Comment

Comment on Anderson and Ones (2008)

LEWIS R. GOLDBERG[1]*, KIBEOM LEE[2] and MICHAEL C. ASHTON[3]

[1]*Oregon Research Institute, Eugene, OR, USA*
[2]*Department of Psychology, University of Calgary, Calgary, AB, Canada*
[3]*Department of Psychology, Brock University, St Catharines, ON, Canada*

Abstract

The 'erratum and addendum' by Anderson and Ones (2008) does not state unambiguously that participants' Hogan Personality Inventory (HPI) scale scores were incorrectly matched with their scores on the other inventories' scales, nor does it mention the existence of other errors in the scoring of the Occupational Personality Questionnaire (OPQ) and Business Personality Indicator (BPI) scales. We demonstrate these errors and we recommend the retraction of the articles by Anderson and Ones (2003) and Ones and Anderson (2002). Copyright © 2007 John Wiley & Sons, Ltd.

Key word: personality scales and inventories

INTRODUCTION

The 'erratum and addendum' by Anderson and Ones (2008) acknowledges the 'possibility' of a major clerical error in the dataset that was the basis of their earlier articles (Anderson & Ones, 2003; Ones & Anderson, 2002). This acknowledgement is an important step forward,[1] but it remains problematic in three respects. First, the misalignment of the Hogan Personality Inventory (HPI) and Occupational Personality Questionnaire (OPQ) scales in

---

*Correspondence to: Lewis R. Goldberg, Oregon Research Institute, 1715 Franklin Boulevard, Eugene, OR 97403-1983, USA. E-mail: lewg@ori.org

[1]The findings described by Anderson and Ones (2003) were previously reported in a symposium at the 1998 meeting of the American Psychological Association in San Francisco. Lew Goldberg attended that symposium, and after the presentation of the Ones and Anderson paper he asked to speak. He stated that for decades he had been correlating scales from various personality inventories, and that he was convinced that there had to be an error in Ones and Anderson's data. Specifically, he explained that there had to be one or more errors in matching the data from inventory to inventory in order for there to be near-zero correlations between all of the scales from any two different inventories. Goldberg then made a public bet with Deniz Ones that her data had been mismatched. He suggested that if she would examine the data carefully, she would find the mismatching error; and if she did not find it, she could send him the data and he would find it for her. Ones responded that she had of course carefully checked the data and that all was well.

the Anderson and Ones dataset is not described forthrightly as an error. Second, no mention is made of some serious problems with the scoring of the OPQ scales. Third, no mention is made either of equally serious problems with the scoring of the scales of the Business Personality Indicator (BPI). We discuss these points below.

MISMATCHED HPI SCALE SCORES

In their 'erratum and addendum', Anderson and Ones (2008) explained that we had brought their attention to the 'potential' of a 'possible' misalignment, and described the results computed from re-aligned data as being based on a '*post-hoc*' analysis of a 'proposed' data shift. That is, Anderson and Ones did not plainly describe the mismatch problem as an error; instead they presented the new results merely as an alternative, supplementary re-analysis.

Contrary to Anderson and Ones (2008), the mismatch between the HPI and OPQ scale scores is much more than a mere 'possibility'. To understand this, consider the following analysis: There are 504 possible ways in which the 504 participants' OPQ scores can be matched with the HPI scores. (We assume that the order of the 504 participants' scores is consistent for the two inventories.) Of these 504 alignments, only the correct one should show a meaningful pattern of correlations between the two inventories; all other alignments should show matrices containing uniformly near-zero correlations.

Using the dataset provided by Anderson and Ones, we examined all 504 of the OPQ/HPI cross-inventory correlation matrices that result from these 504 different alignments. That is we repeated the initial step of pushing the HPI data down one row a total of 503 additional times, until at the last step we returned to the original alignment of the HPI and OPQ that Anderson and Ones (2003) had used.

To summarise the cross-inventory correlations within each of these 504 matrices, we computed for each matrix the standard deviation of those cross-inventory correlations, using mean substitution for the few subjects who had missing data on one or the other inventory. Note that, for the incorrect alignments, these standard deviations should all be very small, because each of those matrices will consist entirely of near-zero correlations. For the one correct alignment, the standard deviation should be much larger, because there will be a wider dispersion of values in the correlation matrix. When we computed the 504 alternative correlation matrices, and then computed the standard deviation of the cross-inventory correlations of each matrix, we obtained the results shown in Figure 1. Notice that the only large standard deviation (0.13) is that for the matrix that was based on what we indicated to be the correct alignment. For the other 503 matrices, the standard deviations were very small, ranging from 0.03 to 0.06; the value for the matrix based on the alignment that Anderson and Ones have used was 0.05. These results eliminate the possibility that the original alignment reported by Anderson and Ones could have been the correct one, and demonstrate that our suggested alignment is the correct one.

ANOMALOUS OPQ SCALE SCORES

Anderson and Ones (2008) did not mention another serious problem with the dataset of their earlier articles. This problem involves some anomalies involving participants' scores on the scales of the OPQ. One such anomaly is very simple: Although the Anderson and Ones dataset does reproduce the correlations reported by Anderson and Ones (2003), it
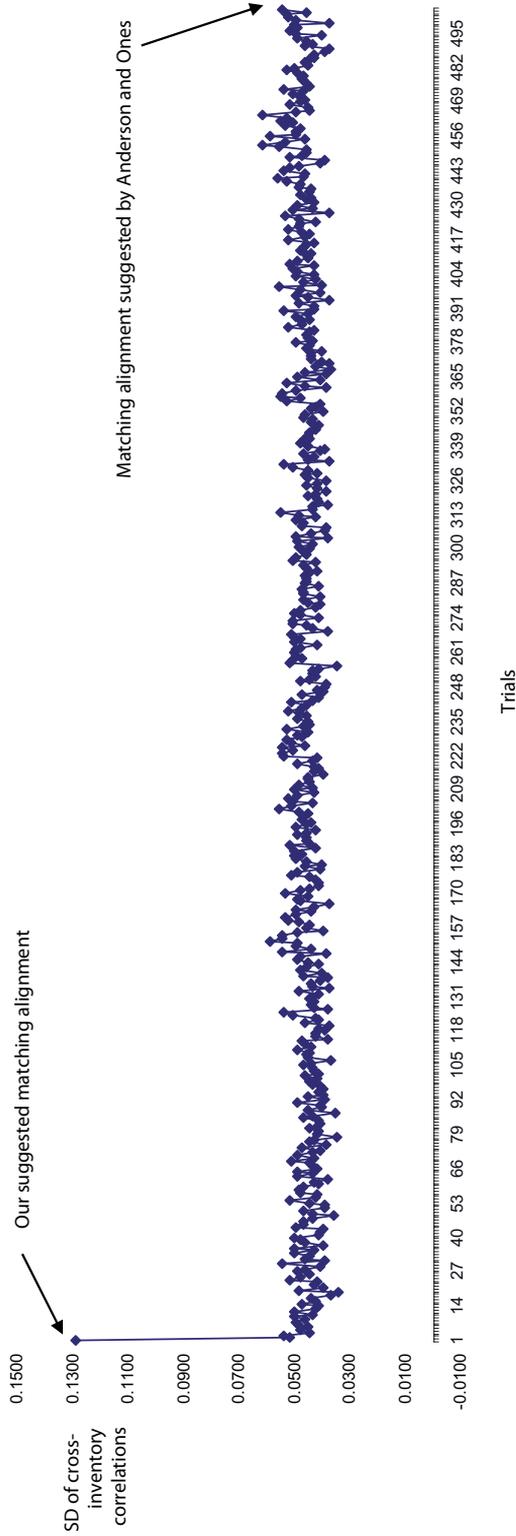
Figure 1. Standard deviations of cross-inventory correlations from the 504 alignments of OPQ and HPI scores.

does *not* reproduce the means and standard deviations of the OPQ scales as reported in Anderson and Ones (2003). For example, the OPQ scale score means as reported in Anderson and Ones ranged from 18.14 to 29.78, but those calculated from their dataset ranged from −3.65 to 11.40; similarly, the standard deviations reported by Anderson and Ones ranged from 3.96 to 8.10, whereas those calculated from their dataset ranged from 2.21 to 5.15. (There is no single linear transformation that allows conversion between the two sets of scale scores.)

There is another serious anomaly involving the OPQ scale scores of the Anderson and Ones dataset. A simple visual inspection of the dataset reveals a striking difference in the OPQ scale score distributions between (a) the *first* 202 participants of the dataset and (b) the *last* 302 participants. However, there was no such pattern of differences in the HPI scale score distributions. This observation became clear when we computed means and standard deviations of the OPQ and HPI scales. Among the first 202 participants, the average standard deviation of the 17 OPQ scales was only 57% as large as the average standard deviation among the last 302 participants; for the seven HPI scales, the average standard deviation was 98% as large. This is a most implausible finding, as it suggests a massive restriction of range across the diverse array of traits measured by the OPQ, but no such restriction of range across the similar range of traits measured by the HPI.

By far the most parsimonious explanation for these results is simply a miscalculation of the OPQ scale scores within one of the two subsamples. This hypothesis suggests that the two subsamples would show a dramatically different pattern of correlations involving the OPQ scales, and this is exactly what was observed. For the first 202 participants, the OPQ/HPI cross-inventory correlations were generally very small, and the few correlations of moderate size generally did not make any theoretical sense. For example, the highest absolute correlation observed was the −0.32 value between OPQ Sociable and HPI Ambition, which is opposite to the direction that would be expected on the basis of the scales' content. In contrast, when we computed the OPQ/HPI cross-inventory correlations for the remaining subset of 302 participants, many of the correlations were much stronger and these were consistent with the content of the scales involved. For example OPQ Sociable correlated above 0.50 (i.e. in the positive direction) with HPI Ambition, Sociability and Likeability. The results based on this subset of participants are more plausibly a reflection of the true patterns of relations between the HPI scales and most of the OPQ scales.

The hypothesis of a miscalculation of the OPQ scales for the first 202 sub-sample is also consistent with the pattern of within-inventory correlations of both the OPQ and the HPI within both subsamples. For the 202-participant subsample, the within-inventory correlations for the OPQ scales were generally very small and made little theoretical sense. For the 302-participant subsample, however, the correlations among the OPQ scales were much larger and were theoretically meaningful. Note that for the HPI, the within-inventory correlations were similar across both participant subsamples (and are theoretically meaningful in both subsamples). These results cannot be explained by any characteristics either of the participant subsamples (given that the relations among the HPI scales were plausible for the first 202 participants) or of the personality inventories (given that the relations among most of the OPQ scales were plausible for the last 302 participants).

We should add that the division between the first 202 and the last 302 participants is not arbitrary. Initially, we noticed a marked increase in OPQ scale scores beginning with the 203rd participant, who had higher scores than did any of the first 202 participants on three

OPQ scales (Influential, Empathic, and Conceptual) that are theoretically linked to three *different* Big Five factors; this participant's scores on each of these scales were exceeded by those of many subsequent participants. Later, we learned that the first 202 participants did in fact constitute a distinct sample: A table of BPI normative data provided from ASE (the publisher of the BPI) is based on a sample of 202 respondents whose age and sex distribution matches exactly that of the first 202 participants of the Anderson and Ones dataset.

ANOMALOUS BPI SCALE SCORES

Also not addressed by Anderson and Ones (2008) were some anomalous findings involving the BPI scale scores in their dataset. First, variation in scale scores was extremely small, with standard deviations for all 11 scales that were only 0.45 to 0.65 times as large as the normative sample values, even though there was no comparable restriction of range in any HPI scale scores. Also, correlations among the 11 BPI scales were all virtually zero in size—ranging from -0.08 to 0.15 (as opposed to values of -0.27 to 0.61 in the $N = 1016$ normative sample; Feltham & Woods, 2003)—even though Anderson and Ones (2003) reported moderately high levels of internal-consistency reliability for these scales (averaging 0.67). For anyone familiar with personality structure, it is extremely implausible that 11 reasonably reliable scales, each measuring a characteristic of personality, would all be almost perfectly mutually uncorrelated. Furthermore, the Anderson and Ones dataset also shows near-zero correlations of the 11 BPI scales with the seven HPI basic scales in the correctly matched data. That is, all 11 BPI scales are virtually uncorrelated with the entire set of seven HPI basic scales: If we regress each BPI scale on those seven HPI scales, the resulting adjusted multiple correlations range from 0.00 to 0.20 ($M = 0.10$).

To appreciate the implausibility of this result, consider the dataset of Goldberg's Eugene-Springfield Community Sample. This dataset contains scores for over 500 participants on the scales of many widely used omnibus personality inventories. When we regressed each of the 219 lower level scales of these inventories on the seven HPI basic scales, only two scales showed adjusted multiple correlations under the 0.20 limit observed for the 11 BPI scales in Anderson and Ones's (2003) dataset. These two scales—MPQ Variable Response Inconsistency (adjusted $R = 0.20$) and MPQ True Response Inconsistency (adjusted $R = 0.00$)—are not personality scales as such; instead, they are designed to detect random or inconsistent response patterns. *All* of the remaining 217 personality inventory scales had multiple correlations higher than the *highest* multiple correlation observed for the 11 BPI scales.

These results obviously raise a serious problem about the plausibility of the results reported by Anderson and Ones (2003) for the BPI scales. Moreover, the content of the BPI scales cannot solve this problem, given that these scales assess an array of constructs broadly similar to those of other inventories examined above. Neither can the participant sample provide a resolution, given that the HPI scale intercorrelations were similar to those observed elsewhere.

SUMMARY

Taken together, our re-analyses make three points clear. First, although the HPI scale scores were apparently calculated correctly, those scale scores were improperly aligned with the

scores from the other inventories. Second, the OPQ scale scores were miscalculated at least for the first 202 participants in the dataset. Third, the BPI scale scores were miscalculated for most if not all of the participants in the sample. Although these conclusions follow from the analyses that we have described above, the second and third conclusions would also be verified directly by simple inspection of the *item-level* dataset. Unfortunately, we have been unable to obtain that dataset from Anderson and Ones.

IMPLICATIONS FOR ONES AND ANDERSON'S (2002) STUDY

The dataset of Anderson and Ones (2003) had previously been used in another investigation (Ones & Anderson, 2002). In that study, Ones and Anderson examined gender and ethnic group differences on the three personality inventories, and the reported results will necessarily have been influenced by the anomalies identified in our re-analyses of these data. For example, it is clear that there should be grave concerns about all participants' BPI data and many participants' OPQ data. Equally, the mismatch problem indicates that the sex and ethnic group variables were matched incorrectly with some or all of the personality variables in the dataset.

CONCLUSION

The anomalous results of Anderson and Ones (2003; see also Ones & Anderson, 2002) appeared to contradict the well-established finding that personality characteristics can be assessed consistently by different instruments. However, our re-analysis demonstrates that the extraordinary findings reported by Anderson and Ones are attributable to clerical errors. We therefore conclude that the convergent validity of the inventories is not undermined by those results, and we recommend the immediate retraction of those articles.

REFERENCES

Anderson, N., & Ones, D. S. (2003). The construct validity of three entry level personality inventories used in the UK: Cautionary findings from a multiple-inventory investigation. *European Journal of Personality*, *17*, s39–s66.

Anderson, N., & Ones, D. S. (2008). The construct validity of three entry level personality inventories used in the UK: A cautionary case study: Erratum and addendum. *European Journal of Personality*. (this issue). DOI: 10.1002/per.651

Feltham, R., & Woods, J. (2003). *PeopleMapper: The workplace personality questionnaire*. London: ASE.

Ones, D. S., & Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology*, *75*, 255–276.