Making Personality Powerful

Kimberly A. Barchard, University of Nevada, Las Vegas

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Abstract

Personality tests are usually designed to ensure reliability. However, to compare groups, you should instead maximize power. Measure areas where groups differ, not areas where individuals differ. Also, avoid using tests that removed items to maximize reliability – they may have removed the items with the largest group differences.

Reliability and Power Serve Different Goals

Personality tests are usually designed using methods that guarantee adequate reliability (e.g., factor analysis, corrected item-total correlations). Reliability is the ratio of true score variance to total variance (see Figure 1). To get a high reliability coefficient we need high variability in true scores (so that people are different from each other) and low variability in error scores (so that the observed scores are close to the true score for each person). When the reliability coefficient is high, the test scores allow us to make relative comparisons between the test takers.

Reliability (Allen & Yen, 1979; Lord & Novick, 1968) is important if you want to distinguish between *individuals*. However, to distinguish between *groups* (men and women, treatment and control, stimulus 1 and 2), you need statistical power (Cohen, 1988). To obtain high power, you need large differences between the groups and little variability within each group. See Figure 2. Psychologists often try to distinguish between groups, and thus they need to know how to design and select tests that will result in high power.

To obtain high power, you should measure areas where groups differ. If you are unsure where they differ, use a truly comprehensive measure. Do not use tests that were developed using methods that maximize reliability (Stewart & Archbold, 1992, 1993). Such tests will omit areas that have little within-group variance in the development sample; however, little within-group variation does not guarantee little between-group variation. Instead, use tests that are truly comprehensive, or tests that where items were selected to distinguish between the groups of interest. See Table 1.

To obtain high power, you should also use tests with strong validity. Subject Matter Experts can suggest the areas where the groups are most likely to differ, and can assist in designing the most valid measure. Open-ended questions, interview, or behavioral observations may be more valid than written closed-ended questions. In some circumstances, blood samples, MRIs, or other data collection methods may provide the most valid data.

Will Increasing Reliability Increase Power?

You may have heard that increasing reliability will increase power. This is incorrect. Many psychometricians have discussed the relationship between reliability and power. After several decades of discussion, the consensus is clear: Other things being equal, more reliability leads to more power. This position is epitomized by two quotes. Humphreys and Drasgow (1989) conclude, "Increases in reliability—or its obverse, decreases in random measurement error in the marginal distribution of the dependent measures in an experiment—always increase power for fixed effect size" (p. 424). Similarly, Zimmerman, Williams, and Zumbo (1993) conclude, "there is no doubt of the importance of augmenting reliability in experimental settings whenever possible" (p. 16). This has led many people (and textbooks) to the over-simplification that greater reliability is associated with greater power. This conclusion is false.

Let us examine the argument in more detail. Increasing test length will increase reliability if the new items are similar to the existing items (Lord & Novick, 1968), and this increase in reliability will increase power (Cleary & Linn, 1969; Williams & Zimmerman, 1989). However, there are other ways to increase reliability, which might or might not increase power. One method of increasing reliability is to eliminate some items in order to create a more homogenous test. For example, in a test of extraversion, we can eliminate items that measure anything besides sociability. Eliminating items can increase reliability, but this may or may not increase power. Imagine we want to compare two groups. If the extra content is irrelevant to group differences, removing those items will increase power. But if those extra items capture important group differences, eliminating them will probably decrease power. A second method of increasing reliability is to change the construct that is being measured. If the new construct has a weaker relationship with the phenomenon of interest, the study might have less power even if the new measure has higher reliability.

Figure 1
Reliability allows us to distinguish between individuals

$$X_{ij} = T_i + E_{ij}$$

where X_{ij} = the observed score for person i on measurement j,
T_i = the true score for person i, and
E_{ij} = the error score for person i on measurement j.

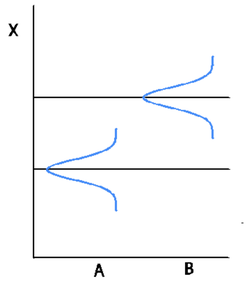
$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

where σ_X^2 = the variance of observed scores across all test takers,
σ_T^2 = the variance of true scores across all test takers, and
σ_E^2 = the variance of error scores across all test takers.

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

where $\rho_{XX'}$ = the reliability coefficient.

Figure 2
Power allows us to distinguish between groups



Psychometricians have concluded that increases in reliability result in increases in power, *other things being equal*. However, when researchers design experiments, they do not need to keep other things equal. There are many aspects of the test and the experimental design that researchers can change in order to increase power, and because of this, tests with lower reliability can result in higher power. Other researchers have noted the fact that lower reliability is sometimes associated with higher power: For example, Humphreys (1991, 1993) and Humphreys and Drasgow (1989) noted that a sample that has restriction of range may sometimes result in higher power, even though reliability is lower. In Table 1, I explain how we can design and select tests to maximize the power of our studies.

Conclusions

If we want to make relative comparisons between individuals, then it is beneficial to have a high Reliability Coefficient. But sometimes we have other goals. For example, if we want to obtain statistically significant results, we need power. This paper describes how to design and select tests that will increase our power.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Prospect Heights, IL: Waveland Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cleary, T. A., & Linn, R. L. (1969). *Effect of error of measurement on the power of statistical tests. Project No. 6-8574-24.* Princeton, NJ: Educational Testing Service.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika. 16,* 297-334. doi:10.1007/BF02310555

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Fabringar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272-299. doi:10.1037/1082-989X.4.3.272

Harmon, H. H. (1976). *Modern factor analysis* (3rd ed. revised). Chicago: University of Chicago Press.

Hotelling, H. (1931). The Generalization of Student's Ratio. *Annals of Mathematical Statistics, 2,* 360-378.

Humphreys, L. G. (1991). The relationship of power of statistical tests to range of talent: A correction and amplification. *Applied Psychological Measurement, 15*(3), doi:10.1177/014662169101500306

Humphreys, L. G. (1993). Further comments on reliability and power of significance tests. *Applied Psychological Measurement, 17* (1), 11-14. doi:10.1177/014662169301700102

Humphreys, L. G., & Drasgow, F. (1989). Paradoxes, contradictions, and illusions. *Applied Psychological Measurement, 13*(4), 429-431. doi:10.1177/014662168901300409

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: Sage.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* New Jersey: Lawrence Erlbaum Associates, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Maxell, S.E., Cole, D. A., Arvey, R. D., & Salas, E. (1991). A comparison of methods for increasing power in randomized between-subject designs. *Psychological Bulletin, 110,* 328-337.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8*(4), 283-298. doi:10.1016/S0001-2998(78)80014-2

Ones, D. S., & Viswesvara, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior, 17*(6), 609-626.

Stewart, B. J., & Archbold, P. G. (1992). Nursing intervention studies require outcome measures that are sensitive to change: Part one. *Research in Nursing & Health, 15,* 477-481.

Williams, R. H., & Zimmerman, D. W. (1989). Statistical power analysis and reliability of measurement. *Journal of General Psychology, 116*(4), 359-369. doi:10.1080/00221309.1989.9921123

Zimmerman, D. W., Williams, R. H., & Zumbo, B. D. (1993). Reliability of measurement and power of significance tests based on differences. *Applied Psychological Measurement, 17*(1), 1-9. doi:10.1177/014662169301700101

Table 1
Maximizing Reliability and Power

	To design a measure with high…	
	Reliability	Power
Rationale	There should be variability among ordinary people.	There should be differences between Group A and Group B, or between two sets of scores for one group (e.g., pretest vs. posttest).
Target constructs	Study areas were individuals differ. Constructs selected by researcher.	Study areas where groups differ. If we don't know where they differ, measure all areas (Ones & Viswesvara, 1996). Constructs selected by subject matter experts.
Number of Measurements	Development study includes a large item pool.	Development study might include only a few measures. Experimental psychologists often include just a single item.
Type of Measurements	Use written items with closed-ended response options. Items are similar but not identical. Measures are created by researchers.	Use the best possible measures of desired constructs to obtain high validity. Measures might be closed-ended, open-ended, behavioral observations, interviews, blood sample, MRI, etc. Sometimes researchers use the aggregate of multiple identical measures (e.g., reaction times). Measures are created by researchers in consultation with subject matter experts.
Standardize	Standardizing procedures will reduce variation and increase reliability.	Reducing unwanted (random) variations between raters will increase power. Therefore, standardize rating procedures. Be careful that you only standardize unwanted variation. Do not eliminate important content.
Aggregate	Increasing test length by adding items with similar content will increase internal consistency (Cleary & Linn, 1969; Williams & Zimmerman, 1989).	Increasing test length will increase power (Maxell, Cole, Arvey, & Salas, 1991). Aggregate multiple measures (from multiple raters, occasions, or items) to increase power (Lipsey, 1990)
Development sample	Large heterogeneous sample (often undergraduates or general population)	Theoretically based sampling. Often contrasts groups who are very different from each other. Within each group, people are relatively homogenous. Sample sizes are often modest.
Statistical analyses	Analyses are done at the level of individual items, to determine which items to retain. Analyses may include: (a) calculating the correlation between the item and the total scores from the remaining items (b) calculating the value of coefficient alpha (Cronbach, 1951) when the item has been removed from the test (c) using factor analysis (Fabringar, Wegener, MacCallum, & Strahan, 1999; Harmon, 1976) to determine which items are most closely associated with each other (d) examining item-characteristic curves (Embretson & Reise, 2000; Lord, 1980) to check that scores on the item increase as the level of the underlying trait increases	Significance testing to determine if there are differences between two or more groups. Analyses may be done at the level of individual items or at the scale level. Analyses may include (a) t-tests or ANOVAs to determine whether groups differ on individual items or scales (b) Hotelling's (1931) T-squared or discriminant function analysis to determine if some linear combination of the items or scales can be used to distinguish between groups (c) Receiver operating characteristics (ROC, Metz, 1978) to determine which cut-off scores to use to distinguish between groups, in order to maximize sensitivity and specificity
Item selection	Discard items upon which there is little variability in the derivation sample. Future studies will not be able to compare groups on these areas, because these areas are no longer included on the test.	To create a comprehensive test, report whether groups differ on each item (or scale), but keep all items. To create a test that has maximum power for distinguishing between two specific groups, discard items (or scales) for which the groups did not differ; however, future studies cannot use this test to compare other groups, because some areas are no longer on the test.
Final Scale	Either a single scale that is internally consistent or multiple subscales that are each internally consistent	A set of items (or scales) that significantly distinguishes between the groups of interest.
Published development paper	Reports the methods used to design the study, evidence of internal consistency (such as coefficient alpha or principle components analysis), and evidence of convergent validity (this test correlates with another test of the same construct). This paper has no practical implications beyond "Researchers should use this test in additional research."	Hypothesizes differences between target groups, based upon theory and practical experience from subject matter experts. Evaluates those differences. The paper has theoretical and practical implications.
Future research	Future studies usually use these same items, regardless of the purpose of their study.	Future researchers sometimes use the same items because the items are theoretically meaningful and distinguish between the target groups. Future researchers sometimes modify the items and scales to make them suitable to other groups.

Presented at the January 2012 Society for Personality and Social Psychology conference, San Diego, CA