Journal of Occupational and Organizational Psychology (2005), 78, 387–409

The British Psychological Society

www.bpsjournals.co.uk

The predictive validity of cognitive ability tests: A UK meta-analysis

Cristina Bertua[1], Neil Anderson[2]* and Jesús F. Salgado[3]

[1] Independent Consultant, London, UK
[2] University of Amsterdam, The Netherlands
[3] University of Santiago de Compostela, Spain

A meta-analysis on the validity of tests of general mental ability (GMA) and specific cognitive abilities for predicting job performance and training success in the UK was conducted. An extensive literature search resulted in a database of 283 independent samples with job performance as the criterion ($N = 13,262$), and 223 with training success as the criterion ($N = 75,311$). Primary studies were also coded by occupational group, resulting in seven main groups (clerical, engineer, professional, driver, operator, manager, and sales), and by type of specific ability test (verbal, numerical, perceptual, and spatial). Results indicate that GMA and specific ability tests are valid predictors of both job performance and training success, with operational validities in the magnitude of .5–.6. Minor differences between these UK findings and previous US meta-analyses are reported. As expected, operational validities were moderated by occupational group, with occupational families possessing greater job complexity demonstrating higher operational validities between cognitive tests and job performance and training success. Implications for the practical use of tests of GMA and specific cognitive abilities in the context of UK selection practices are discussed in conclusion.

Several recent surveys indicate that tests of general mental ability (GMA) and tests of specific cognitive abilities (e.g. numerical, verbal, spatial, etc.) are increasingly popular amongst employer organizations in the UK for selection and assessment purposes (e.g. Hodgkinson & Payne, 1998; Keenan, 1995; Ryan, MacFarland, Baron, & Page, 1999; Salgado & Anderson, 2002; Salgado, Ones, & Viswesvaran, 2001). Whereas in the USA numerous meta-analytic studies have provided predictive and criterion-related validity evidence to support the use of GMA tests in selection (e.g. Hunter & Hunter, 1984; Schmidt, 2002; Schmidt & Hunter, 1998), there has been a notable absence of validity generalization studies in the UK. This is a serious shortcoming in our understanding of the predictive efficacy of such tests. Given their increasing popularity amongst

* *Correspondence should be addressed to Neil Anderson, Department of Work and Organizational Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands (e-mail: N.R.Anderson@uva.nl).*

DOI:10.1348/096317905X26994

employers, selection psychologists and test suppliers in the UK are potentially open to claims of relying upon tests which have not been fully validated through independent meta-analytic procedures combining multiple proprietary tests.

US meta-analyses of cognitive ability tests

A number of meta-analyses have been carried out in the USA investigating the criterion-related validity of GMA and cognitive ability tests (see Schmidt, 2002, Appendix A, for a comprehensive summary of past findings). Amongst these, the largest meta-analyses based on occupational samples are those conducted by Hartigan and Wigdor (1989), Hunter (1986), Hunter and Hunter (1984), and Levine, Spector, Menon, Narayanon, and Canon-Bowers (1996). Overall, these have shown that the average operational validity for GMA and cognitive ability tests ranges from .38 to .47 for overall job performance and from .54 to .62 for training success (re-estimated using Hunter & Hunter's criterion reliability and range restriction estimates). Furthermore, Hunter and Hunter (1984) demonstrated that despite differences in jobs and organizations, the predictive validity of GMA and cognitive ability tests generalizes across samples and settings. Consequently, it has been concluded that GMA and cognitive ability tests are robust predictors for all types of jobs (Salgado, 1999; Salgado, Ones, & Viswesvaran, 2001; Schmidt & Hunter, 1998), and that their validity generalizes across occupations in the USA. However, despite the large body of evidence supporting the validity of GMA and cognitive ability tests, there are a number of limitations within the current body of research. Firstly, there has been a general tendency towards examining general mental ability as a predictor of work behaviour, as opposed to the predictive validity of specific cognitive abilities. Secondly, only limited research has examined the predictive validity of GMA and specific cognitive abilities across different occupational groups. Finally, and perhaps most importantly, in examining these issues, there has been a general reliance on predominantly US samples (Anderson, Born, & Cunningham-Snell, 2001; Schmidt, 2002). As highlighted by Herriot and Anderson (1997), the findings from US meta-analyses have been unreservedly cited as being generalizable to the UK, without consideration of possible cultural, social, legislative, and recruitment and appraisal differences between countries. These differences, it can be argued, may well impact on the magnitude of validities observed in GMA test validity between the USA and UK (see also Salgado & Anderson, 2002, 2003).

European and UK meta-analyses of cognitive ability tests

A comprehensive review of the published studies revealed that no previous meta-analysis which considered the criterion-related validity of GMA tests in the UK has been published. Robertson and Kinder (Robertson & Kinder, 1993; see also Salgado, 1996) published a meta-analysis using data collected in the UK, but this meta-analysis focused on the validity of personality measures. Their meta-analysis did, however, examine the incremental validity of personality measures after partialling-out the variance in the criterion measure attributable to cognitive tests. In their series of recently published papers, Salgado, Anderson, and colleagues have investigated the criterion-related validity of cognitive tests across other countries in the European Union, but no UK-specific meta-analysis appears to have been published to date (Salgado & Anderson, 2003; Salgado, Anderson, Moscoso, Bertua, & de Fruyt, 2003a; Salgado *et al.*, 2003b). This is undoubtedly a notable shortcoming in our understanding of the efficacy of

cognitive ability tests for employee selection in the UK. According to Levy-Leboyer (1994), there are important differences between the US and European organizations in how selection procedures are carried out. This is borne out by subsequent analyses by Salgado and Anderson (2002) into the popularity of cognitive ability tests in Britain, Europe, and the USA as indicated by previous surveys of GMA test use in these countries. Across 16 major surveys conducted over the last 25 years, Salgado and Anderson found that organizations in the UK tended to use GMA measures substantially more than organizations in the USA, despite the dearth of British meta-analytic evidence to support this widespread popularity. Viswesvaran and Ones (2002) have further pointed out that countries in the European community, if considered individually, are relatively homogeneous compared with the USA as they have less within-country diversity. Of any European country, of course, it can be argued that the UK is closest to the USA in terms of its employment legislation (having opted out of the EU Social Chapter, for instance), hours of work, job security, and human resource management practices. As noted by Roe (1989) selection practices and perspectives in other European countries follow less the classical American predictivist model. Instead, they emphasize the social negotiation perspective (e.g. Herriot & Anderson, 1997), prospective employee rights in the procedure, and applicant privacy and expectations of equitable and fair treatment by the prospective employer organization (Levy-Leboyer, 1994). Other researchers have argued that another relevant difference is the difference in size typically between US and European organizations (see, for instance, Salgado *et al.*, 2003a, 2003b). Again, comparisons between the UK and the USA are particularly interesting given the cultural differences between the UK and other European countries, and the adoption by UK organizations of American HR procedures and working practices. Several of the tests upon which primary studies were based in our dataset were either developed in the USA but are popular in the UK for GMA measurement (e.g. the Minnesota Clerical Test, the Differential Aptitude Test, Bennett's Mechanical Comprehension Test), or were UK-developed but are now used also in the USA (e.g. Raven's Progressive Matrices: Jensen, 1998). These overlaps further suggest that similar predictor–criterion relations could be expected across both countries.

Issues concerning the theoretical groundings, development, and use of cognitive ability measures for employee selection have been at the forefront of debate in US industrial, work, and organizational psychology recently (e.g. Ones & Viswesvaran, 2002; Viswesvaran & Ones, 2002). Indeed, the journal *Human Performance* has published a seminal special issue entirely dedicated to the role of GMA in selection and job performance. Given that cognitive tests are used considerably more extensively for selection in Britain than in the USA, it is timely and fitting that debate in the cultural and legislative context of the UK is encouraged. Indeed, major issues such as criterion-related validity, adverse impact, test construction, validation procedures, and claims for the efficacy of cognitive tests for employee selection in the UK have received scant attention (see for instance, Murphy, 2002; Ones & Anderson, 2002; Reeve & Hakel, 2002). As will be highlighted in the following sections, such limitations necessitate a comprehensive analysis of these issues. What is more, in view of the lack of comparable meta-analyses conducted on British samples, a country specific analysis of the validity of GMA and specific cognitive ability tests is warranted in order to accurately assess the predictive validity of such tests in the UK. Therefore, the current investigation sought to address these limitations by conducting the first independent and comprehensive meta-analysis of GMA and specific cognitive ability tests across a range of occupations consisting exclusively of UK samples.

General versus specific cognitive abilities

An extensive body of research conducted over the last 50 years has led to the general consensus that cognitive abilities manifest a hierarchical structure (see for example, Carretta & Ree, 2000; Carroll, 1993; Jensen, 1998; Ree & Carretta, 1998). In conjunction with this, many tests have been developed to measure both GMA and specific cognitive abilities, such as numerical, spatial, verbal, and perceptual ability. However, even in the USA, in contrast to the extensive research regarding the predictive validity of GMA, very little research has been conducted examining the predictive validity of specific cognitive ability tests. For example, Hunter and Hunter (1984) and Hartigan and Wigdor (1989) partially examined this issue by examining the predictive validity of a cognitive ability composite and a perceptual ability composite (as assessed by the GATB) within civil settings. The results from both of these studies revealed that the perceptual ability composite had generally lower predictive validity than the cognitive ability composite. For example, in Hunter and Hunter's (1984) presentation of the US Employment Service validation studies, the mean validities found for the cognitive ability composite ranged from .23 to .58 for job performance, and from .50 to .65 for training success (depending on the job complexity). However, in the case of the perceptual ability composite, mean validities ranged from .24 to .52 for job performance and from .26 to .53 for training success. A further piece of research which supports the conclusion that perceptual ability tests have generally lower predictive validity than general cognitive ability is Hunter's (1980, 1984, cited in Hunter, 1986) reanalysis of Ghiselli's data (1966, 1973). These results revealed that for general cognitive ability validities ranged from .27 to .61 for job performance, and from .37 to .87 for training success (corrected for measurement error and range restriction). However, for perceptual ability, lower estimates ranging from .20 to .46 were found for job performance.

On the question of general versus specific cognitive abilities as predictors of subsequent job performance, findings from meta-analyses conducted in the US have been unequivocal. Several studies indicate GMA to be the most robust predictor with specific abilities adding little or no incremental validity to predictor-criterion relationships (e.g. Carretta & Ree, 1996; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Olea & Ree, 1994; Ree & Carretta, 1994; Ree & Earles, 1991; Ree, Earles, & Teachout, 1994). However, tests of specific cognitive ability are highly popular for selection purposes in the UK, with for instance, many organizations using notionally separate tests of verbal, numerical, and abstract reasoning (that is, regardless of underlying construct correlations with *g*). Meta-analyses in the USA have typically examined the issue of the incremental validity of tests of specific abilities, however, not their 'stand-alone' validity if used by selection practitioners as multiple tests of different aspects of cognitive ability. This is typically the way in which specific ability tests are used for selection in the UK, regardless of existing findings that specific abilities correlate very highly with GMA.

Although not examining the validity of specific cognitive ability tests across a range of job groupings, some research has been conducted within narrower job groupings (e.g. Hirsh, Northrop, & Schmidt, 1986; Levine, Spector, Menon, Narayanan, & Cannon-Banister, Slater, & Radzan, 1962.; Pearlman, Schmidt, & Hunter, 1980: Vinchur, Schippmann, Switzer, & Roth, 1998). For example, Levine *et al.* (1996) examined the criterion validity of perceptual and cognitive ability tests for craft jobs in the utility industry. In their study, they found that perceptual tests demonstrated a corrected validity of .34 when predicting job performance, and .36 when predicting training success. However, these validity estimates may not accurately represent the predictive

validity of perceptual ability tests, since the classification of tests under their perceptual test category is problematic.

The main conclusion to be drawn from these US results is that the magnitude of the predictive validities estimated varies according to the type of cognitive ability test used, and that GMA or overall cognitive ability generally appears to be a better predictor of future job performance and training success than specific cognitive ability tests. In addition, as indicated by Hirsh *et al.*'s results, validity generalization may not be evident for all tests in all cases. However, as mentioned at the outset, the current body of research is limited by the relative paucity of studies comprehensively examining the predictive validity of a range of specific cognitive ability tests across a range of job groupings. Therefore, one of the main aims of the current research was to provide a more detailed examination of the predictive validity of specific cognitive ability tests across a range of occupational groups. Also, in view of the variability in the validity magnitudes reported in these American meta-analytic investigations an important issue was to ascertain a more accurate estimate of the predictive validities of GMA and specific cognitive ability tests in the UK.

Criterion validity across occupational groups

One of the first examinations of the predictive validity of GMA and cognitive ability tests across different occupational groups is Hunter's reanalysis (1986) of Ghiselli's data (1966, 1973). These covered a range of occupational groups including managerial, clerical, sales, protective professions, service workers, vehicle operators, sales clerks, trades and crafts jobs and elementary industrial jobs. Following corrections for sampling error, measurement error, and range restriction, Hunter reported a range of validities from .61 for sales persons to .27 for sales clerks when predicting job performance. For training success, validities ranged from .87 for protective professionals to .37 for vehicle operators. However, due to the unavailability of sample size details and information concerning the variability of the coefficients, Hunter was unable to establish the generalizability of the results across each job family.

Despite this, additional studies have subsequently been conducted which do address this limitation. For example, in Hunter and Hunter's (Hunter, 1986; Hunter & Hunter, 1984) examination of the validity of the GATB in the US Department of Employment, corrected validities were estimated across broad categories of jobs defined by their level of complexity. Overall, they found that the criterion validity of cognitive tests when predicting job performance was moderated by occupational group membership. That is, they found that operational validity was highest for high-complexity jobs, and decreased as the level of job complexity decreased. For example, corrected validities for job performance ranging from .58 for high-level complexity general job groups down to .23 for low-level complexity industrial job groups were reported. For training success, corrected validities ranged from .65 for high-level complexity industrial job groups to .50 for lower complexity general job groups. Therefore, the criterion validity of cognitive ability tests appears to be moderated by job complexity for both job performance and training success, but particularly for the former.

Taken as a whole, these studies indicate that occupational group may be a relevant moderator of the predictive validity of cognitive ability tests. Yet, the moderating effect of occupational group is an issue which has not been comprehensively examined in previous meta-analyses even in the USA, let alone in the UK.

In order to achieve these goals, two work-related criteria were examined, overall job performance ratings and training success. This choice was based on three principal factors: (1) US meta-analyses have only used these two criteria and therefore, since one of the current aims is to compare these results with those of previous US meta-analyses the same criteria were used here; and (2) the practical consideration that these criteria are the most frequently reported in the literature; (3) the scarcity of primary studies including alternative criteria (such as turnover, absenteeism, promotion etc.) would have meant that meta-analyses including such criteria would not have been possible.

To summarize, the current meta-analytic investigation addressed four main research questions:

(1) Are GMA and cognitive ability tests valid predictors of job performance and training success in UK samples?
(2) Does operational validity of GMA and cognitive ability tests generalize across UK samples and settings?
(3) Does operational validity of GMA tests generalize across different occupational groups?
(4) Are the results obtained from this UK investigation comparable to those found in previous US and other European country meta-analyses?

Method

Compilation of database

The process of compiling a database of sufficient scope and size to permit investigation of the current issues entailed a number of key stages. The first of these involved conducting an exhaustive literature search for potential studies to be included. Firstly, an extensive search was conducted using PsycInfo and BIDS databases. Secondly, a manual article-by-article search was performed through major journals and other publications in the field of organizational psychology. For example, the *Journal of Occupational and Organizational Psychology*, *International Journal of Selection and Assessment*, *Journal of the National Institute of Industrial Psychology*, *Occupational Psychology, Personnel Journal*, *Journal of Applied Psychology*, *European Journal of Applied Psychology*, *Psychological Review*, *Human Factor*, *Occupational Psychologist*, *British Journal of Psychology*, and the *Guidance and Assessment Review*, amongst others. Thirdly, test manuals and books thought likely to include data were also inspected for potential studies. Fourthly, individual well-known researchers, practitioners and test publishing companies were contacted and asked for reports containing criterion-related validity data. Finally, the reference sections of obtained articles were also inspected for additional papers not located by other means. Following the collection of studies, two researchers served as judges, independently coding and classifying the studies and the information contained within. The inclusion criteria stipulated were that: (a) studies report a validity coefficient relating to GMA and/or cognitive ability measures and overall job performance and/or training success criteria, (b) only UK samples should be included, (c) samples should consist of employees or trainees, and not students (unless these were part of a formal occupational apprenticeship training programme), (d), there should be sufficient information to enable appropriate classification of the cognitive ability tests (e.g. GMA, verbal and numerical ability) and criterion measures used (i.e. overall job performance, training success).

Classification of GMA and cognitive ability tests

The first step in coding the study details involved classifying the mental ability test measures used in primary studies into the GMA or cognitive ability test type categories of interest within the present investigation. These consisted of measures of general mental ability (*g* or GMA), numerical, verbal, spatial-mechanical, and perceptual-clerical ability tests. As in previous studies (e.g. Ghiselli, 1966), GMA and cognitive ability tests were classified in line with Philip Vernon's classification of tests, according to the construct or ability factors measured (see, for example, Vernon, 1956, 1961; Vernon & Parry, 1949). It is important to note that Vernon's model suggested that two levels captured the hierarchy of abilities, and that more recently, the massive factor analytic work by Carroll (1993) suggested that three levels can better capture the hierarchy of abilities. In both models, the third level corresponds to GMA and Vernon's first level is very similar to Carroll's second level. Ree and Carretta (1994) have also found that Vernon's model arose from factor analyses of the ASVAB in the US army. To enable the classification of measures, descriptions and test information available within individual articles were consulted. Where such information was lacking, or insufficient, clarification was sought from the psychometric literature. This included consulting relevant books (e.g. Carroll, 1993; Ghiselli, 1966; Vernon, 1961, 1972; Vernon & Parry, 1949), articles and test manuals which contained test descriptions or statistical information relating to the underlying ability factors measured. Each mental ability test was classified by each researcher into one of the categories mentioned previously (see Appendix A for a listing of the tests included under each test type categories).

Classification of jobs into occupational categories

The classification of jobs into occupational categories involved using a number of information sources. Firstly, job and occupational category descriptions from individual articles included within the database were used to group jobs according to naturally occurring job types (e.g. all clerical samples were categorized under the clerical job category). In cases where there was insufficient information or where such explicit similarities were not available, additional information was sought to clarify the appropriate classification. This included using information such as: (1) job and task descriptions, for the jobs contained within the individual studies, from the *Dictionary of Occupational Titles* (DOT: US Department of Labor, 1977); (2) job category classifications used in previous studies (e.g. Hunter & Hunter, 1984; Pearlman *et al.*, 1980). Overall, this resulted in the classification of jobs according to seven broad categories for the job performance ratings criterion database: clerical and administrative jobs, engineers, professionals, drivers, operators and spotters, managers and supervisors, sales and advisors. The training success criterion database consisted of six broad categories: clerical and administrative, engineers, health professionals, drivers, operators, coders and air traffic control, trade and skilled workers. In addition to these, a further category for each criterion database was added, including mixed occupational groups cited as such within the original studies (see Appendix B for a listing of the jobs included within each occupational category).

Compilation of validity distributions

The next stage in developing the current database involved compiling validity distributions upon which each meta-analysis could be conducted. Only one validity

coefficient was included from each sample for each ability test and occupational category combination. In cases where more than one coefficient from the same sample was reported (e.g. two numerical ability tests), these were combined using one of two methods. Where correlations between the measures were available, a composite was calculated using Mosier's formula to correct for attenuation (see Hunter & Schmidt, 1990, for a full description). In cases where intercorrelation information was unavailable, average correlations were calculated. The resulting single coefficients were those used within the meta-analyses.

Database

The resulting database consisted of 56 individual papers and books reporting 283 independent samples for the ability test-criteria combination database, including 60 independent samples with overall job performance as the criterion ($N = 13{,}262$), and 223 independent samples with training success as the criterion ($N = 75{,}311$). For the ability test-occupation-criteria combination database, there was a total of 105 independent samples, 43 with overall job performance as the criterion ($N = 6{,}644$), and 62 with training success as the criterion ($N = 20,005$). It is important to note that a number of studies were conducted before 1960 and this could suggest to some readers that possible changes in the nature of jobs, the type of applicants, and other factors might potentially affect the validity of the tests. This problem was exhaustively examined by the panel of the National Sciences Foundation (Hartigan & Wigdor, 1989) and they found no evidence of a decline in validity over time. Also, it must be noted that an examination of the studies included in the database did not reveal that specific tests (e.g. spatial/mechanical tests) were more often used with an occupational group than with other groups.

Procedure

Once the database had been compiled, the psychometric meta-analytic formulas developed by Hunter and Schmidt (1990, 2000) were applied. These allow the estimation of the percentage of variance in observed validities which can be attributed to artifactual errors, and the operational validity one can expect, once artifactual error sources are removed. The artifactual errors considered within the current investigation included, direct range restriction in the predictor scores, predictor and criterion unreliability and sampling error. However, since our interest lies in the operational validity of GMA and cognitive ability tests (as opposed to their theoretical value), the observed mean validity is only corrected for criterion unreliability and range restriction in the predictor. Predictor unreliability estimates are only used to eliminate artifactual variability in the calculation of the standard deviation of the operational validity (*SD*rho; see Hunter & Schmidt, 1990, 2000, for further explanation).

Artifact distributions

To correct for artifactual errors within the meta-analyses, the most common technique is the development of specific artifact distributions for the error sources of interest. Within the current investigation, this involved recording and collating all relevant information pertaining to range restriction and predictor and criterion unreliability, by consulting a number of information sources: (1) primary studies, (2) general references, and (3) test manuals.

Sufficient data regarding range restriction and predictor reliabilities were available to develop specific empirical artifact distributions. These provided a sample-weighted average of .60 ($SD = .24$) for range restriction. This value is similar to the one used by Hunter and Hunter (1984), and Hermelin and Robertson (2001). For predictor reliability, the average test–retest reliability was used (as recommended by Schmidt & Hunter, 1999), resulting in an estimate of .85 ($SD = .05$). In the case of criterion reliability, there was insufficient information to enable the development of specific distributions. Therefore, the alternative of using previously well-established criterion reliabilities was used (see Hunter & Schmidt, 1990). Since the estimate of interest in such cases is the inter-rater reliability (Hunter, 1986; Hunter & Hirsh, 1987; Schmidt & Hunter, 1996), the average reliability estimate of .52 ($SD = .09$) was used (Viswesvaran, Ones, & Schmidt, 1996). Although this estimate is slightly lower than the estimate of .60 used by Hunter and Hunter (1984), additional research does suggest that this is an accurate estimate of job performance reliability (Rothstein, 1990; Salgado & Moscoso, 1996; Salgado *et al.*, 2003a, 2003b). For training success, the sample-weighted average reliability estimate of .80 ($SD = .10$) used by Hunter and Hunter (Hunter & Hunter, 1984; see also Hunter, 1986) was used. Note that these artifact distributions were drawn from previous meta-analyses conducted in the USA. There may, of course, be differences in artifact values across studies conducted in other countries, including the UK. However, one previous UK meta-analysis similarly used these distribution values (Hermelin & Robertson, 2001).

Results

GMA and specific cognitive ability tests

The first series of meta-analyses examined the predictive validity of GMA and specific cognitive ability tests as predictors of job performance and training success. Tables 1, 2, respectively present the results for each ability test – job performance, and training success combination. These show (from left to right) the number of validity coefficients (K) and total sample size (N) upon which the analysis was based. Also shown are the mean observed validities (r) and their standard deviation (SDr), the operational validities one can expect once artifactual error from range restriction in predictor scores and criterion unreliability has been removed (rho), and their standard deviation (SDrho).

Table 1. Meta-analysis results for GMA tests – job performance combinations

Ability test	*K*	*N*	*r*	*SDr*	rho	*SD*rho	%VE	90% CV	Lrho	NSD	LCV
GMA	12	2,469	.22	.15	.48	.24	45	.17	.44	.27	.09
Verbal	14	3,464	.17	.11	.39	.15	61	.20	.35	.18	.12
Numerical	20	3,410	.19	.11	.42	.12	75	.26	.38	.17	.17
Perceptual	7	1,968	.23	.00	.50	.00	242	.50	.45	.14	.27
Spatial	7	1,951	.15	.04	.35	.00	348	.35	.32	.10	.19
Sample total	60	13,262									
Average validity			.19		.42						

Note. K = Number of correlations; N = Total sample size; r = Mean observed validity (sample size weighted); SDr = Standard deviation of observed validity (sample size weighted); rho = Operational validity (observed validity corrected for criterion unreliability and range restriction); SDrho = rho Standard deviation; %VE = Percentage of variance accounted for by artifactual errors; 90% CV = 90% Credibility value).

Table 2. Meta-analysis results for GMA tests – training success combinations

Ability test	*K*	*N*	*r*	*SDr*	rho	*SD*rho	%VE	90% CV	Lrho	NSD	LCV
GMA	53	17,982	.29	.13	.50	.13	64	.33	.45	.19	.21
Verbal	33	12,679	.29	.12	.49	.10	72	.36	.45	.17	.23
Numerical	46	15,925	.32	.12	.54	.09	81	.43	.49	.18	.26
Perceptual	41	13,134	.30	.13	.50	.12	66	.35	.45	.18	.22
Spatial	50	15,591	.24	.07	.42	.00	149	.42	.38	.12	.23
Sample total	223	75,311									
Average validity			.29		.49						

Note. K = Number of correlations; N = Total sample size; r = Mean observed validity (sample size weighted); *SDr* = Standard deviation of observed validity (sample size weighted); rho = Operational validity (observed validity corrected for criterion unreliability and range restriction); *SD*rho = rho Standard deviation; %VE = Percentage of variance accounted for by artifactual errors; 90% CV = 90% Credibility value; Lrho = Lowest (hypothetical) rho; NSD = hypothetical standard deviation of Lrho; LCV = Lowest (hypothetical) CV.

The next two columns present the percentage of variance explained by artifactual errors (%VE) and the 90% credibility values (90% CV). This last figure denotes the validity value at or above which 90% of all true validities lie and, consequently, the minimum value one can expect in 9 out of 10 cases.

Job performance

A total of 60 independent samples with a total sample size of 13,262, contributed to these meta-analyses. The number of independent samples contributing to each ability test – job performance combination meta-analysis ranged from a maximum of 20 for numerical ability tests to a minimum of seven for both perceptual and spatial ability tests. As indicated by the operational validities reported, all ability tests demonstrate good predictive validity for overall job performance. Perceptual ability tests emerged as the best predictors, with an operational validity of .50 (SD = .00). All the variance in the observed validity was explained by artifactual errors and consequently the 90% credibility value was also .50. This indicates that the validity of perceptual tests does generalize across samples and settings. The percentage of variance explained is also indicative of second order sampling error, in which case the sample of coefficients included within the current analysis may not be totally representative of the general population. However, as highlighted by Hunter and Schmidt (1990), the main impact of second-order sampling error is not on the estimation of means or operational validities, but rather its main impact is on the estimates of standard deviations. In view of this, the 90% credibility value observed may change as the number of studies and sample sizes increase.

The next highest predictor was GMA, which also showed a high validity, since the operational validity was .48 (SD = .24). In this case, the 90% credibility value of .17 also indicated that the validity of GMA tests generalizes across samples and settings. However, in addition to this the percentage of variance explained by artifactual errors (45%) and the standard deviation of the operational validity (*SD*rho = .24) indicates that other factors may moderate the operational validity magnitude of GMA tests.

The third best predictors of job performance ratings were numerical ability tests, which showed an operational validity of .42 (SD = .12) and a 90% credibility value of

.26, indicating that the validity of numerical validity tests also generalizes across samples and settings. The percentage of variance explained by artifactual errors (75%) also indicates that the remaining variance can be considered attributable to additional artifactual error sources, not considered within the current analyses (e.g. imperfect construct measurement, range restriction in criterion scores and clerical errors. See also, Hunter & Schmidt, 1990 for full listings of possible error sources). Verbal and spatial ability tests showed slightly lower operational validities of .39 ($SD = .15$) and .35 ($SD = .00$), respectively. Nonetheless, in both cases the 90% credibility values indicate that both have generalized validity across samples and settings. However, there was evidence of second order sampling error.

As can be seen in Table 1, the standard deviation of rho for GMA, verbal ability and numerical ability is larger than the standard deviation of the observed validity. This is due to the fact that not all the observed variability was explained for the artifactual errors, and that the residual variance is corrected for the effects of predictor and criterion reliability in order to have an unbiased estimate of the standard estimate of rho.

Some cells in Table 1 have a relatively small number of studies although the number is still acceptable for meta-analysis. However, we conducted a so-called 'file-drawer analysis' (Rosenthal, 1979; Hirsh *et al.*, 1986). With regard to this point, Ashworth, Callender, Osburn, and Boyle (1992) have developed a method for assessing the vulnerability of validity generalization results to unrepresented or missing studies. Ashworth *et al.* (1992) suggested calculating the effects on validity when 10% of studies are missing and their validity is zero. Therefore, we calculated additional estimates to represent what the validity would be if we were unable to locate 10% of the studies carried out and if these studies showed zero validity. The last three columns in Table 1 report these new (hypothetical) estimates for every design cell: the lowest rho value, new standard deviation, and lowest 90% CV. As can be seen, adding 10% of the studies with zero validity has no effect on our conclusion that there is validity generalization for GMA and specific cognitive ability for predicting job performance.

Training success

As reported in Table 2 the total number of validity coefficients ($K = 223$) and sample size ($N = 75,311$) contributing to this series of analyses was larger than that for the job performance analyses. Across the range of ability type test – training success combinations, the number of coefficients ranged from a maximum of 53 to a minimum of 33, with sample sizes ranging from 17,982 to 12,679. Consequently, the large number of coefficients and huge total sample size can be expected to assure the stability of the results.

The results indicate that all ability tests are good predictors of training success. Numerical ability tests emerged as the best predictors with an operational validity of .54 ($SD = .09$). About 81% of the variance was explained by artifactual error and the 90% credibility value was .43. Consequently, it can be concluded that the validity of numerical tests does generalize across samples and settings and furthermore, there is little room for moderators. The next best predictors were GMA and perceptual ability tests, both showing an operational validity of .50 ($SD = .13$ and .12, respectively for GMA and perceptual tests). A similar percentage of the variance was explained in both cases, with 64% explained for GMA tests and 66% for perceptual tests. Finally the 90% credibility values were also similar, with .33 for GMA tests and .35 for perceptual ability

tests. Therefore, the validity of both GMA and perceptual ability tests can be seen to generalize across samples and settings.

Verbal ability tests were also found to have a high operational validity (.49, $SD = .10$), and the 90% credibility value of .36 indicates that their validity generalizes across samples and settings. The final ability test type analysed was spatial ability tests. These showed an operational validity of .42 ($SD = .00$) and the 90% credibility value was identical as all of the observed variance was explained by artifactual errors. Thus, the validity of these tests also generalizes across samples and settings. However, there was also evidence of second order sampling error.

The results of the file-drawer analysis using the Ashworth *et al.* (1992) method appear in the last three columns of Table 2. Although the number of studies and the total sample size did not require this analysis, it was carried out as an additional confirmation of our conclusions. The results of these file-drawer analyses also showed that, for training success, there is validity generalization and that the magnitudes of the new rho estimates are very similar to the original ones.

Occupational groups

The following series of analyses examined the predictive validity of GMA tests as predictors of both job performance and training success across the different occupational groups represented within the current database. In this series of meta-analyses we used the same studies included in the previous meta-analyses, but we have not included studies in which specific cognitive ability tests were used as an estimate of GMA. This was done because there were not a sufficient number of studies to examine the validity of specific cognitive ability for each occupational group. This decision resulted in a smaller number of studies in comparison with the meta-analyses reported in Tables 1 and 2. The first series of meta-analyses, looking specifically at job performance, are presented in Table 3, while the results for the training success criterion are presented in Table 4.

Table 3. Meta-analysis results for GMA tests: occupation – job performance combinations

Occupation	*K*	*N*	*r*	*SDr*	rho	*SD*rho	%VE	90% CV	Lrho	NSD	LCV
Clerical	5	628	.14	.07	.32	.00	177	.32	.29	.09	.17
Engineer	5	542	.33	.24	.70	.42	30	.16	.64	.45	.06
Professional	4	348	.36	.18	.74	.23	61	.45	.67	.31	.28
Driver	2	293	.16	.09	.37	.00	109	.37	.34	.11	.20
Operator	9	3,105	.24	.05	.53	.00	365	.53	.48	.15	.29
Manager	5	302	.33	.01	.69	.00	200	.69	.63	.20	.37
Sales	6	483	.25	.20	.55	.31	46	.15	.50	.34	.07
Miscellaneous	7	943	.18	.08	.40	.00	166	.40	.36	.11	.22
Sample Total	43	6,644									

Note. *K* = Number of correlations; *N* = Total sample size; *r* = Mean observed validity (sample size weighted); *SDr* = Standard deviation of observed validity (sample size weighted); rho = Operational validity (observed validity corrected for criterion unreliability and range restriction); *SD*rho = rho standard deviation; %VE = Percentage of variance accounted for by artifactual errors; 90% CV = 90% Credibility value; Lrho = Lowest (hypothetical) rho; NSD = Hypothetical standard deviation of Lrho; LCV = Lowest (hypothetical) CV.

Table 4. Meta-analysis results for GMA: occupation – training success combinations

Occupation	*K*	*N*	*r*	*SDr*	rho	*SD*rho	%VE	90% CV	Lrho	NSD	LCV
Clerical	8	1,989	.33	.13	.55	.11	75	.41	.50	.19	.26
Engineer	5	1,381	.39	.15	.64	.14	68	.46	.58	.23	.29
Professional	3	295	.35	.14	.59	.08	88	.49	.54	.19	.30
Driver	3	1,674	.28	.06	.47	.00	206	.47	.43	.14	.25
Operator	17	4,322	.32	.12	.54	.07	86	.45	.49	.17	.27
Skilled	12	3,086	.33	.14	.55	.14	65	.37	.50	.21	.24
Miscellaneous	14	7,258	.33	.10	.55	.00	104	.55	.50	.16	.30
Sample total	62	20,005									

Note. *K* = Number of correlations; *N* = Total sample size; *r* = Mean observed validity (sample size weighted); *SDr* = Standard deviation of observed validity (sample size weighted); rho = Operational validity (observed validity corrected for criterion unreliability and range restriction); *SD*rho = rho Standard deviation; %VE = Percentage of variance accounted for by artifactual errors; 90% CV = 90% Credibility value; Lrho = Lowest (hypothetical) rho; NSD = Hypothetical standard deviation of Lrho; LCV = lowest (hypothetical) CV.

Job performance

The database used for this series of analyses consisted of 43 coefficients with a total sample size of 6,644. The largest operational validity found was for professional occupations. For this group of jobs the operational validity was .74 (*SD* = .23). Furthermore, the 90% credibility value of .45 indicates that validity generalizes across professional jobs. However, although a large percentage of the variance was explained by artifactual errors (61%), the results indicate that there may be scope for an examination of possible moderating factors amongst this occupational group. The operational validities estimated for engineer and manager jobs were also high, with GMA tests showing an operational validity of .70 (*SD* = .42) and .69 (*SD* = .00) for engineers and managers, respectively. In the case of managers, all of the observed variability in validities was explained by artifactual errors and consequently the 90% credibility value was also .69. Thus, the validity of GMA for predicting overall job performance generalizes across managerial occupations. However, there was also evidence of second order sampling error. The 90% credibility value of GMA tests for engineer occupations (.16) also indicates that their validity generalizes across all engineering occupations. However, this value along with the percentage of variance explained (30%) and the standard deviation of rho (.42) indicates that moderators may impact on the validity observed for these measures.

The next highest ranking operational validities were for sales and operator occupations. For these occupations GMA tests were found to have an operational validity of .55 (*SD* = .31) and .53 (*SD* = .00) for sales and operator occupations, respectively. The 90% credibility values for both occupations also indicated that validity generalizes across both occupations (90% CV = .15 and .53, respectively). However, in the case of sales occupations, additional moderators may impact on the validity of GMA tests. Furthermore, there is also evidence of second order sampling error for operator occupations.

The final three occupational groups analysed were driver, clerical and mixed occupational groups. Amongst these groups, GMA tests were found to have moderate to high operational validities ranging from .32 (*SD* = .00) for clerical jobs, .37 for driver jobs (*SD* = .00), to .40 (*SD* = .00) for mixed occupations. In all cases, all of the variance

was accounted for by the artifactual error sources considered here, and consequently validity generalized across all occupations. However, there was also evidence of second order sampling error.

The file drawer analysis showed that the addition of 10% of new studies with zero validity had no significant effects on the validity magnitude and that, therefore, the conclusions remain the same for all occupations.

Training success

The operational validity magnitudes of GMA tests were all large, ranging from .64 for engineering occupations to .47 for driver occupations (see Table 4). Moreover, apart from indicating that they are very good predictors of training success, the 90% credibility values indicated that their validity generalizes across occupational groups. Ninety-percent credibility values of .46 and .49 were found for engineer and professional occupations (respectively). In both cases the percentage of variance explained was high, with 68% and 88% being explained for engineer and professional occupations, respectively. Furthermore, in the case of professional occupations the remaining variability in the validity of GMA tests can be considered attributable to additional potential error sources.

Clerical, skilled, operator and mixed occupations all showed very similar operational validity magnitudes (.55 for clerical, skilled and mixed occupational groups and .54 for operator jobs). Furthermore, since the variance in validities was largely, if not totally, accounted for by artefactual errors, validity generalized across each occupational group, with 90% credibility values of .41, .37, .45, and .55 for clerical, skilled, operator and mixed occupational groups, respectively. The lowest operational validity of .47 ($SD = .00$) was for driver jobs, although as in all other occupations examined, this validity is still of sufficient magnitude to be of practical value. All of the variance in GMA's validity was accounted for by artefactual error and consequently, the 90% credibility value was identical to the operational validity. Therefore, it can be concluded that the validity of GMA tests generalizes across driver occupations. Nevertheless, the evidence of second order sampling error indicates that the 90% credibility value may vary as sample sizes increase.

As was found in the previous meta-analyses reported in this article, the results of the file-drawer analyses also showed that, for training success, there is validity generalization and that the magnitude of the new rho estimates was very similar to the original ones. Therefore, the conclusions remain the same after this analysis was done.

Discussion

Taken as a whole, the results of the present investigation indicate that GMA and cognitive ability tests are robust predictors of job performance and training success across a wide range of occupations in the UK. Furthermore, while some differences were observed across different occupational groups and different criteria, the findings from the present study are largely in line with those found in earlier meta-analytic studies in the USA.

General verses specific mental abilities

The crucial overall finding from this series of meta-analyses is that all GMA and cognitive ability tests included within the present investigation were found to be valid predictors of job performance and training success. For job performance, the variation of operational validities observed ranged from .50 (perceptual ability tests) to .35

(spatial tests), indicating that all tests demonstrate moderate to high predictive validity. For training success, operational validities were even greater, with validities ranging from .54 for numerical ability tests to .42 for spatial ability tests. The larger operational validities observed for training success appear consistent with previous research, which reveals a tendency for higher predictive validities for training criteria compared with job performance (e.g. Pearlman *et al.*, 1980). Furthermore, contrary to previous research (Hirsh *et al.*, 1986) which failed to find validity generalization for some tests, all the tests analysed here showed positive credibility values, which were substantially different from zero, thus indicating that validity does generalize when predicting job performance and training success criteria. It is interesting to note that, when comparing the differences in validity for GMA versus specific cognitive ability tests, the 90% credibility intervals are completely overlapped for job performance and training success. In other words, the GMA credibility interval included the respective intervals for verbal, numerical, perceptual and spatial abilities.

A note of caution is warranted with regard to direct comparisons between findings emerging from meta-analyses computed using different databases of primary studies. While such comparisons are possible and valuable we should be mindful of differences in the composition and distribution of primary studies, especially concerning differences in the distribution of job complexity across primary studies (Salgado & Anderson, 2002, 2003). Note, for instance that Schmidt (2002) also found operational validities in the region of .50 for perceptual ability tests for jobs of similar complexity to those we included in the present UK-based meta-analysis. Job complexity has emerged from several meta-analyses internationally as the principal moderator of predictor-criterion relationships, and indeed this was the case in the present meta-analysis of UK studies of tests of GMA and specific abilities. We do not argue that different meta-analyses internationally cannot be compared *per se*, simply that some caution is warranted in comparing the distributions of primary studies especially in terms of job complexity differences.

An interesting finding of the present study concerns the variability in the magnitudes of validities observed for the different ability tests examined. For example, when predicting job performance, GMA and perceptual ability tests demonstrated the highest predictive validities (rho = .48 and .50, respectively). This pattern was similar for the training success criterion, where both GMA and perceptual ability tests showed an operational validity of .50, and numerical ability tests showed an operational validity of .54. Both sets of results are slightly surprising in view of previous research demonstrating that perceptual ability tests demonstrate lower predictive validities than GMA tests (e.g. Hartigan & Wigdor, 1989; Hunter, 1980, 1984, cited in Hunter, 1986; Hunter & Hunter, 1984). However, an important point to note, within the current analyses, is that both the *SD*rho and the % VE for GMA tests (particularly when predicting job performance) indicate that there is room for moderators.

With respect to the findings for perceptual tests, a further point to note is the possibility that tests included within the perceptual-clerical test category may have been more *g* saturated than for the other types of tests we examined in this study. For example, factorial analysis of clerical and instructions tests has revealed that such measures can prove to be as good a measure of general mental ability as abstraction and matrices tests (see for example, Vernon, 1949). Consequently, the high operational validities observed here for perceptual tests may be partly due to the tests' measurement of general mental ability in addition to pure clerical and perceptual ability.

A second point to note is the relative consistency of the operational validity magnitudes for the different ability tests examined here. These results are of particular

interest, in view of current research examining the incremental validity of specific cognitive abilities. This has shown that, for training success and job performance, GMA is the best predictor, with little incremental validity for specific cognitive abilities (e.g. Carretta & Ree, 1996; McHenry *et al.*, 1990; Olea & Ree, 1994; Ree & Carretta, 1994; Ree & Earles, 1991; Ree *et al.*, 1994). However, it was not possible to investigate this hypothesis directly within the current study, and therefore, any hypothesis must remain speculative.

Operational validity across occupational groups

Analysing the validity of GMA tests across different occupational job groups also provided support for their validity as predictors for job performance and training success. In predicting job performance, GMA tests produced moderate to high operational validities across all occupational categories, with values ranging from .74 for professional occupations, to .32 for clerical occupations. Furthermore, GMA tests demonstrated generalized validity across all occupational groups. Comparing the magnitude of these operational validities with those from previous US meta-analyses reveals a number of differences. For example, GMA tests demonstrated operational validity magnitudes of .69 for managers, .55 for sales, and .37 for drivers, all of which are values greater than the validities reported for comparable US samples (e.g. Hunter & Hunter, 1984; Vinchur *et al.*, 1998). Conversely, for clerical occupations the operational validity of .32 reported here is lower than the validity of .52 reported by Pearlman *et al.* (1980). It should be acknowledged, of course, that these differences may be due to differences in the samples used in primary studies in the USA compared with the current UK meta-analyses, and variations in the artifact corrections applied. Interestingly, and in accordance with earlier meta-analyses carried out in the USA, tests of GMA showed considerably higher operational validity for professional and managerial job roles. Paradoxically, practitioners may believe, and indeed may have experienced, that such tests are less popular for senior appointments due to a misbelief that they lack job-related validity; the results of our meta-analysis on a large sample of UK occupational groups strongly refutes this erroneous belief.

One point to note is that, as with any meta-analysis, the primary studies we included in the present study were published historically and over many years. In fact, primary studies date from the 1950s up until the present day (see asterisked papers in the References). What effect might this historic dependency have upon the applicability of our findings to present-day and future selection practices with regard to GMA test validity? As we noted, firstly this will be the case with any meta-analysis, as of course such quantitative reviews can only be done upon previously published and conducted primary studies. Secondly, and in addition, over this period job roles and working practices have obviously changed, and indeed will continue to do so (e.g. Parker & Wall, 2001). Organizations are becoming more complex, delayered, and decentralized, for instance. Job roles are becoming less stable, less routinized, and less specialized as a consequence. This has led some authors to speculate that, if anything, general mental ability is likely to become more rather than less important in the future (Anderson, Lievens, van Dam, & Ryan, 2004). We therefore suggest that measures of GMA in selection are likely to remain as important predictors of job performance, although future research is called for to examine any changes in criterion-related validity as a result of changes in the nature of work and job design in organizations. The currently available meta-analytic findings to the present day suggest that operational validity remains very stable over time (Hartigan & Wigdor, 1989).

Implications for practice

In view of the lack of previous meta-analysis in this country, practitioners have relied upon either primary in-house validity studies or upon the assumption that US meta-analysis findings will generalize unabated to the context of selection and assessment in the UK. However, the findings presented here are particularly informative since they provide unambiguous support for the use of tests of GMA and specific cognitive abilities in the UK, based upon a large-scale meta-analysis of primary validity studies which had not been conducted previously in this country. Such findings therefore have obvious implications for selection practices in UK organizations (Anderson, Herriot, & Hodgkinson, 2001). In their meta-analysis of countries across the European community, Salgado and Anderson (2003) conclude that, 'The magnitude of the operational validities found suggests that GMA measures may be the best single predictor for personnel selection for all occupations' (pp. 16). The findings of the present meta-analysis specifically into UK studies further supports this practical implication: Selection practitioners and HR professionals in UK organizations should be encouraged to use psychometrically developed cognitive ability tests regardless of job type, hierarchical seniority, potential future changes in job role composition, or whether the tests are principally for general of specific cognitive abilities. Moreover, these findings highlight the importance of research-based practice in selection psychology and provide unequivocal evidence for the continued and expanded use of GMA tests for employee selection in UK organizations (Anderson, 2005). The present findings also serve to further understanding of GMA's contribution to the prediction of job performance and training success across different occupational groups. For example, as found in the present analysis, the validity magnitudes one can expect can vary substantially depending on occupational group. Therefore, such issues need to be taken into consideration when incorporating GMA tests into the selection process. The present results further suggest that within particular occupational groups additional factors may also serve to moderate the predictive validity of GMA, particularly for predicting job performance. Thus, additional research will be needed to further examine and quantify the factors that moderate criterion-related validity of GMA across difference occupational groups.

Another set of practical implications stems from our findings regarding the validity of tests of GMA versus tests of specific cognitive abilities for selection in the UK. Our results are unambiguous and in general concur with those reported by Salgado *et al.* (2003a) for other European countries. We found that both tests of GMA and tests of specific cognitive abilities are strong predictors of job performance and training success. In most cases operational validities were found to be in the magnitude of .4–.5, with only the operational validity of spatial tests for predicting job performance falling somewhat below this level (rho = .35, see Table 1). These operational validities can be taken to suggest that tests of GMA and of specific cognitive abilities are robust predictors of subsequent job success in the UK cultural context, and that selection practitioners could justifiably use either type of test purely on the grounds of criterion-related validity.

In conclusion, the present meta-analytic investigation sought to address a number of limitations in the existing body of research by investigating whether, and to what extent, GMA and cognitive ability tests are valid predictors of job performance and training success in UK samples, and occupational groups in the UK. The results demonstrated that GMA tests and tests of specific cognitive abilities are valid predictors of both job performance and training success, and that validity generalizes across samples and settings in the UK. Occupational grouping was also found to moderate the predictive validity GMA tests, with higher operational validities found for occupational groups with

higher job complexity. However, particularly when predicting job performance, the results also indicated that there is scope for additional moderating factors, in particular occupational families. Finally, while some differences were observed, when taken as a whole, the present results are comparable to those observed in previous meta-analyses conducted in the USA and other European countries.

Acknowledgements

We wish to express our thanks to Paul Sparrow and John Arnold as action editors on this paper. We also acknowledge with gratitude the valuable comments of Deniz Ones and the anonymous reviewers on an earlier version of this paper. All three authors contributed equally to this paper which is based upon the first author's MSc dissertation at Goldsmith College, University of London. The preparation of this manuscript was partially supported by funding from the Army Personnel Research Establishment to Neil Anderson and a grant BSO2001-3070 from the Ministerio de Ciencia y Tecnología (Spain) to Jesús F. Salgado. The opinions presented in this paper are the authors' and do not relate to either of these funding agencies.

References

Anderson, N. (2005). Relationships between practice and research in personnel selection: Does the left hand know what the right hand is doing? In A. Evers, N. Anderson, O. Smit-Voskuyl (Eds.), *The Blackwell handbook of selection*. Oxford: Blackwell.

Anderson, N., Born, M., & Cunningham-Snell, N. (2001). Recruitment and selection: Applicant perspectives and outcomes. In N. Anderson, D. S. Ones, H. K. Sinargil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology*, Vol. 1. London/New York: Sage.

Anderson, N., Herriot, P., & Hodgkinson, G. P. (2001). The practitioner–researcher divide in industrial, work and organizational (IWO) psychology: Where are we now and where do we go from here? *Journal of Occupational and Organizational Psychology*, *74*, 391–411.

Anderson, N., Lievens, F., van Dam, K., & Ryan, A. M. (2004). Future perspectives on employee selection: Key directions for future research and practice. *Applied Psychology: An International Review*, *53*, 487–501.

Ashworth, S. D., Osburn, H. G., Callnder, J. C., & Boyle, K. A. (1992). The effects of unrepresented studies on the robustness of validity generalization results. *Personnel Psychology*, *45*, 341–361.

*Banister, D., Slater, P., & Radzan, M. (1962). The use of cognitive tests in nursing candidate selection. *Occupational Psychologist*, *36*, 75–78.

*Bartram, D., & Dale, H. C. A. (1982). The Eysenck Personality Inventory as a selection test for military pilots. *Journal of Occupational Psychology*, *55*, 287–296.

Carretta, T. R., & Ree, M. J. (1996). Factor structure of the air force officer qualifying test: Analysis and comparison. *Military Psychology*, *8*, 29–42.

Carretta, T. R., Ree, M. J. (2000). General and specific cognitive and psychomotor abilities in personnel selection: The prediction of training and job performance. In J. F. Salgado (Ed.), Special issue on: Personnel selection at the beginning of a new millennium: A global and international perspective. *International Journal of Selection and Assessment, Vol. 8*(4) 227–236.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

*Castle, P. F. C., & Garforth, F. I. (1951). Selection, training and status of supervisors. *Occupational Psychologist*, *25*, 109–123.

*Cleary, T. A. (1968). Test Bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124.

*Farmer, E. (1930). A note on the relation of certain aspects of character to industrial proficiency. *British Journal of Psychology*, *21*, 46–49.

*Feltham, R. (1988). Validity of a police assessment centre: A 1–19 year follow-up. *Journal of Occupational Psychology*, *61*, 129–144.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. London: Wiley.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology*, *26*, 461–477.

Gottfredson, L. S. (1997). Why *g* matters: The complexity of everyday life. *Intelligence*, *24*(1), 70–132.

*Handyside, J. D., & Duncan, D. C. (1954). Four years later: A follow-up of an experiment in selecting supervisors. *Occupational Psychology*, *28*, 9–23.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization minority and the general aptitude test battery*. Washington, DC: National Research Council.

*Heim, A. W. (1946). An attempt to test high-grade intelligence. *British Journal of Psychology*, *37–38*, 71–81.

*Henderson, P., & Boohan, M. (1987). SHL's Technical Test Battery: Validation. *Guidance and Assessment Review*, *3*(6), 3–4.

*Henderson, P., & Hopper, F. (1987). SHL's TTB: Development of norms. *Guidance and Assessment Review*, *3*(6), 5–6.

*Henderson, P., Lockhart, H., & O'Reilly, S. (1987). SHL's Technical Test Battery: Test–retest reliability. *Guidance and Assessment Review*, *3*(3), 4–5.

*Henderson, P., & O'Hara, R. (1990). Norming and validation of SHL's Personnel Test Battery in Northern Ireland. *Guidance and Assessment Review*, *6*(4), 2–3.

Hermelin, E., & Robertson, I. T. (2001). A critique and standardization of meta-analytic validity coefficients in personnel selection. *Journal of Occupational and Organizational Psychology*, *74*, 253–277.

Herriot, P., & Anderson, N. (1997). Selecting for change: How will personnel selection psychology survive? In N. Anderson & P. Herriot (Eds.), *International handbook of selection and assessment*. London, UK: Wiley.

Hirsh, H. R., Northrop, L. C., & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology*, *39*, 399–420.

Hodgkinson, G. P., & Payne, R. L. (1998). Graduate selection in three countries. *Journal of Occupational and Organizational psychology*, *71*, 359–365.

Hofstede, G. (1980). *Culture's consequences. International differences in work-related values*. Beverly Hills, CA: Sage.

Hunter (1980). *The dimensionality of the General Aptitude Test Bettery (GATB) and the dominance of general factors over specific factors in the prediction of job performance.* Washington, DC: U. S. Employment Service, U. S. Department of Labor.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Intelligence*, *29*, 340–362.

Hunter, J. E. & Hirsh, H. R. (1987). Applications of meta-analysis. *International Review of Industrial and Organizational Psychology, 2*, 321–357. Chichester, UK: Wiley.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72–98.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. London: Sage Publications.

Hunter, J. E. & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. In J. F. Salgado (Ed.), Special issue on: Personnel selection at the beginning of a new millennium: A flobal and international perspective. *International Journal of Selection and Assessment, Vol. 8*(4) 275–292.

Jensen, A. R. (1998). *The* g *factor: The science of mental ability*. Westport, CT: Praeger.

*Jones, E. S. (1917). The Woolley-test series applied to the detection of ability in telegraphy. *Journal of Educational Psychology*, *8*, 27–34.

Keenan, T. (1995). Graduate recruitment in Britain: A survey of selection methods used by organizations. *Journal of Organizational Behavior*, *16*, 303–317.

Levine, E. L., Spector, P. E., Menon, P. E., Narayanon, L., & Cannon-Bowers, J. (1996). Validity generalisation for cognitive, psychomotor, and perceptual tests for craft jobs in the utility industry. *Human Performance*, *9*, 1–22.

Levy-Leboyer, C. (1994). Selection and assessment in Europe. In H. C. Triardis, M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology*, Vol 4. Palo Alto, CA: Consulting Psychologists Press.

*Lummis, C. (1946). The relation of school attendance to employment records, army conduct and performance in tests. *British Journal of Educational Psychology*, *16*, 13–19.

*Mace, C. A. (1950). The human problems of the building industry: Guidance, selection and training. *Occupational Psychology*, *24*, 96–104.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology*, *43*, 335–354.

*Montgomery, G. W. G. (1962). Predicting success in engineering. *Occupational Psychologist*, *36*, 59–80.

Murphy, K. R. (2002). Can conflicting perspectives on the role of *g* in personnel selection be resolved? *Human Performance*, *15*, 173–186.

*Newman, S. H., & Howell, M. A. (1961). Validity of forced choice items for obtaining references on physicians. *Psychological Reports*, *8*, 367.

*Nyfield, G., Gibbons, P. J, Baron, H., & Robertson, I. (1995, May). *The cross-cultural validity of management assessment methods*. Paper presented at the 10th Annual SIOP Conference, Orlando, USA.

Parker, S. K., & Wall, T. D. (2001). Work design: Learning from the past and mapping a new terrain. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology* (Vol. 1, pp. 90–109). London, UK: Sage.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology*, *79*, 845–851.

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.

Ones, D. S., & Viswesvaran, C. (2002). Introduction to the special issue: Role of general mental ability in industrial, work, and organizational psychology. *Human Performance*, *15*, 1–3.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalisation results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, *65*, 373–406.

*Petrie, A., & Powell, M. B. (1951). The selection of nurses in England. *Journal of Applied Psychology*, *35*, 281–286.

Pettersen, N., & Tziner, A. (1995). The cognitive ability test as a predictor of job performance: Is its validity affected by job complexity and tenure within the organization? *International Journal of Selection and Assessment*, *3*, 237–241.

Ree, M. J., & Carretta, T. R. (1994). Factor analysis of the ASVAB: Confirming a Vernon-like structure. *Educational and Psychological Measurement*, *54*, 459–463.

Ree, M. J., & Carretta, T. R. (1998). General cognitive ability and occupational performance. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology*, Vol. 3. London: Wiley.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than g. *Personnel Psychology*, *44*, 321–332.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology*, *79*, 518–524.

Reeve, C. L., & Hakel, M. D. (2002). Asking the right questions about *g*. *Human Performance*, *15*, 47–74.

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*, 225–241.

Roe, R. A. (1989). Designing selection procedures. In H. P. Herriot (Ed.), *Assessment and selection in organizations*, Chichester, UK: Wiley.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 238–241.

*Ross, J. (1962). Predicting practical skill in engineering apprentices. *Occupational Psychologist*, *36*, 69–74.

Rothstein, H. R. (1990). Inter-rater reliability of job performance ratings: Growth to asymptote level with increasing. *Journal of Applied Psychology*, *75*, 322–327.

Ryan, A. M., MacFarland, L., Baron, H., & Page, R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology*, *52*, 359–391.

Salgado, J. F. (1996). Personality and job competences: A comment on the Robertson and Kinder (1993) study. *Journal of Occupational and Organizational Psychology*, *69*(4), 346–351.

Salgado, J. F. (1999). Personnel selection methods. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology*, 14 (pp. 1–54). UK: Wiley.

Salgado, J. F., & Anderson, N. (2002). Cognitive GMA testing in the European community: Issues and evidence. *Human Performance*, *15*, 75–96.

Salgado, J. F., & Anderson, N. (2003). Validity generalization of GMA tests across countries in the European community. *European Journal of Work and Organizational Psychology*, *12*, 1–17.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities as predictors of work behaviors: A European meta-analysis. *Personnel Psychology*, *56*, 573–605.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology*, *88*, 1068–1081.

Salgado, J. F., Ones, D., & Viswesvaran, C. (2001). Predictors used for personnel selection: An overview of constructs, methods and techniques. In N. Anderson, D. S. Ones, H. K. Sinargil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology*, Vol. 1. London/New York: Sage.

Salgado, J. F., & Moscoso, S. (1996). Meta-analysis of the interrater reliability in validity studies of personnel selection. *Perceptual and Motor Skills*, *83*, 1195–1201.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance*, *15*, 187–210.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199–223.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, *27*, 183–198.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981). Validity generalisation results for two job groups in the petroleum industry. *Journal of Applied Psychology*, *66*(3), 261–273.

Schmidt, F. L., Ocasio, B. P., Hillery, J. M., & Hunter, J. E. (1985). Further within-setting empirical tests of the situational hypothesis in personnel selection. *Personnel Psychology*, *38*, 509–524.

*SHL (1989). *Validation review*. UK: Saville & Holdsworth.

*Smith, M. C. (1976). A comparison of the trainability assessments and other tests for predicting the practical performance of dental students. *International Review of Applied Psychology*, *25–26*, 125–130.

*Sneath, F., Thakor, M., & Medjuck, B. (1976). *Testing of people at work*. London: Institute of Personnel Management.

*Srinivasan, V., & Weinstein, A. G. (1973). Effects of curtailment on an admissions model for a graduate management program. *Journal of Applied Psychology*, *58*(3), 339–346.
*Stanbridge, R. H. (1936). The occupational selection of aircraft apprentices of the Royal Air Force. *Lancet*, *230*, 1426–1430.
*Stevenson, M. (1942). Summaries of researches reported in degree theses. *British Journal of Psychology*, *12*, 182.
*Stratton, G. M., McComas, H. C., Coover, J. E., & Bagby, E. (1920). Psychological tests for selecting aviators. *Journal of Experimental Psychology*, *3*(6), 405–423.
*Timpany, N. (1947). Assessment for foremanship. *British Journal of Psychology*, *38*, 23–28.
Department of Labor. (1977). *Dictionary of occupational titles* (4th ed.). Washington, DC: US Government Printing Office.
*Vernon, P. E. (1946). Statistical methods in the selection of navy and army personnel. *Journal of the Royal Statistical Society*, *8*(2), 139–153.
*Vernon, P. E. (1947). Research on the selection in the Royal Navy and British Army. *American Psychologist*, *2*, 35–51.
Vernon, P. E. (1949). The structure of practical abilities. *Occupational Psychology*, *23*, 81–96.
*Vernon, P. E. (1950). The validation of civil service selection board procedures. *Occupational Psychology*, *24*, 75–95.
Vernon, P. E. (1956). The measurement of abilities, 2nd ed. London: University of London Press.
Vernon, P. E. (1961). *The structure of human abilities* (2nd ed.). London: Methuen.
*Vernon, P. E., & Parry, J. B. (1949). *Personnel selection in the British forces*. London: University of London Press.
Vernon, P. E. (1972). *Intelligence and cultural environment*. London: Methuen.
Vinchur, A. J., Schippmann, J. S., Switzer, F. S., III. & Roth, P. L. (1998). A meta-analytic review of predictors of job performance for salespeople. *Journal of Applied Psychology*, *83*, 586–597.
Viswesvaran, C., & Ones, D. S. (2002). Agreements and disagreements on the role of general mental ability in industrial, work and organizational psychology. *Human Performance*, *15*, 211–231.
Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574.
*Wickham, M. (1949). Follow-up of personnel selection in the A.T.S. *Occupational Psychology*, *23*, 153–168.

Articles included in the analyses are marked with an asterix in the reference list.

Received 28 February 2002; revised version received 9 July 2004

Appendix A. Ability tests included within each test type category

General mental ability (g) tests	Non-Verbal Intelligence, Intelligence, Group Test 70, NIIP Intelligence Group Test 33, NIIP Intelligence Group Test 70\23, Moray House General Intelligence Advanced (5), INS B (Instrument Comprehension), Intelligence Test, APU Abstractions Test, Raven's Progressive Matrices, Abstractions Test, Shipley's Abstractions Test (1), AH4, Intelligence selection test A, Penrose Pattern Perception, Kent Shakow Performance, Intelligence Test (I & II)
Verbal ability tests	Verbal Intelligence, English, Verbal test, ATS Spelling, Mill Hill Vocabulary Intelligence Test, SP17/25, PTB VP5, TTB VT1, Dictation Test 70, Verbal Test 25, Dictation Test 71, Reading Comprehension

Numerical ability tests	Maths, General Mathematics, Arithmetical approximation, Arithmetic reasoning, Calculations: Arithmetic, Arithmetic Test, ATS Arithmetic, Vernon's Arithmetic Test, PTB NP6, TTB NT2, Mathematical Test, Test 3a Arithmetic, SP3a, Mathematics and Arithmetic, Mathematics 3b
Perceptual-clerical ability tests	Minnesota Clerical Test, Instructions Test, Instructions SP21, CP4, CP73, Clerical Instructions Test (12 or 21), Clerical, Group Choice Reaction Time, Oral Directions, Judgment of speed, NIIP Group Test 25, Sale and graph reading (119), Oscilloscope Reading (118), Dial reading
Spatial-mechanical ability tests	Figure construction, Spatial Intelligence, NIIP Squares Test, Squares, Squares, Test 4, Spatial Relations Test, NIIP Forms Relations, Planning & Drawing Test, APU Mechanical Comprehension, DAT Mechanical Aptitude, DAT Space Relations Test, TTB MT4, TTB ET3, Bennett's Mechanical Comprehension, Memory for designs (97), Judgment of distances, Judgment of Ellipses

Note. NIIP = National Institute of Industrial Psychology; DAT = Differential Aptitude Tests; TTB = Technical Test Battery; ATS = Auxiliary Territorial Service; DAT = Differential Aptitude Test Battery; PTB = Personnel Test Battery.

Appendix B. Jobs included within each occupational category

Occupational categories for job performance ratings	
Clerical jobs	Administrative, general duties (ATS), foreign service
Engineers	Engineers and engineering apprentices
Professionals	Nurses and student nurses, surgeons and accountants
Drivers	Transport station staff and drivers (ATS)
Operators	ATS: telephonist, radar operators, special wireless, predictors, height takers, plotters, and spotters. military: anti aircraft and army infantry
Managers	Catering and banking managers. diverse managers and engineering supervisors
Sales	Sales staff and order takers. insurance claim assessor, financial consultants and people jobs
Miscellaneous	ATS: store women and orderlies. assembly workers, production line and working operatives. police (all ranks)
Occupational categories for training success	
Clerical jobs	Administrative. ATS general duties and auxiliary officers. military: clerical, writers and assistants
Engineers	Building and engineering apprentices
Professionals	Nurses, student nurses and dental students
Drivers	ATS drivers and military lorry drivers
Operators	ATS: height takers, spotters, & kine theodol operators. military, Asdic, radar and naval operators, naval coders. radar, wireless, teleprinter and switchboard operators, cipher operators, air traffic controllers
Skilled	Electricians (military and civilian). ATS: motor and general fitters, instrument and driver mechanics. military naval mechanics, radio and electrical mechanics, skilled workers
Miscellaneous	ATS: surveyors, cooks, draughts women, technical store women, and tinsmiths. military: clerks, store men, signal men, drivers, instrument and radio mechanics. protective military: pilots, gunners, riflemen, gunnery instructors, anti-air craft, anti-submarine. police (all ranks). unknown civilian occupations

Note. Unless otherwise stated all jobs are civilian jobs; ATS = Auxiliary territorial service.