Response to Anderson and Ones

Lewis R. Goldberg
Oregon Research Institute

Kibeom Lee
University of Calgary

Michael C. Ashton
Brock University
Abstract

We respond to the rejoinder by Anderson and Ones (2008). Contrary to the claims of the rejoinder, the evidence described in our earlier comment (Goldberg, Lee, & Ashton, 2008) does indeed constitute proof of errors in the data sets in question. These errors do undermine the conclusions drawn in the original reports by Anderson and Ones, and we continue to recommend the retraction of those articles.
In their rejoinder to our comment (Goldberg, Lee, & Ashton, 2008), Anderson and Ones (2008) denied the existence of several errors in their earlier articles. We inferred these errors from several extraordinarily unlikely anomalies in their data set, which suggested a mismatching of participants’ HPI scale scores with their OPQ and BPI scale scores, as well as an incorrect scoring of OPQ scale scores for the first 202 of the 504 participants of their data file, and of BPI scale scores across the entire participant sample. We concluded that the results of the earlier articles were undermined by these errors, and we recommended the retraction of those articles. We explain below that the arguments denying these errors are invalid.

Mismatching of Participants’ HPI and OPQ Scale Scores

Recall that the data set of Anderson and Ones (2003) indicated that the cross-inventory correlations between the HPI and OPQ scales were all close to zero. We noted, however, that if participants’ HPI scores are shifted down one row, then the cross-inventory correlations between HPI and OPQ scores became large and theoretically meaningful. We pointed out that our proposed alignment produced a matrix of cross-inventory correlations whose standard deviation was .13, whereas the corresponding results for the other 503 alignments (including that reported by Anderson and Ones) ranged from .03 to .06. We concluded that our proposed alignment was the correct one, and that the alignment suggested by Anderson and Ones was incorrect.

In their response, Anderson and Ones (2008) argued that the data “can be shifted around in countless ways” and that evidence “that one particular shift in the data produces results more in line with expectations is in no way irrefutable evidence of a
clerical matching error.” Thus, Anderson and Ones reject our conclusion that our alternative alignment must be the correct one.

To evaluate the plausibility of the claim by Anderson and Ones, let us consider the null hypothesis that our proposed alignment is not the correct one. If our proposed alignment were not the correct one, then the matrix of cross-inventory correlations produced by this alignment should show no more variability than would the other 503 such matrices produced by the other 503 alignments. How plausible is this? In terms of the distribution of standard deviations of the other 503 correlation matrices, the standard deviation of our proposed matrix would yield a test statistic of $z = 18.6$, which roughly corresponds to $p < 10^{-76}$. We believe that a $p$-value of this infinitesimally small size is indeed “irrefutable evidence” that the participants’ HPI and OPQ scores had been improperly aligned in the original data set of Anderson and Ones.

Scoring of OPQ and BPI Scores

In our comment, we also drew attention to the extremely small standard deviations for the OPQ scales (among the last 302 of the 504 participants of the Anderson and Ones data set) and the BPI scales (among all 504 participants of that data set). We pointed out that these results indicated an error in the computation of these scale scores, because it would be extraordinarily unlikely that such a huge restriction in scale score variability would simultaneously be observed (a) among all participants for one of three omnibus personality inventories (i.e., the BPI), (b) among some participants for a second such inventory (i.e., the OPQ), and (c) among no participants for the third inventory (i.e., the HPI).
In response, Anderson and Ones argued that “upon more careful consideration, differential variability for HPI and OPQ in the two subsets of the data set is entirely unremarkable. In this particular sample, the OPQ and HPI are not correlated. Therefore, there is no inherent reason that variability on one inventory should mirror the variability on the other.”

The fallacy of this argument should be obvious: The near-zero correlations between HPI and OPQ scales cannot support the claim that the OPQ scales were scored correctly, because those correlations would themselves be influenced by any errors in scoring of those scales. This is a classic case of “begging the question”—that is, of using one’s proposition as one of the premises to that same proposition.

Also in response to our point regarding the implausibly severe restrictions in scale score variability, Anderson and Ones (2008) suggested that “Personality differences in who is selected into and remains in certain disciplines and universities can produce markedly direct and indirect range restriction in the data.” However, the processes of selection and attrition cannot plausibly explain two features of the restricted range of BPI scale scores. One feature is the finding of such a massive restriction in variability—with standard deviations averaging about 50% as large as in the normative sample—that is nearly uniform across all 11 scales of an omnibus personality inventory. The other feature is the finding of near-zero correlations between every pair of scales in that inventory, in spite of what are reported to be moderately high reliabilities for all scales. These findings are not “counterintuitive” or “lower than expected”; they are wildly implausible.
Implications for Conclusions

Anderson and Ones (2008) argued that their original conclusions about low convergent correlations across batteries remained intact, despite the evidence that we reported concerning errors in the scale scores. They reported results based on a re-analysis of data after re-alignment of participants’ HPI and OPQ scale scores, and after exclusion of the 202 participants whose OPQ scale scores we had identified as anomalous. They indicated that, on average, the cross-inventory convergent correlations remained very small.

Again, the flaw in this argument should be obvious. We argued that the results involving the BPI scale scores were invalid for all 504 participants, as a consequence of scoring errors across the entire sample. Therefore, any re-analysis of cross-inventory convergent correlations that is sincerely intended to address our concerns must exclude the BPI scale scores. Because the BPI scales are involved in the large majority of the cross-inventory correlations that were averaged by Anderson and Ones, the uniformly near-zero correlations produced by the BPI scale scores will obviously reduce the mean convergent correlations to very low levels. If BPI scale scores are excluded—as must obviously done if one genuinely means to address the question at hand—the mean cross-inventory convergent correlations (as based on the same pairs of HPI and OPQ scales identified by Anderson and Ones) become considerably larger. (The question of whether or not the scales identified by Anderson and Ones as conceptually “convergent” is a different issue, and beyond the scope of this article.)

Anderson and Ones (2008) also argued that the clerical errors described in our earlier comment did not have implications for their article in which sex differences and
ethnic group differences on the personality scales were reported. Contrary to their claim, these errors do undermine those results and the conclusions drawn from them. The errors in the scoring of the BPI and OPQ scales will obviously affect the sex differences in those scale scores. Also, when we correct the alignment of participants’ HPI scores with the other variables of the data file (including participant sex and ethnicity), this also changes the pattern of sex and ethnic group differences.

“Irrefutable Proof”

Finally, we note that Anderson and Ones (2008) use the phrase “irrefutable proof” in four places in their article, arguing that our evidence did not constitute irrefutable proof of a clerical error in the scoring of their data. We maintain that our evidence does indeed constitute irrefutable proof (recall, for example, the value of $p < 10^{-76}$). However, we also suggest that even if researchers were to believe that the evidence of errors in their data were not “irrefutable,” their sense of scientific responsibility should still compel them to retract their articles whenever they learn of evidence that casts such profound doubts on the integrity of those data.
References


Author Notes

Lewis R. Goldberg, Oregon Research Institute, Eugene, Oregon, USA; Kibeom Lee, Department of Psychology, University of Calgary, Calgary, Alberta, Canada; Michael C. Ashton, Department of Psychology, Brock University, St. Catharines, Ontario, Canada.

Correspondence should be addressed to Lewis R. Goldberg, Oregon Research Institute, 1715 Franklin Boulevard, Eugene, OR 97403-1983, USA (e-mail: lewg@ori.org).