Industrial and Organizational Psychology, *8(3), pp 1–6 July 2015.*
 doi:10.1017/iop.2015.64

Commentary

The Determinacy and Predictive Power of Common Factors

Q1 James J. Lee and Nathan R. Kuncel
University of Minnesota—Twin Cities

Ree, Carretta, and Teachout's (2015) arguments for recognizing the importance of general factors are mostly on point, but they neglect two broad issues: (a) an important theoretical problem introduced by the presence of multiple factors (general, group, specific) and (b) the criterion validity of group factors in certain settings.

The theoretical problem is one known in the psychometric literature as *factor indeterminacy* (McDonald & Mulaik, 1979). Consider Figure 1, which represents an assignment of scores to a population of N individuals as a vector in N-dimensional space. Suppose that the vector X_1 represents the best estimates of general cognitive ability (g) in our population. Because no estimate is perfectly reliable, there is some correlation—smaller than 1—between X_1 and whatever the true population values of g may be. Suppose that the correlation happens to be .80. In Figure 1, the correlation between two vectors is represented by the cosine of the angle between them, and therefore a cone is traced in this N-dimensional space by all possible orderings of the examinees whose correlations with X_1 are equal to .80.

The question arises: Which of the vectors making up the cone corresponds to the "real" g? By standard psychometric theory, as the test is made more reliable by increasing the number of indicators (subtests, items), X_1 should approach that part of the original cone containing the true g. If the domain of indicators measuring g is not defined in advance, however, there is no reason to suppose that two independent research teams increasing the reliability of the same "seed" tests in this way will converge on the same part of the cone. Suppose that one team decides to add measurements of

James J. Lee and Nathan R. Kuncel, Department of Psychology, University of Minnesota—Twin Cities.

Correspondence concerning this article should be addressed to James J. Lee, Department of Psychology, University of Minnesota—Twin Cities, N409 Elliott Hall, 75 East River Parkway, Minneapolis, MN 55455. E-mail: leex2293@umn.edu

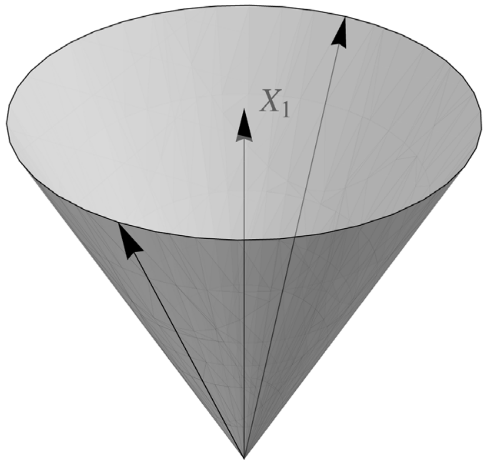

Figure 1. The cone representing the locus of all vectors having a fixed angle (correlation) with vector X_1. Two vectors on opposite sides of the cone are explicitly highlighted. Notice that two vectors can each be highly correlated with X_1 but not with each other.

reaction time to the original test battery; as a purely mathematical matter, this will increase calculated reliability because reaction time is correlated with IQ. Now suppose that another team, similarly unconstrained, decides to add anthropometric measurements to the original test battery; after all, height and similar variables are also correlated with IQ. As two measurements of the same quantity become more reliable, their correlation should increase, but in this case there is no logical reason to expect that seed tests + reaction time will become more highly correlated with seed tests + anthropometric measurements as the two sets are extended. In fact, the two vectors representing these extended sets may veer toward opposite sides of the cone, in which case a basic trigonometric identity shows that their correlation is a paltry .28. The calculated reliability of each extension may get closer to 1, but clearly the extensions are becoming increasingly accurate measures of different traits.

The extension may continue to share the original name but function differently as predictors. Not incidentally, this point—that two variables may show a high correlation with each other while having markedly (even sign
Q2 revered) correlations with a third—is much the same as the one made by McCornack (1956) in the context of whether two highly correlated variables can be assumed to be interchangeable for purposes of criterion validity. Mathematically this is not a safe assumption even when the correlation between two variables exceed .90. It is possible for two highly correlated sets of indicators to have external correlations that differ enough to be of practical significance.

This indeterminacy in factors that are empirically rather than conceptually grounded is an argument for why the domain of permissible indicators for the measurement of a psychological trait should be defined a priori (Lee, 2012; McDonald, 2003). Thus we only follow Ree et al. so far in their criticism of "naming [factors by] apparent content … frequently supported by consensus rather than by empirical evidence." The authors appear to envision cases where empirically observed correlations might trump content validity in determining whether a candidate indicator measures a certain factor. But this appears to invite precisely the drift of trait meaning that defines the problem of factor indeterminacy. In our opinion, the problem is not entirely fanciful. Some of the divisions between researchers over whose version of the Big Five is "really" measuring personality may be owed to excessive degrees of freedom in item selection. We are also somewhat perturbed by a trend in certain kinds of collaborative research for different groups to claim that their heterogeneous and often unreliable cognitive tasks are in fact measurements of the same common factor *g*.

There seems to exist a strong consensus regarding the a priori contours of the domains corresponding to certain group factors, such as the verbal and quantitative abilities measured by the SAT and GRE. The designers of these tests have produced thousands of items for operational use over the decades, and the high reliabilities of long but disjoint samples of items from these vast behavior domains indicate that any indeterminacy in these group factors is a very remote concern: No two face-valid tests of verbal ability, say, can show a correlation deviating all that far from unity as their item numbers go to infinity (Cook, Dorans, & Eignor, 1988). Moreover, two distinct tests of this kind do indeed show comparable magnitudes and patterns of criterion validity (e.g., Kuncel, Hezlett, & Ones, 2001, 2004). In this way group factors may possess a theoretical advantage over the general factor. In a hierarchical model with a general factor at the top level, it is the number of *group factors* rather than the number of items that must become large in order to beat down the indeterminacy of the general factor (Guttman, 1955).

Unfortunately, a strong consensus regarding the conceptually appropriate group factors to measure *g*, similar to one implicitly guiding applied psychometricians in their work on operational tests of group factors, does not yet exist. Psychologists of such stature as Lloyd Humphreys and John Carroll would certainly fail to see eye to eye here, if more group factors beyond a core of verbal, quantitative, and spatial factors were required. There is thus a worry that a greater focus on *g* rather than group factors is a greater focus on an object that is not mathematically unique.

Having laid out the cause for concern, we now give some reasons why *g* may be reasonably determinate after all. Upon Schmid-Leiman transformation of a hierarchical factor model, the loadings of the indicators on the *g* and the group factors obey a certain proportionality constraint. Removal of this constraint leads to the bifactor model, where indeterminacy may no longer be as much of a problem. For example, if a certain subset of indicators is characterized by strong loadings on *g* and negligible loadings on their group factor, then this subset can be given greater weight in the estimation of individual *g* scores. However, because the frequent excellent fit of the hierarchical model indicates that (for whatever reason) ability tests do usually come close to satisfying the proportionality constraint, it is desirable to seek another means of assuring the determinacy of *g*. In this light the study of Segall (2001) is quite interesting because one of its simulations of multidimensional computer adaptive testing of verbal and quantitative ability was able to measure the general factor in a hierarchical model with a reliability of .95, exceeding the figures obtained with more conventional methods. This result hints that certain features of this setting, possibly including the nonlinearity of the item response theory characteristic surfaces and the implicit individualized weighting of the item scores, can drive the reliability to one even when selecting items from a consensus domain. This intriguing suggestion is one that we plan to investigate in future work.

Ree et al. may feel that their Table 1 already alleviates any concerns over the indeterminacy of any general factor. The appropriate measure of determinacy (reliability) when there are multiple factors, however, is not the sum over indicators of the variance associated with the first principal component (or Cronbach's α, also mentioned by the authors). The appropriate measure is rather McDonald's ω, which is the squared correlation between *g* and the appropriate weighted sum of the indicators. (The reliability reported by Segall was, essentially, McDonald's ω.) An especially helpful tutorial regarding the calculation and interpretation of ω has been given by Brunner, Nagy, and Wilhelm (2012).

Even if we take the determinacy of *g* for granted, we must address the issue of criterion validity. Although *g* may indeed often be the "predominant source of predictiveness in cognitive tests," a substantial body of work has shown that certain group factors do predict important outcomes in a manner affording both practical utility and psychological insight. For instance, even within the top 1% of SAT scorers at age 13, those whose later achievements fall within a certain family of criteria (tenure-track faculty positions in the humanities, literary publications) show higher relative scores on the verbal subtest. Similarly, those in the top 1% whose later achievements fall within a contrasting family of criteria (tenure-track faculty in STEM, patents) show higher relative scores on the mathematics subtest (Park, Lubinski, & Benbow, 2007). This finding dovetails with those reported in a meta-analysis of GRE criterion validity: The verbal subtest and appropriate subject matter tests show higher correlations with graduate-school grade point average in the humanities, whereas the quantitative subtest and appropriate subject matter tests show higher correlations in mathematical/physical science (Kuncel et al., 2001). We also note that the verbal factor specifically appears to add more criterion validity to the prediction of performance on comprehensive exams (Kuncel et al., 2001, 2004). A particularly provocative finding is that, in a group of individuals with high and comparable levels of *g*, it is those with *more* spatial ability who find school to be *less* interesting and who are more likely to *discontinue* education for the sake of entering the workforce (Gohm, Humphreys, & Yao, 1998).

To conclude, because it is intellectually unsatisfying to place a strong emphasis on a general factor that is in fact ontologically ill defined (Meehl, 1993), more attention should be paid to whether a behavior domain is merely measuring several correlated things or can justifiably be said to be measuring *one* thing in a certain limit. At the very least, regardless of whether such a limit is attainable, the criterion validity of group factors demonstrates that a psychology of abilities is impoverished if the inherently plural nature of the abilities is too swiftly bypassed.

References

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality, 80*, 796–846.

Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT-Verbal test editions. *Journal of Educational Statistics, 13*, 19–43.

Gohm, C. L., Humphreys, L. G., & Yao, G. (1998). Underachievement among spatially gifted students. *American Educational Research Journal, 35*, 515–531.

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology, 8*, 65–81.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*, 162–181.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*, 148–161.

Lee, J. J. (2012). Correlation and causation in the study of personality (with discussion). *European Journal of Personality, 26*, 372–412.

McCornack, R. L. (1956). A criticism of studies comparing item-weighting methods. *Journal of Applied Psychology, 40*, 343–344.

McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, *49*, 212–230.

McDonald, R. P., & Mulaik, S. A. (1979). Determinacy of common factors: A nontechnical review. *Psychological Bulletin, 86*, 297–306.

Meehl, P. E. (1993). Four queries about factor reality. *History and Philosophy of Psychology Bulletin, 5*, 4–5.

Park, G., Lubinski, D., & Benbow, C. P. (2007). Contrasting intellectual patterns predict creativity in the arts and sciences: Tracking intellectually precocious youth over 25 years. *Psychological Science, 18*, 948–952.

Ree, M. J., Carretta, T. R., & Teachout, M. S. (2015). Pervasiveness of dominant general factors in organizational measurement. *Industrial and Organizational Psychology: Perspectives on Science and Practice*. **Q5**

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika, 66*, 79–97.