the range of means for a measure with 1 item and a sample size of $L \times N$. Thus, by multiplying the sample size by the number of items in the scale, composite measures can be analyzed using the GRIM technique in the same way as single items, although as the number of scale items increases, the maximum sample size for which this analysis is possible is correspondingly reduced as the granularity decreases toward 0.01. We use the term *GRIM-testable* to refer to variables whose granularity (typically, 1 divided by the product of the number of scale items and the number of participants) is sufficiently large that they can be tested for inconsistencies with the GRIM technique. For example, a 5-item measure with 25 participants has the same granularity (0.008) as a 1-item measure with 125 participants, and hence scores on this measure are not typically GRIM-testable.

Figure 1 shows the distribution of consistent (shown in white) and inconsistent (shown in black) means as a function of the sample size. Note that only the two-digit fractional portion of each mean is linked to consistency; the integer portion plays no role. The overall pattern is clear: As the sample size increases, the number of means that are consistent with that sample size also increases, and so the chance that any single incorrectly reported mean will be detected as inconsistent is reduced. However, even with quite large sample sizes, it is still possible to detect inconsistent means if an article contains multiple inconsistencies. For example, consider a study with $N = 75$ and six reported "means" whose values have, in fact, been chosen at random: There is a 75% chance that any one random mean will be consistent, but only a 17.8% (0.75^6) chance that all six will be.

Our general formula, then, is that when the number of participants (N) is multiplied by the number of items composing a measured quantity (L, commonly equal to 1), and the means that are based on N are reported to D decimal places, then if $(L \times N) < 10^D$, there exists some number of decimal fractions of length D that cannot occur if the means are reported correctly. The number of inconsistent values is generally equal to $(10^D - N)$; however, in the analyses reported in the present

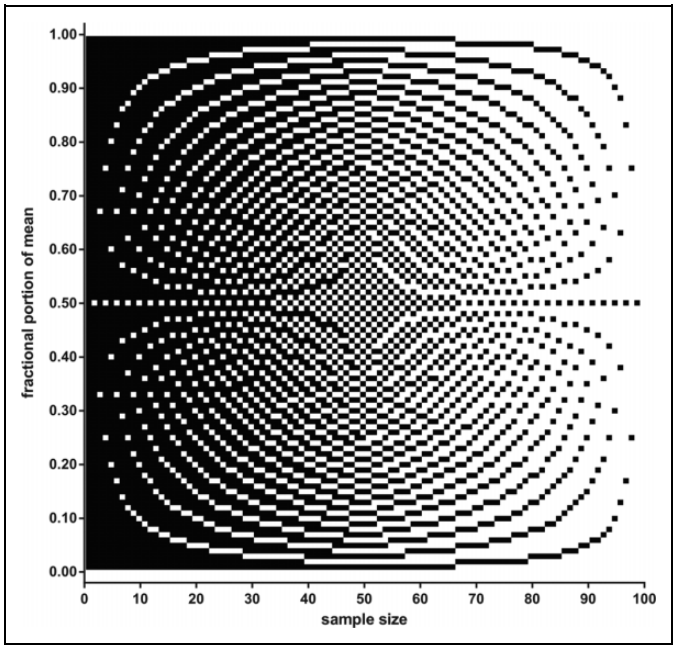

Figure 1. Plot of consistent (white dots) and inconsistent (black dots) means reported to two decimal places. *Note.* As the sample size increases toward 100, the number of means that are consistent with that sample size also increases, as shown by the greater number of white (vs. black) dots. Thus, granularity-related inconsistency of means works better with smaller sample sizes, as the chance of any individual incorrectly reported mean being consistent by chance is lower. The *Y*-axis represents only the fractional portion of the mean (i.e., the part after the decimal point), because the integer portion of the mean plays no role. That is, for any given sample size, if a mean of 2.49 is consistent with the sample size, then means of 0.49 or 8.49 are also consistent. This figure assumes that means ending in 5 at the third decimal place (e.g., 10/80 = 0.125) are always rounded up; if such means are allowed to be rounded up or down, a few extra white dots will appear at sample sizes that are multiples of 8.

article, we conservatively allowed numbers ending in exactly 5 at the third decimal place to be rounded either up or down without treating the resulting means as inconsistent, so that some values of N have fewer possible inconsistent means than this formula indicates.

Using the GRIM technique, it is possible to examine published reports of empirical research to see whether the means have been reported correctly.[1] Psychological journals typically require the reporting of means to two decimal places, in which case the sample size corresponding to each mean must be less than 100 in order for its consistency to be checked. However, since the means of interest in experimental psychology are often those for subgroups of the overall sample (e.g., the numbers in each experimental condition), it can still be possible to apply the GRIM technique to studies with overall sample sizes substantially above 100. (Note that percentages reported to only one decimal place can typically be tested for consistency with a sample size of up to 1,000, as they are, in effect, fractions reported to three decimal places.)

We now turn to our pilot trial of the GRIM test.

Method

We searched recently published (2011–2015) issues of *Psychological Science* (*PS*), *Journal of Experimental Psychology: General* (*JEP:G*), and *Journal of Personality and Social Psychology* (*JPSP*) for articles containing the word "Likert" anywhere in the text. This strategy was chosen because we expected to find Likert-type data reported in most of the articles containing that word (although we also checked the consistency of the means of other integer data where possible). We sorted the results with the most recent first and downloaded at most the first 100 matching articles from each journal. Thus, our sample consisted of 100 articles from *PS* published between January 2011 and December 2015, 60 articles from *JEP:G* published between January 2011 and December 2015, and 100 articles from *JPSP* published between October 2012 and December 2015.

We examined the Method section of each study reported in these articles to see whether GRIM-testable measures were used and to determine the sample sizes for the study and, where appropriate, each condition. A preliminary check was performed by the first author; if he did not see evidence of either GRIM-testable measures or any (sub)sample sizes less than 100, the article was discarded. Subsequently, each author worked independently on the retained articles. We examined the table of descriptives (if present), other result tables, and the text of the Results section, looking for means or percentages that we could check using the GRIM technique. On the basis of our tests, we assigned each article a subjective "inconsistency level" rating. A rating of 0 (*no problems*) meant that all the means we were able to check were consistent, even if those means represented only a small percentage of the reported data in the article. We assigned a rating of 1 (*minor problems*) to articles that contained only one or two inconsistent numbers, where we believed that these were most parsimoniously explained by typographical or transcription errors, and where an incorrect value would have little effect on the main conclusions of the article. Articles that had a small number of inconsistencies that might impact the principal results were given a rating of 2 (*moderate problems*); we also gave this rating to articles in which the results seemed to be uninterpretable as described. Finally, we applied a rating of 3 (*substantial problems*) to articles with a larger number of inconsistencies, especially if these appeared at multiple points in the article. Finally, ratings were compared between the authors and differences resolved by discussion.

Results

The total number of articles examined from each journal, the number retained for GRIM analysis, and the number to which we assigned each rating, are shown in Table 1. A total of 260 articles were initially examined. Of these, 189 (72.7%) were discarded, principally because either they reported no GRIM-testable data or their sample sizes were all sufficiently large that no inconsistent means were likely to be detected. Of the

Table 1. Journals and Articles Consulted.

Journal	*PS*	*JEP: G*	*JPSP*	Total
Number of articles	100	60	100	260
Earliest article date	January 2011	January 2011	October 2012	
Articles with GRIM-testable data	29	15	27	71
Level 0 articles (no problems detected)	16	8	11	35
Level 1 articles (minor problems)	5	3	7	15
Level 2 articles (moderate problems)	1	1	3	5
Level 3 articles (substantial problems)	7	3	6	16

Note. *PS* = *Psychological Science*; *JEP: G* = *Journal of Experimental Psychology: General*; *JPSP* = *Journal of Personality and Social Psychology*; GRIM = granularity-related inconsistency of means.

remaining 71 articles, 35 (49.3%) reported all GRIM-testable data consistently and were assigned an inconsistency level rating of 0. That left us with 36 articles that appeared to contain one or more inconsistency. Of these, we assigned a rating of 1–15 articles (21.1% of the 71 in total for which we performed a GRIM analysis), a rating of 2–5 articles (7.0%), and a rating of 3–16 articles (22.5%). In some of these "Level 3" articles, over half of the GRIM-testable values were inconsistent with the stated sample size.

Next, we e-mailed[2] the corresponding authors of the articles that were rated at Levels 2 or 3, asking for their data. In response to our 21 initial requests, we received 11 replies within 2 weeks. At the end of that period, we sent follow-up requests to the 10 authors who had not replied to our initial e-mail. In response to either the first or second e-mail, we obtained the requested data from eight authors, while a ninth provided us with sufficient information about the data in question to enable us to check the consistency of the means. Four authors promised to send the requested data but have not done so to date. Five authors either directly or effectively refused to share their data, even after we explained the nature of our study; interestingly, two of these refusals were identically worded. In another case, the corresponding author's personal e-mail address had been deleted; another author informed us that the corresponding author had left academia and that the location of the data was unknown. Finally, two of our requests went completely unanswered after the second e-mail.

Our examination of the data that we received showed that the GRIM technique identified one or more genuine problem in each case. We report the results of each analysis briefly here, in the order in which the data were received.

Data Set 1

Our GRIM analysis had detected two inconsistent means in a table of descriptives as well as eight inconsistent *SD*s.[3] Examining the data, we found that the two inconsistent means and one of the inconsistent *SD*s were caused by the sample size for that cell not corresponding to the sample size for the column of data in question; five *SD*s had been incorrectly rounded because the default (three decimal places) setting of SPSS had caused a value of 1.2849 to be rounded to 1.285, which the authors had subsequently rounded manually to 1.29; and two further *SD*s appeared to have been incorrectly transcribed, with values of 0.79 and 0.89 being reported as 0.76 and 0.86, respectively. All of these errors were minor and had no substantive effect on the published results of the article.

Data Set 2

Our reading of the article in this case had detected several inconsistent means, as well as several inconsistently reported degrees of freedom and apparent errors in the reporting of some other statistics. Examination of the data confirmed most of these problems and indeed revealed a number of additional errors in the authors' analysis. We subsequently discovered that the article in question had already been the subject of a correction in the journal, although that had not addressed most of the problems that we found. We intend to write to the authors to suggest a number of points that require (further) correction.

Data Set 3

In this case, our GRIM analysis had shown a large number of inconsistent means in two tables of descriptives. The corresponding author provided us with an extensive version of the data set, including some intermediate analysis steps. We identified that most of the entries in the descriptives had been calculated using a Microsoft Excel formula that included an incorrect selection of cells; for example, this resulted in the mean and *SD* of the first experimental condition being included as data points in the calculation of the mean and *SD* of the second. The author has assured us that a correction will be issued.

Data Set 4

In the e-mail accompanying their data, the authors of this article spontaneously apologized in advance (even though we had not yet told them exactly why we were asking for their data) for possible discrepancies between the sample sizes in the data and those reported in the article. They stated that, due to computer-related issues, they had only been able to retrieve an earlier version of the data set rather than the final version on which the article was based. We adjusted the published sample sizes using the notes that the authors

provided and found that this adequately resolved the GRIM inconsistencies that we had identified.

Data Set 5

The GRIM analyses in this case found some inconsistent means in the reporting of the data that were used as the input to a number of *t* tests, as well as in the descriptives for one of the conditions in the study. Analysis revealed that the former problems were the result of the authors having reported the *N*s from the output of a repeated-measures analysis of variance in which some cases were missing, so that these *N*s were smaller than those reported in the Method section. The problems in the descriptives were caused by incorrect reporting of the number of participants who were excluded from the analyses. We were unable to determine to what extent this difference affected the results of the study.

Data Set 6

Here, the inconsistencies that we detected were mostly due to the misreporting by the authors of their sample size. This was not easy to explain as a typographical error, as the number was reported as a word at the start of a sentence (e.g., ''Sixty undergraduates took part''). Additionally, one inconsistent *SD* turned out to have been incorrectly copied during the drafting process.

Data Set 7

This data set confirmed numerous inconsistencies, including large errors in the reported degrees of freedom for several *F* tests, from which we had inferred the per-cell sample sizes. Furthermore, a number that was meant to be the result of subtracting 1 Likert-type item score from another (thus giving an integer result) had the impossible value of 1.5. We reported these inconsistencies to the corresponding author but received no acknowledgment.

Data Set 8

The corresponding author indicated that providing the full data set could be complicated, as the data were taken from a much larger longitudinal study. Instead, we provided a detailed explanation of the specific inconsistencies we had found. The author checked these and confirmed that the sample size of the study in question had been reported incorrectly, as several participants had been excluded from the analyses but not from the reported count of participants. The author thanked us for finding this minor (to us) inconsistency and described the exercise as ''a good lesson.''

Data Set 9

In this case, we asked for data for three studies from a multiple-study article. In the first two studies, we found some reporting problems with *SD*s in the descriptives and some other minor problems to do with the handling of missing values for some variables. For the third study, however, the corresponding author reported that, during the process of preparing the data set to send to us, an error in the analyses had been discovered that was sufficiently serious as to warrant a correction to the published article.

For completeness, we should also mention that in one of the cases above, the data that we received showed that we had failed to completely understand the original article; what we had thought were inconsistencies in the means on a Likert-type measure were due to that measure being a multiple-item composite, and we had overlooked that it was correctly reported as such. While our analysis also discovered separate problems with the article in question, this underscores how careful reading is always necessary when using the GRIM technique.

Discussion

We identified a simple method for detecting discrepancies in the reporting of statistics derived from integer-based data and applied it to a sample of empirical articles published in leading journals of psychology. Of the articles that we were able to test, around half appeared to contain one or more errors in the summary statistics. (We have no way of knowing how many inconsistencies might have been discovered in the articles with larger samples, had it been standard practice to report means to three decimal places.) Nine data sets were examined in more detail, and we confirmed the existence of reporting problems in all nine, with three articles requiring formal corrections.

We anticipate that the GRIM technique could be a useful tool for reviewers and editors. A GRIM check of the reported means of an article submitted for review ought to take only a few minutes. (Indeed, we found that even when no inconsistencies were uncovered, simply performing this check enhanced our understanding of the methods used in the articles that we read.) When GRIM errors are discovered, depending on their extent and how the reviewer feels they impact the article, actions could range from asking the authors to check a particular calculation, to informing the action editor confidentially that there appear to be severe problems with the manuscript.

When an inconsistent mean is uncovered by this method, we of course have no information about the *true* mean value that was obtained; that can only be determined by a reanalysis of the original data. But such an inconsistency does indicate, at a minimum, that a mistake has been made. When multiple inconsistencies are demonstrated in the same article, we feel that the reader is entitled to question what else might not have been reported accurately. Note also that not all incorrectly reported means will be detected using the GRIM technique, because such a mean can still be consistent by chance. With reporting to two decimal places, for a sample size $N < 100$, a random mean value will be consistent in approximately $N\%$ of cases. Thus, the number of GRIM errors detected in an article is likely to be a conservative estimate of the true number of such errors.

A limitation of the GRIM technique is that, with the standard reporting of means to two decimal places, it cannot

reveal inconsistencies with per-cell sample sizes of 100 or more, and its ability to detect such inconsistencies decreases as the sample size (or the number of items in a composite measure) increases. However, this still leaves a substantial percentage of the literature that can be tested. Recall that we selected our articles from some of the highest impact journals in the field; it might be that other journals have a higher proportion of smaller studies. Additionally, it might be the case that smaller studies are more prone to reporting errors (e.g., because they are run by laboratories that have fewer resources for professional data management).

A further potential source of false positives is the case where one or more participants are missing values for individual items in a composite measure, thus making the denominator for the mean of that measure smaller than the overall sample size. However, in our admittedly modest sample of articles, this issue only caused inconsistencies in one case. We believe that this limitation is unlikely to be a major problem in practice because the GRIM test is typically not applicable to measures with a large number of items, due to the requirement for the product of the per-cell sample size and the number of items to be less than 100.

Concluding Remarks

On its own, the discovery of one or more inconsistent means in a published article need not be a cause for alarm; indeed, we discovered from our reanalysis of data sets that in many cases where such inconsistencies were present, there was a straightforward explanation, such as a minor error in the reported sample sizes, or a failure to report the exclusion of a participant. Sometimes, too, the reader performing the GRIM analysis may make errors, such as not noticing that what looks like a single Likert-type item is in fact a composite measure.

It might also be that psychologists are simply sometimes rather careless in retyping numbers from statistical software packages into their articles. However, in such cases, we think it is legitimate to ask how many other elementary mistakes might have been made in the analysis of the data, and with what effects on the reported results. It is interesting to compare our experiences with those of Wolins (1962), who asked 37 authors for their data, obtained these in usable form from seven authors, and found "gross errors" in three cases. While the numbers of studies in both Wolins' and our cases are small, the percentage of severe problems is, at an anecdotal level, worrying. Indeed, we wonder whether some proportion of the failures to replicate published research in psychology (Open Science Collaboration, 2015) might simply be due to the initial (or, conceivably, the replication) results being the products of erroneous analyses.

Beyond inattention and poorly designed analyses, however, we cannot exclude that in some cases, a plausible explanation for GRIM inconsistencies is that some form of data manipulation has taken place. For example, in the fictional extract at the start of this article, here is what should have been written in the last sentence:

> Participants in the "drank the Kool-Aid" condition did not report a significantly stronger belief in the ability of pigs to fly ($M = 4.79$, $SD = 1.34$) than those in the control condition ($M = 4.26$, $SD = 1.41$), $t(53) = 1.43$, $p = .16$.

In the "published" extract, compared to the above version, the first mean was "adjusted" by adding 0.40 and the second by subtracting 0.40. This transformed a nonsignificant p value into a significant one, thus making the results considerably easier to publish (cf. Kühberger, Fritz, & Scherndl, 2014).

We are particularly concerned about the 8 data sets (out of the 21 we requested) that we believe we may never see (5 due to refusals to share the data, 2 due to repeated nonresponse to our requests, and 1 due to the apparent disappearance of the corresponding author). Refusing to share one's data for reanalysis without giving a clear and relevant reason is, we feel, professionally disrespectful at best, especially after authors have assented to such sharing as a condition of publication, as is the case in, for example, American Psychological Association journals such as *JPSP* and *JEP:G*. We support the principle, currently being adopted by several journals, that sharing of data ought to be the default situation, with authors having to provide strong arguments why their data cannot be shared in any given case. When accompanied by numerical evidence that the results of a published article may be unreliable, a refusal to share data will inevitably cause speculation about what those data might reveal. However, throughout the present article, we have refrained from mentioning the titles, authors, or any other identifying features of the articles in which the GRIM analysis identified apparent inconsistencies. There are three reasons for this. First, the GRIM technique was exploratory when we started to examine the published articles, rather than an established method. Second, there may be an innocent explanation for any or all of the inconsistencies that we identified in any given article. Third, it is not our purpose here to "expose" anything or anyone; we offer our results in the hope that they will stimulate discussion within the field. It would appear, as a minimum, that we have identified an issue worthy of further investigation, and produced a tool that might assist reviewers of future work, as well as those who wish to check certain results in the existing literature.

Acknowledgments

The authors wish to thank Tim Bates and Chris Chambers for their helpful comments on an earlier draft of this article, as well as those authors of articles that we examined who kindly provided their data sets and help with the reanalysis of these.

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.

Supplemental Material

The online supplements are available at http://journals.sagepub.com/doi/suppl/10.1177/1948550616673876.

Notes

1. We have provided a simple spreadsheet at https://osf.io/3fcbr that automates the steps of this procedure.
2. The text of our e-mails is available in the Supplemental Online Material for this article.
3. Standard deviations (*SD*s) exhibit granularity in an analogous way to means, but the determination of (in)consistency for *SD*s is considerably more complicated. We hope to cover the topic of inconsistent *SD*s in a future article.

References

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *3*, 106–116. doi:10.3844/jssp.2007.106.116

Coombs, C. H. (1960). A theory of data. *Psychological Review*, *67*, 143–159. doi:10.1037/h0047773

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, *38*, 1212–1218. doi:10.1111/j.1365-2929.2004.02012.x

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS One*, *9*, e105825. doi:10.1371/journal.pone.0105825

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. doi:10.1126/science.aac4716

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286. doi:10.1037/h0070288

Wolins, L. (1962). Responsibility for raw data. *American Psychologist*, *17*, 657–658. doi:10.1037/h0038819

Author Biographies

Nicholas J. L. Brown is a PhD candidate at the University Medical Center, University of Groningen, The Netherlands.

James A. J. Heathers conducted the bulk of the work described in the attached document while a postdoctoral fellow at the Poznań University of Medical Sciences in Poland. He is currently a postdoctoral fellow at Northeastern University.

Handling Editor: Joseph Simmons