Pergamon

PII: S0191-8869(98)00011-7

Person. individ. Diff. Vol. 24, No. 6, pp. 815–820, 1998

Printed in Great Britain
0191-8869/98 $19.00+0.00

PERSONALITY–PERFORMANCE CORRELATIONS AT WORK: INDIVIDUAL AND AGGREGATE LEVELS OF ANALYSES

Chris J. Jackson[1] and Philip J. Corr[2]*

[1]Department of Psychology, University of Surrey, Guildford GU2 5XH, England; [2]Department of Psychology, Goldsmiths College, University of London, New Cross, London SE14 6NW, England

(Received 24 September 1997)

Summary—In the occupational community, there is a widespread faith in the utility of personality assessment for selection, development, etc. This faith has been immune to arguments, supported by empirical evidence, regarding the poor correlation between personality and performance in the workplace (these correlations rarely exceed the 0.2–0.3 level). The difference between perception of utility and the actual empirical reality is large. We investigated one possible source of this perceived-actual discrepancy. In two separate samples, we compared the magnitude of validity coefficients from individual and aggregate (i.e. organizational) levels. Our results indicated that strong actual personality–performance correlations exist at the aggregate level of analysis, but not at the individual level of analysis. We suggest that this aggregate-individual correlation discrepancy may, in part at least, account for the perceived-actual discrepancy noted above. We conclude that the continued faith in personality testing in the workplace may be a consequence of test users' sensitivity to actual aggregate level personality–performance correlations. However, we warn of the danger of drawing inferences from aggregate level correlations when making decisions about individuals, and point out the statistical artefacts that may account for some of the magnitude increase in aggregate level correlations. Several foci for further research are indicated. © 1998 Elsevier Science Ltd. All rights reserved.

INTRODUCTION

Confidence in the utility of personality assessment for occupational purposes is high amongst practitioners (for a sample of opinions, see Fletcher, 1991); and many companies use personality tests for a wide range of purposes (e.g. selection and development). For example, Shackleton and Newell (1991) found that 37% of U.K. companies used personality for management selection in 1989, a large increase on the 12% previously reported by Robertson and Makin (1986).

The perceived value of personality assessment in the work place stands in stark contrast to the claim of Blinkhorn and Johnson (1990) that "there is precious little evidence that even the best personality tests predict performance" (p. 672). Although evidence is beginning to be published showing the replicability of personality effects on work performance (e.g. Corr and Gray, 1995, 1996), and more sophisticated approaches to formulating and testing personality–performance relations are being proposed (e.g. Robertson and Kinder, 1993), it remains a fact that validity coefficients are relatively low (<0.30) and resistant to significant increment in magnitude (Robertson and Kinder, 1993). The conclusion seems unavoidable that, in comparison with validities for cognitive ability, assessment centres and work samples, personality measures fare rather badly in their power to predict work performance (Schmitt *et al.*, 1984). There are a large number of possible reasons for these low validities (e.g. unreliability in personality and performance measures; criterion insufficiency and contamination; as well as lack of consideration of personality × situation effects).

In this article, we ask the question: in the light of the above findings, why do human resource practitioners continue to place faith in personality assessment? We propose that one reason for this faith is that personality and performance *are* highly correlated from the perspective of the test user;

*To whom all correspondence should be addressed. E-mail: psa01pjc@gold.ac.uk

and that these perceived correlations result from a sensitivity to organizational, or aggregate level, information. In trying to make sense of employees' behaviour at work, we suggest that test users apply implicit personality processes, that look for consistent patterns of behaviour that relate to performance. In their search for work-related behavioural consistency, test users may compute (subjective) correlations at the *aggregate* data level rather than at the *individual* data level.

Aggregation of data tends to increase the magnitude of correlations between variables (for a discussion of the theoretical, methodological and statistical bases of aggregate data, see Ostroff, 1993). Thus, validity coefficients at aggregate and individual levels of analysis are likely to differ markedly; this fact in turn, may account for the marked differences in the perceived value of occupational personality testing. Indeed, test users often talk of personality–performance relations in terms of aggregate data; for example, "the best sales staff are extraverts", "the best accountants are introverts". Such statements are also echoed by leading personality researchers: Costa and McCrae (1992) remarked that, "while it is possible to enjoy sales and to be a good salesperson without being extraverted, it is probably unusual" (p. 36).

One way of aggregating personality data is to shift the focus from individual scores to average scores of individuals (at the coarsest level of analysis, typological classification, e.g., introverts and extraverts). The resulting data reflects an average level of performance for each aggregate personality group (aggregate data correlations are based upon a locus of averages, as with individual data, but with no error variance distributed around these averages). By elimination of error variance alone, aggregate correlations will tend to increase weak and non-significant linear relationships (but it is incorrect to assume that increased magnitudes are exclusively a function of aggregation).

Aggregate data correlations cannot be interpreted as representing correlations between *individuals* and performance — they represent correlations between average *groups* of individuals and performance. The importance of this caveat is illustrated by the following example. If a strong positive correlation is found between aggregate extraversion groups and sales performance, then it is valid to conclude that better salespeople tend, *on average*, to be higher in extraversion; but, it would be invalid to conclude that all extraverted *individuals* make better salespeople. Thus, knowing an individual's extraversion score does not necessarily entail any logical conclusion concerning their likely sales performance (such a conclusion would depend on the degree of variance at each aggregate data point). By eliminating individual differences, correlations based on aggregate data may hold no implications for individuals' scores.

Why should it be assumed that test users compute aggregate correlations when relating personality to performance? Following the work of Heider (1958), it could be assumed that test users adopt "naive" causal attribution models to explain observed correlations between personality and performance. In fact, until comparatively recently, it was common among scientists, such as Galton and Quetelet, to consider the taking of the average (or "typical" score) as a fundamental operation of mind for identifying the generic elements (representativeness) of diverse objects; and it was not until the work of Karl Pearson that current conceptions of correlation and regression took hold in the scientific community (for a fascinating review of the philosophical background to statistical techniques that underlie the above discussion, see Mulaik, 1987). Therefore, the idea that test users may perceive personality–performance relations by taking average scores is far from being a fanciful notion.

An example of the use of the aggregate data level is seen in the area of job analysis, which usually has the aim of identifying average (i.e. generic) components of the job, which are then expressed in terms of job requirements for the *average* individual. For example, repertory grid analysis often requires raters to compare a number of above *average* performers with their below *average* peers.

The aims of this study are (1) to demonstrate the increase in coefficient magnitude as one moves from the individual level to the aggregate level of analysis; and (2) to illustrate how the pattern of aggregate level correlations can be consistent with test users' faith in personality–performance relations (despite the lack of support at the individual level of analysis).

We conducted two studies. The first study examined the effectiveness of the Eysenck Personality Profiler (EPP: Eysenck and Wilson, 1991) in predicting sales success; the second study examined the power of the NEO PI-R (Costa and McCrae, 1992) personality profiler in predicting the performance of rugby referees. We used two samples, with two different personality instruments, in order to demonstrate the generalizability of our conclusions.

STUDY 1: SALES PERFORMANCE

Method

Personality questionnaire: Eysenck Personality Profiler (EPP)

The EPP (Eysenck & Wilson, 1991) is a 440-item normative test, developed by factor analysis and possessing adequate psychometric properties (Eysenck *et al.*, 1992; Costa and McCrae, 1995). The EPP is composed of four major dimensions of personality, Extraversion (E), Neuroticism (N), Psychoticism (P) and Dissimulation or Lie (L), with E, N and P containing 7 lower order traits, giving a total of 21 traits (plus the L scale). In this study we focused on the 4 major dimensions.

Performance criterion

Criterion ratings were collected at the same time as the administration of the EPP, but raters did not have knowledge of the personality scores. Overall ratings of performance for each member of staff were made by consensus discussion by five senior sales staff on a five point scale: 5 = "star performer" ($n = 5$); 4 = "star/steady performer" ($n = 27$); 3 = "steady performer" ($n = 22$); 2 = "steady/poor performer" ($n = 15$); 1 = "poor performer" ($n = 5$).

Procedure

The EPP was administered to seventy-four salespeople of a company that sold cosmetic products to retail outlets. Before the questionnaire was administered, staff were reassured that the results would in no way affect their career. Males comprised 80% of the sample; the average age was 40 yr.

Data aggregation and analysis

Aggregated personality scores were derived from the median scale score corresponding to each of the five rating categories (the median was preferred to the arithmetic mean in order to reduce the possible influence of outliers). Thus, for each personality scale the median score corresponding to a performance rating of 1 was computed, then for a rating of 2, and so on to a rating of 5. This procedure produced as many personality aggregate scores as there were rating categories (i.e. 5). Spearman's (rho) correlation was then computed for ratings and corresponding median personality score (Pearson's parametric equivalent to Spearman's rho would have been highly inappropriate for the type of data reported in this paper because of the non-interval distributions of aggregate personality scores). Significance levels are not of much importance from the perspective of aggregate level correlations: firstly, the elimination of error variance renders their interpretation highly problematic; and secondly, test users are unlikely to be applying tests of statistical significance to their, putative, aggregate level correlations. However, for purposes of illustration and completeness, significance levels have been given.

Results and Discussion

Descriptive statistics, and the correlations between EPP scales and performance, are shown in Table 1.

Table 1. Eysenck Personality Profiler (EPP) means and standard deviations (SD), and individual and aggregate correlations between EPP and performance ratings

Personality	Mean	SD	Individual rho	Aggregate rho
Extraversion (E)	18.1	3.5	0.17	0.90*
Neuroticism (N)	31.8	4.2	0.01	−0.70
Psychoticism (P)	21.0	3.8	−0.04	0.74
Lie (L)	16.3	8.3	0.15	0.79
		N =	74	5

*$P < 0.05$, one-tailed.

Individual level of analysis

All Spearman's rank correlations between personality scales and performance, based on individuals' data, failed to reach the critical level required for significance. As in other comparable studies of sales performance, a positive correlation with extraversion hovered below the 0.20 level.

Aggregate level of analysis

Most of the aggregate data correlations were large, but with a sample size of 5, few correlations exceed the critical value required for statistical significance. However, extraversion was statistically significant, and the correlation of 0.90 was very high. The direction of this correlation was expected: on average, the extraverted groups achieved higher performance ratings than the lower performance groups.

The discrepancy between individual and aggregate correlations was marked; from the individual level correlations, there was no correlation of note; but the aggregate level correlations support the test users' perception of a real and meaningful personality–performance correlation.

STUDY 2: RUGBY REFEREES

Method

Personality questionnaire: NEO-PI

The NEO PI-R (Form S; Costa and McCrae, 1992) is a test based upon the five factor model of personality (Digman, 1990), defining Neuroticism (N), Extraversion (E), Openness (O), Agreeableness (A) and Conscientiousness (C).

Performance criterion

The performance of U.K. Rugby Union Referees was assessed using a 10 point scale (Table 2). Grades were assigned on the basis of decisions made by grading committees, using a national assessment system. The resulting grade is a nationally recognised index of a referee's ability.

Procedure

One hundred and sixty-three Rugby Union Referees were asked to complete the NEO-PI-R. The sample consisted of the entire current active Rugby Football Union elite and "A" List referees in England and a sample of 78 active referees in England. The response rate was 81%, giving a total sample of 132. The percentage of males in the sample was 98%.

Results and Discussion

Descriptive statistics, and the correlations between NEO scales and referees' gradings, are shown in Table 3.

Individual level of analysis

None of the correlations between NEO-PI scores and rugby referees' gradings approached the critical level required for significance.

Table 2. Performance categories used to assess Rugby Union Referee performance

Rating	Category		*n*
10.	"A" List	A1	4
9.		A2	4
8.		A3	23
7.		A4	34
6.	"B" List	B1	4
5.		B2	14
4.		B3	16
3.	"C" List	C1	14
2.		C2	10
1.		C3	9

Table 3. NEO-PI means and standard deviations (SD), and individual and aggregate correlations between NEO and referees' gradings

Personality	Mean	SD	Individual rho	Aggregate rho
Neuroticism (N)	70.11	18.49	−0.01	−0.29
Extraversion (E)	122.10	17.81	−0.03	−0.13
Openness (O)	107.92	20.85	−0.11	−0.65*
Agreeableness (A)	112.49	17.07	−0.02	0.16
Conscientiousness (C)	127.76	19.50	−0.02	−0.42
		$N=$	132	10

*$P<0.05$, two-tailed.

Aggregate level of analysis

Once again, correlations were much higher than for the individual level analysis. One correlation was statistically significant: openness and performance was negatively correlated ($r=-0.65$, $P<0.05$, two-tailed; we did not hypothesize the direction of correlation).

This significant correlation characterises the average high performing U.K. rugby referee as being low on openness. This finding makes sense in terms of the demands placed upon a high-ranking rugby referee. Rugby referees have to maintain discipline and ensure that rules are followed to the letter; these are requirements that referees who score high on openness would find more difficult to impose. Imagine the performance of a referee who was high on fantasy, feelings, aesthetics, etc. (i.e. facets of openness). As the NEO manual (Costa and McCrae, 1992) states, low openness people are "Down-to-earth, practical, traditional, and pretty much set in their ways" (p. 9); in addition, "...their emotional responses are somewhat muted" (p. 15). These are all desirable qualities in a Rugby referee.

GENERAL DISCUSSION

The first aim of this article was to demonstrate that the magnitude of individual level personality–performance correlations are substantially increased by computing aggregate level correlations. Demonstration of this effect is not a new finding in applied psychology; but we have now documented this effect for personality–performance relations in two different occupations.

The second aim of this article was to illustrate how the increase in personality–performance correlations at the aggregate level is consistent with test users' faith in the utility of personality assessment. It is easy to see how the pattern of aggregate level results observed in these studies could be interpreted by test users as indicating the existence of a real and important relationship between personality and performance.

At the individual level of analysis, there were no personality–performance correlations above 0.20, in either of the samples (ignoring sign, the mean correlation across both samples was 0.06). In contrast, at the aggregated level of analysis, the correlations were much higher (the comparable mean correlation was 0.53). These results lend support to our view that test users may be applying implicit personality processes in their search for an explanation of personality-performance associations. If this argument were correct, then test users may be sensitive to the high correlations that exist at the aggregate, or organizational, level of description. However, we have not provided direct confirmation that aggregate data correlations underlie subjective confidence, so, for the present, our conclusions must remain tentative.

Our results point to several hypotheses that may be profitably addressed by future research. First, is there a correlation between test practitioners' faith in personality tests and empirical correlations at the aggregate level of analysis; that is, does confidence follow empirically confirmed high aggregate correlations? Secondly, do practitioners base their estimates of typical (aggregate) personality scores on the basis of means, medians, modes, or some other measure of typical score? Thirdly, what effect does within-cell variance have upon such estimates of typical personality scores?

Although the strong linear correlations reported here for aggregate data appear meaningful, they are, however, somewhat artefactual in the sense that it seems that any weak linear relationship can be transformed into a strong relationship by applying aggregate analysis techniques. Certainly the

use of parametric correlations would have produced spurious and meaningless correlations, as group personality scores tend to lead to a non-interval scale of measurement, with the consequence that one or two outlying data points would lead to a grossly inflated estimate of aggregate coefficients. In contrast, the strong linear relationship in our data represents a real association between aggregate personality and performance data: our analytical approach did not impose restriction on the possibility of zero, positive or negative correlations. But it is true that the probability of a strong correlation is a function of the number of data points; that is, with fewer data points, the influence of individual scores becomes greater. The fact that the aggregate correlations in Study 1 (with 5 data points) were higher than those found in Study 2 (with 10 data points) supports this supposition. This supposition points to another testable hypothesis: is the perceived strength of personality–performance inversely related to number of performance categories used to aggregate data (i.e. the finer the performance grading, the weaker the perceived relationship, irrespective of the actual aggregate correlation)?

If our arguments are valid, then it follows that test users' faith in personality assessment are understandable. However, the conclusion from aggregate-performance correlations that personality can be used for selection, etc. does not logically follow. Aggregate level correlations have very limited implications for individual level use. In most situations, there is simply too much variation around aggregate (average) points for such points to be taken as reliable indicators of the true relationship between personality and performance. Faith and practice must, therefore, be clearly dissociated.

We do not wish to imply that personality effects at work are not important. However, we acknowledge the fact that few studies thus far have shown impressive validities. We suggest that these poor validities result from a lack of work-relevant theoretical models which relate personality processes to work performance. For example, trait × situation interactions are likely to be important, yet these are rarely considered in occupational applications of personality. We suspect that more rigorous theoretical models of personality effects would produce substantially enhanced individual level correlations that would be of considerable practical utility.

In conclusion, weak correlations were found between personality and performance at the individual level of analysis, but strong correlations were found at the aggregate, organizational level of analysis. The nature of and the extent to which test users employ aggregate levels of analysis is not known, but our arguments and data point to a number of testable hypotheses that may shed light upon this problem.

Acknowledgements—We are grateful to Mr Steven Womersley for collection of the data reported in study 2.

REFERENCES

Blinkhorn, S. & Johnson, C. (1990). The insignificance of personality testing. *Nature*, *348*, 671–672.

Corr, P. J. & Gray, J. A. (1995). Attributional style, socialization and cognitive ability as predictors of sales success: A predictive validity study. *Personality and Individual Differences*, *18*, 241–252.

Corr, P. J. & Gray, J. A. (1996). Attributional style as a personality factor in insurance sales performance in the U.K. *Journal of Occupational and Organizational Psychology*, *69*, 83–87.

Costa, P. T. & McCrae, R. R. (1992). *NEO PI-R*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T. & McCrae, R. R. (1995). Primary traits of Eysenck's P-E-N system: Three- and five-factor solutions. *Journal of Personality and Social Psychology*, *69*, 308–317.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*, 417–440.

Eysenck, H. J. & Wilson, G. D. (1991). *The Eysenck Personality Profiler*. London: Corporate Assessment Network Ltd.

Eysenck, H. J., Barrett, P., Wilson, G. D. & Jackson, C. J. (1992). Primary trait measurement of the 21 components of the P-E-N system. *European Journal of Psychological Assessment*, *8*, 109–117.

Fletcher, C. (1991). Personality tests: The great debate. *Personnel Management*, *September*, 38–42.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *Multivariate Behavioural Research*, *22*, 267–305.

Ostroff, C. (1993). Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology*, *78*, 569–582.

Robertson, I. T. & Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational and Organizational Psychology*, *66*, 225–244.

Robertson, I. T. & Makin, P. (1986). Management-selection in Britain: A survey and critique. *Journal of Occupational Psychology*, *59*, 45–57.

Schmitt, N., Gooding, R. Z., Noe, R. A. & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, *37*, 407–422.

Shackleton, V. J. & Newell, S. (1991). Management selection: A comparative survey of methods used in top British and French companies. *Journal of Occupational Psychology*, *64*, 23–36.