

Gender differences in the reliability of the EPQ? A bootstrapping approach

Jeremy N. V. Miles*

Institute of Behavioural Sciences, Derby University

Mark Shevlin

University of Ulster at Magee College, Northern Ireland

Patrick C. McGhee

Psychology Department, Bolton Institute of Higher Education

Reliability indicates the degree of stability or homogeneity of a measurement, but also places an upper limit on the degree of association with other variables. Various methods are available to estimate the reliability of a measurement scale. However, an issue that has rarely been examined is that the reliability of a measurement, as estimated by coefficient alpha, may differ between groups. If a measurement has a different reliability for groups within a sample, spurious moderator effects may occur. The present study examines the reliability of the four subscales of a widely used psychological measurement instrument, the Eysenck Personality Questionnaire-Revised (EPQ-R), across gender. A bootstrapping methodology is employed which allows empirically derived standard errors to be calculated, and therefore tests of significance of difference to be computed. No significant differences were found in the reliability of the EPQ-R across sexes.

Introduction

Reliability and coefficient alpha

In psychological research, measures are usually taken indirectly, often by means of multiple item tests, such as intelligence or aptitude tests, or personality questionnaires. These indirect methods of measurement are likely to contain random measurement error. According to classical test theory, an observed score on a variable X is comprised of two uncorrelated components: a true score T , and measurement error e . The correlation between X and T is referred to as reliability, and is commonly estimated using coefficient alpha (Cronbach, 1951). Nunnally & Bernstein (1994) write of alpha: 'It is so pregnant with meaning that it should routinely be applied to all new tests' (Nunnally & Bernstein, 1994, p. 235).

* Requests for reprints should be addressed to Jeremy N. V. Miles, Institute of Behavioural Sciences, Derby University, Mickleover, Derby DE3 5GX, UK.

The lack of a perfect relationship between the true score and the measured score leads to attenuation of the estimates of relationships between those measures. Figure 1 shows two observed variables X_1 and X_2 , with corresponding true scores T_1 and T_2 . The correlation between the true scores is 0.70. However, the estimated correlation between the observed scores will be 0.448 (the product of the two reliabilities and the correlation between the true scores, i.e. $0.8 \times 0.7 \times 0.8 = 0.448$). This attenuation of the correlation occurs because the observed variables are imperfect indicators of their true scores. It is possible to estimate the size of the correlation between the true scores by using the attenuation-corrected correlation coefficient. However, this correction relies upon several untested assumptions and results derived from this procedure should be treated with extreme caution (Cohen & Cohen, 1983).

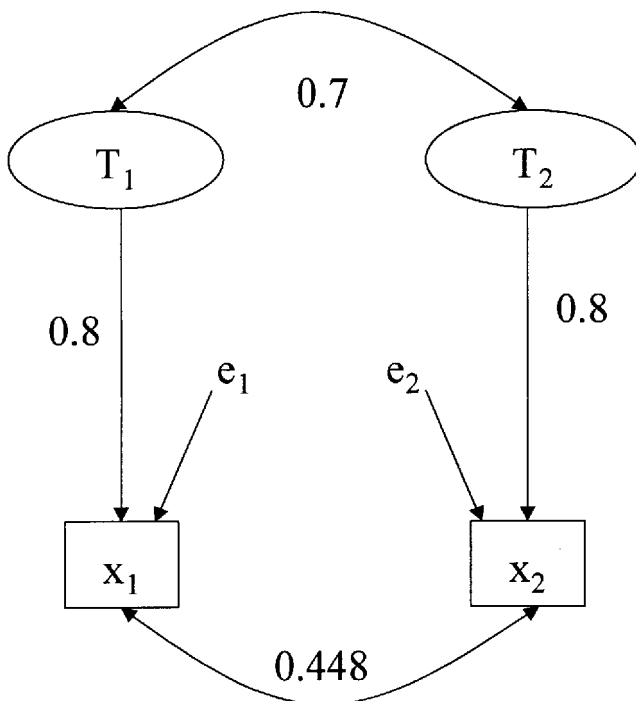


Figure 1. Path diagram showing the attenuation of correlation due to lack of reliability.

Psychologists have not paid a great deal of attention to this attenuation of effect size because hypotheses derived from psychological theory and tested in psychological research tend to be ordinal in nature (Frick, 1997); that is, they propose that an effect will exist and will be greater than zero rather than that an effect with a specific size of f will exist (where f is a value in terms of an effect size). Increasingly, psychologists are being encouraged to think in terms of effect sizes, due to considerations of value for money, clinical gain and ethical concerns (APA, 1996).

It is possible that the reliability of a psychological measurement instrument will vary across different groups of individuals. This difference in reliability may occur if the meanings of any questions could be interpreted differently according to cultural context or expectations. If this occurs it is possible that the reliability, and therefore validity, may differ across groups, which would lead to two problems. First, researchers or applied psychologists may find themselves using a test that is not appropriate for a proportion of the people they are testing. Lower reliability may lead to misclassification, a problem that may raise questions, for example, in relation to equality in recruitment, and for the long-term usefulness of psychometric tests in the selection process.

Second, an interaction effect is being tested. An interaction effect occurs when the level of one independent variable moderates the effect of a second independent variable. Experimental psychologists commonly analyse their data to examine the existence of interaction effects; psychologists studying individual differences do this less commonly, although recently there has been an increasing amount of interest in the investigation of interaction effects in continuous independent variables (Aguinis & Stone-Romero, 1997; Aiken & West, 1991; Baron & Kenny, 1986; Jaccard, Turrisi & Wan, 1990; Jaccard & Wan, 1996*a, b*; McLelland & Judd, 1993; Maxwell & Delaney, 1993).

When moderator analyses are carried out, the effects of attenuation due to measurement error are increased. Consider the example of an analysis of gender differences in the relationship between a personality measure, such as self-esteem, and a dependent variable, such as blood pressure. If the self-esteem scale has a lower reliability for one gender, it is likely that statistical analysis will reveal that a moderated relationship exists: that is, the analysis may show that self-esteem and blood pressure correlate to a greater degree in one of the two groups than the other, when the actual relationship between the two variables is equal for both groups. The moderator effect that would be detected is spurious. That is, the estimated correlation between self-esteem and blood pressure may appear to differ across the groups, but this appearance is only due to the differences in the reliability of the self-esteem scale.

To carry out a statistical procedure that examines the difference between the coefficient alpha of a test for two groups, the distributional properties of alpha must be known. These can be used to calculate standard errors and confidence intervals. The distributional properties of alpha have been presented by Feldt (1965) and Hakstian & Whalen (1976), although the procedures are known to be accurate only under restrictive assumptions. Barchard & Hakstian (1997) have investigated the accuracy of both these procedures and found that, for the confidence intervals to be accurate, the data must satisfy the condition of sphericity. If the sphericity assumption is violated, poor type I error control occurs using either method.

An alternative approach to significance testing which does not require a knowledge of the distributional properties of a statistic is the use of the bootstrap (Efron, 1979).

The bootstrap

The aim of statistical inference is to estimate a value of a population (θ), with a sample point estimate ($\hat{\theta}$). For example, the mean of a sample is considered the best estimator of a measure of the mean for that population, given that certain assumptions hold. As the sampling distribution of the mean is known to be normal it is possible, from Central Limit Theorem, to calculate standard errors and therefore confidence intervals for the mean. Because the confidence intervals are known, a range of statistics may be calculated and this is the basis of parametric statistical significance testing. Efron & Tibshirani (1993) refer to estimates such as the mean, median or standard deviation as ‘plug-in’ estimates of a sample parameter because they are calculated by ‘plugging in’ a formula to a dataset.

Where the distributional properties of an estimator are not known, it is not possible to make these sorts of comparisons of estimates using plug-in estimates. The distributional properties of an estimate may not be known because they have not been calculated or they could not be calculated, or the calculations rely upon assumptions that are violated. It is possible, however, to estimate empirically the sampling distributions of any test statistic using the bootstrap technique (Efron, 1979).

Bootstrapping is a conceptually simple but very powerful tool for use in inferential statistics. It is a computationally intensive procedure, related to Monte Carlo simulation and Jack-Knife Analyses, and it has become more feasible in recent years due to the advances in the speed at which desktop computers can carry out analyses.

This paper can necessarily give only a brief introduction to the reasoning behind the bootstrap procedure (the interested reader is referred to Mooney & Duval, 1993; Efron & Tibshirani, 1993; Davison & Hinkley, 1997; or for a briefer and more mathematical treatment, Efron, 1982). Bootstrap methods are so called because they seem to use the data to generate more data, in a way ‘analogous to a trick used by the fictional Baron Munchausen, who when he found himself at the bottom of a lake, got out by pulling himself out by his bootstraps,’ (Davison & Hinkley, 1997, p. 14).

Bootstrapping works because it does not rely upon knowing, or assuming, the form of the underlying probability distribution function (F) of the variable. Instead, it relies on the empirical distribution function of the variable (\hat{F}), which is the non-parametric maximum likelihood estimate of F .

Given a sample $X = (x_1, x_2, x_3, x_4 \dots x_N)$ of size N , bootstrapping involves creating a new sample (X^*) by sampling, with replacement, from the original sample. By sampling with replacement, we refer to a process in which one randomly selected case is sampled and added to the bootstrap sample, but is then replaced in the original sample, from which it may be sampled again. The process is shown diagrammatically in Figure 2, in which the value 1 was selected three times, the value 4 was selected twice, and the value 5 was selected once. The values 2, 3 and 6 were not selected in this replication.

The statistic of interest is then calculated from the X^* , referred to as $\hat{\theta}^*$. The process is repeated many times. The results of the calculation of each $\hat{\theta}^*$ are used in further calculation. The standard error of $\hat{\theta}$ is then estimated by calculating the standard deviation of the bootstrapped estimates $\hat{\theta}^*$, or alternatively, the confidence intervals can be calculated by using the percentiles of the distribution of $\hat{\theta}^*$.

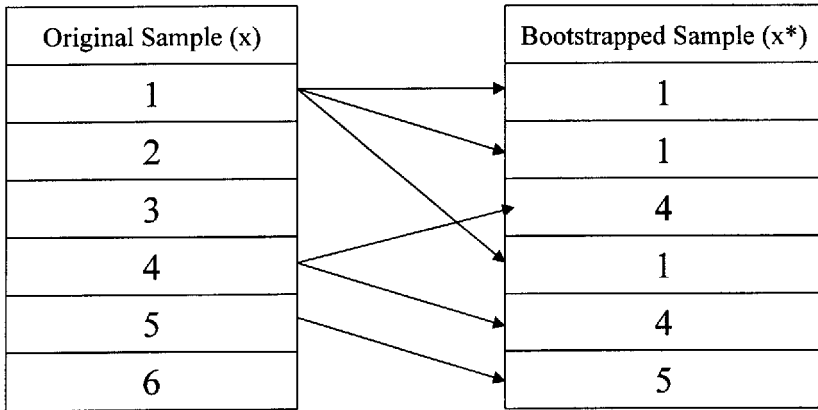


Figure 2. The process of sampling with replacement.

The following analyses demonstrate the use of bootstrapping techniques for estimating the reliability, and testing for differences, of the Eysenck Personality Questionnaire-Revised (EPQ-R; Eysenck, Eysenck, & Barrett, 1985) for males and females. The results are compared to traditional ‘plug-in’ estimates for the same data, and those reported by Eysenck *et al.* (1985). The EPQ-R was selected as a widely used scale in theoretical and applied settings, and it is generally acknowledged as having satisfactory psychometric properties.

Method

Sample

The sample was a group of people who were suffering from various skin diseases, such as psoriasis, vitiligo and eczema, and who were taking part in a longitudinal research project. Participants were contacted by postal questionnaire. The sample comprised of 118 males and 256 females. The mean age for females was 42.2 years (SD = 18.1), and for males 44.6 years (SD = 21.9). The age difference was not significant ($t = 1.16$, d.f. = 364; $p = .247$).

Measures

The short form of the EPQ-R was administered by means of a postal questionnaire (Eysenck, Eysenck, & Barrett, 1985). This questionnaire consists of four scales, E (extraversion), N (neuroticism), P (psychoticism) and L (social desirability). Each scale is measured using 12 items and has a dichotomous (yes/no) response format.

Results

Plug-in results

Table 1 shows the plug-in estimates of mean, standard deviation and alpha. The estimates of the current study and the norm data published by Eysenck *et al.* (1985)

are shown. Eysenck *et al.* (1985) presented results from two samples for psychoticism (Sample a and Sample b).

Table 1. Plug in estimates of mean, SD and alpha, from current study and norms sample presented by Eysenck *et al.* (1985)

Sample	Extroversion		Neuroticism		Lie		Psychoticism		
	Current study	Eysenck <i>et al.</i>	Current study	Eysenck <i>et al.</i>	Current study	Eysenck <i>et al.</i>	Current study	Eysenck <i>et al.</i> (Sample a)	Eysenck <i>et al.</i> (Sample b)
Mean									
Total	6.67		6.95		5.25		1.88		
Male	6.56	6.36	6.42	4.95	4.84	3.86	2.31	2.73	3.33
Female	6.73	6.46	7.2	5.90	5.45	3.69	1.69	2.02	2.61
s.d.									
Total	3.35		3.33		2.70		1.74		
Male	3.46	3.8	3.44	3.44	2.60	2.71	1.75	2.19	2.18
Female	3.54	3.27	3.25	3.14	2.74	2.55	1.70	1.69	1.97
Alpha									
Total	0.86		0.83		0.70		0.61		
Male	0.84	0.88	0.83	0.84	0.71	0.77	0.55	0.68	0.62
Female	0.86	0.84	0.82	0.80	0.70	0.73	0.63	0.51	0.61

The plug-in estimates of the means, standard deviations and alphas for this study are similar to those reported by Eysenck *et al.* (1985).

Sphericity tests

Before confidence intervals can be calculated, using either the Feldt (1965) or Hakstian–Whalen (Hakstian & Whalen, 1976) method, the assumption of sphericity must be checked by using a sphericity test, which results in a χ^2 distributed test statistic, which in turn provides the probability that the data are drawn from a spherical covariance matrix (Field, 1998, provides a detailed description of sphericity). Table 2 shows the results of the test for each of the four scales on each group. As each of the tests is significant, the hypothesis of sphericity is rejected, and the data therefore violate the assumptions required for the confidence interval tests. Without this assumption being satisfied, the distributional properties of coefficient alpha are not known.

Bootstrap results

A bootstrap test does not make any assumptions about the distribution of the test statistic, and is therefore appropriate when assumptions about the distributions are violated. SPSS for Windows (SPSS Inc., 1996) was used to carry out bootstrap replication (see Appendix). Males and females were separated and 1000 bootstrap samples generated for each group. For each of the four subscales, alpha was calculated for each bootstrap sample.

The standard deviation of the distribution of the bootstrapped estimates of coefficient alpha provides an estimate of the standard error of the alpha. The standard

Table 2. Results of sphericity tests for each scale, for female and male groups

Scale	Group	χ^2 all with 65 d.f. (<i>p</i>)
Extroversion	Females	342.0 (< .005)
	Males	179.1 (< .005)
Neuroticism	Females	301.1 (< .005)
	Males	129.2 (< .005)
Lie	Females	694.9 (.005)
	Males	234.1 (< .005)
Psychoticism	Females	167.5 (< .005)
	Males	93.4 (.012)

Table 3. Tests of significance of difference of alpha between males and females, using bootstrapping

Scale	Group	Mean	Standard error	95 % confidence interval (lower)	95 % confidence interval (upper)	<i>t</i> -value (two-tailed significance)
Extroversion	Male	0.849	0.019	0.811	0.888	0.48 (0.63)
	Female	0.862	0.019	0.838	0.885	
Neuroticism	Male	0.833	0.024	0.789	0.877	0.31 (0.75)
	Female	0.824	0.016	0.793	0.855	
Lie	Male	0.705	0.041	0.625	0.785	0.07 (0.94)
	Female	0.701	0.037	0.627	0.774	
Psychoticism	Male	0.543	0.058	0.430	0.656	1.06 (0.29)
	Female	0.622	0.047	0.530	0.714	

error can be used to calculate the 95 % confidence intervals and to test for significance.

To calculate the significance of any difference in the levels of alpha, a parametric bootstrap difference test (Mooney & Duval, 1993) was used. The standard deviation of the bootstrapped distribution is used as the estimate of the standard error in the calculation of the *t*-value. The bootstrapped mean, standard errors, 95 % confidence intervals and *t*-values (with associated probability) are shown in Table 3. The difference between the alpha values for males and females is not significant for each of the four scales.

Discussion

The results of the bootstrap comparison of the reliability of the four scales show no significant differences in the reliability of the EPQ-R between males and females in this sample. The use of plug-in estimates does not allow such conclusions to be drawn, as statistical tests of differences are not reliable when the assumptions of the tests are violated. Testing differences in reliability is important because, as previously

noted, the failure of a psychological measure to be equally reliable across groups would have implications for the research findings based on that measure. For self-report measures, significant differences in reliability may indicate that the language used in the items, or the values and aspirations assumed in the items do not validly apply to members of both groups.

Consistent with previous findings, the reliability of the psychoticism subscale is lower than the reliability of the other scales (Eysenck & Eysenck, 1985). In addition, the bootstrap analysis shows that the standard deviation of the sampling distribution of the reliability of the psychoticism scale is larger than that of the other scales. Therefore, varying estimates for the reliability of the psychoticism subscale are to be expected.

Although this paper has focused on testing differences in reliability across gender, the bootstrapping technique has a range of applications for testing differences in reliability estimates. Furthermore, bootstrapping can be used to test hypotheses or calculate confidence intervals for samples drawn from populations that have non-normal or unknown distributions. Differences in a test statistic can be examined across qualitatively different groups, such as race, nationality or age groups.

Despite the obvious benefits of the bootstrapping procedure, its application has been limited in psychological research. This can be attributed in part to previous hardware and software limitations, and in part to a lack of awareness. Although bootstrapping is computationally intensive, it is now viable using commercially available statistical packages on desktop computers.

Acknowledgements

This research was partly funded by the Psoriasis Association of Great Britain. We thank Henry Potts and two anonymous reviewers for comments on an earlier draft of this paper, and the associate journal editor who has dealt with this paper, Steven Sutton.

References

- Aguinis, H. & Stone-Romero, E. F. (1997). Methodological artefacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology*, **82**, 197–206.
- Aiken, L. & West, S. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- APA (1996). *Task force on statistical inference initial report (DRAFT)*. Washington: Board of Scientific Affairs, American Psychological Association.
- Barchard, K. A. & Hakstian, A. R. (1997). The robustness of confidence intervals for coefficient alpha under violations of the assumption of essential parallelism. *Multivariate Behavioral Research*, **32**, 169–191.
- Baron, R. M. & Kenny, D. A. (1996). The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51**, 1173–1182.
- Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioural sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297–334.
- Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap methods & their applications*. Cambridge: Cambridge University Press.
- Efron, B. (1979). Bootstrap methods: another look at the jack-knife. *Annals of Statistics*, **7**, 1–26.

- Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
- Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.
- Eysenck, H. J. & Eysenck, M. W. (1985). *Personality and individual differences: A natural sciences approach*. London: Plenum.
- Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, **6**, 21–29.
- Feldt, L. S. (1965). The approximate sampling distribution of Kuder–Richardson reliability coefficient twenty. *Psychometrika*, **30**, 357–370.
- Field, A. (1998). A bluffer's guide to ... sphericity. *The British Psychological Society Mathematical, Statistical and Computing Section Newsletter*, **6**, 13–22.
- Frick, R. W. (1997). The appropriate use of null hypothesis significance testing. *Psychological Methods*, **1**, 379–390.
- Hakstian, A. R. & Whalen, T. E. (1976). A k -sample significance test for independent alpha coefficients. *Psychometrika*, **51**, 393–413.
- Jaccard, J., Turrissi, R. & Wan, C. K. (1990). *Interaction effects in multiple regression*. Newbury Park, CA: Sage.
- Jaccard, J. & Wan, C. K. (1996a). *LISREL approaches to interaction effects in multiple regression*. Newbury Park, CA: Sage.
- Jaccard, J. & Wan, C. K. (1996b). Measurement errors in the analysis of interaction effects between continuous predictors using multiple regression: multiple indicator and structural equation approaches. *Psychological Bulletin*, **117**, 348–357.
- McLelland, G. H. & Judd, C. M. (1993). Statistical difficulties in detecting interactions and moderator effects. *Psychological Bulletin*, **14**, 376–383.
- Maxwell, S. E. & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, **113**, 181–190.
- Mooney, C. Z. & Duval, R. D. (1993). *Bootstrapping: A non-parametric approach to statistical inference*. Newbury Park, CA: Sage.
- Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric theory*, 3rd ed. New York: McGraw Hill, Inc.
- SPSS Inc. (1996). *SPSS for Windows*, Version 7.0 [Computer software]. Chicago: SPSS Inc.

Received 14 July 1997; revised version received 14 May 1998

Appendix

The SPSS macro to create a data-set comprising of bootstrapped samples, set up for analysis, provided by David Marso, of SPSS Inc.

```

DEFINE !BOOTSMP (BKPFIL !CHAREND ("/")
    / NSAMP !CHAREND ("/")
    / NCASE !CMDEND).
COMPUTE @ID = $CASENUM.
SAVE OUTFILE !QUOTE (!BKPFIL).
INPUT PROGRAM.
LOOP @SAMPLE = 1 to !NSAMP.
LOOP #V = 1 to !NCASE.
COMPUTE @ID = TRUNC (UNIFORM (!NCASE))+ 1.
END CASE.
LEAVE @SAMPLE.
END LOOP.
END LOOP.
END FILE.
END INPUT PROGRAM.
SORT CASES BY @ID.
MATCH FILES / FILE* / TABLE !QUOTE (!BKPFIL) / BY @ID.

```

!SORT CASES BY @SAMPLE.

!SPLIT FILE BY @SAMPLE.

!ENDDEFINE.

* Example:*

Data list free / a b c.

Begin data

1 2 3 2 4 1 3 4 3 1 2 4 2 1 4

1 2 4 3 1 2 4 3 1 4 3 1 2 3 4

End data.

!BOOTSMF BKPFILE C:\TEMP\BOOTSAMP

/ NSAMP 100

/ NCASE 10.

Substituting C:\TEMP\BOOTSAMP with a valid filename of your choice NSAMP with the number of desired bootstrap samples and NCASE with the number of cases in the data file. New variables called @SAMPLE and @ID will appear in the data file. @SAMPLE is the bootstrap sample number, @ID is the original position of the case in the file.

Disclaimer:

THIS MACRO IS LICENSED 'AS IS' WITHOUT WARRANTY AS TO ITS PERFORMANCE. THERE ARE NO WARRANTIES EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, AND ALL SUCH WARRANTIES ARE EXPRESSLY DISCLAIMED. IN NO EVENT SHALL SPSS INC. BE RESPONSIBLE FOR ANY INDIRECT OR CONSEQUENTIAL DAMAGES OR LOST PROFITS, EVEN IF SPSS INC. HAD BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE. SOME STATES DO NOT ALLOW THE EXCLUSION OR LIMITATION OF IMPLIED WARRANTIES OR LIABILITY FOR INCIDENTAL OR CONSEQUENTIAL DAMAGES, SO THE ABOVE LIMITATION MAY NOT APPLY TO YOU.

SPSS Inc. retains ownership rights to this Macro.