Pearson
TalentLens

The Science Behind
Predicting Job Performance at Recruitment

Authors:
Wyn Davies, Global Product Manager, Pearson TalentLens
Angus McDonald, Chartered Psychologist

Date: May 2018

Introduction

Recruitment is a game of risk, where the chances of identifying a candidate who will turn out to be an effective employee are balanced against the risks of spotting someone who will not work out as expected. Recruiters use a wide variety of tools and processes to help them manage the risks associated with hiring decisions. If risks are effectively managed, recruiters are more likely than not to identify candidates who perform well in the job, though even under optimal conditions this process is far from perfect.

The complexity of understanding candidates and evaluating whether they really have the capabilities, characteristics and motivations necessary for any given job, means that incorrect decisions are easy to make. Factor into this issues such as deliberate deception on the part of the candidate and the difficulties for recruiters only increase. If recruiters get it wrong they risk costs including the need to quickly replace employees and the potential disruption that poorly-fitting employees may cause to the organisation.

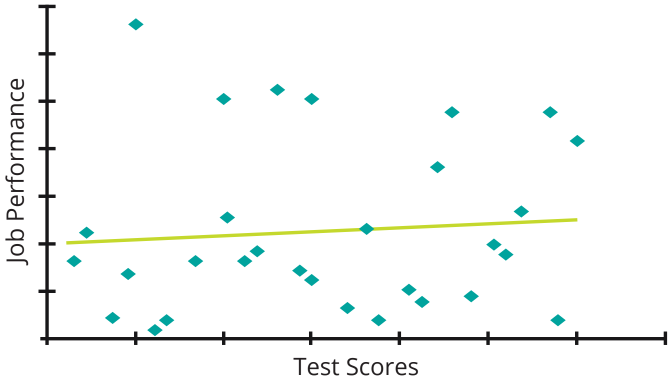
In response to this recruitment challenge, there is now over 100 years of scientific evidence on the effectiveness of different tools that can be used in the recruitment process. This evidence is based on the association between evaluations of candidates made during the recruitment process and how they subsequently perform in the role for which they are hired. In turn, this evidence should feed back into the hiring process, guiding recruiters as to which tools will most effectively allow them to manage risk.

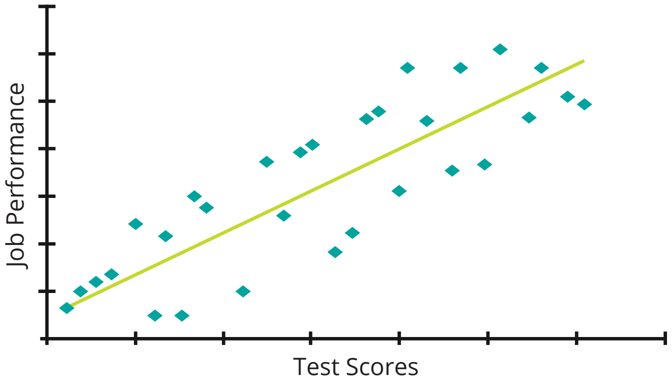
This paper summarises scientific research on the effectiveness of different recruitment tools, explains key concepts and identifies considerations to be taken when using recruitment tools. It is recognised that the processes and legislative frameworks recruiters work within will vary between organisations and regions, so the information given here will necessarily need to be applied in a way that is consistent with these local frameworks.

Correlation

Many different statistical techniques can be used to determine how effective any assessment tool is for candidate selection, but the majority of these are based on the idea of correlation. Correlation is a statistical technique that looks at the strength of association between two variables, typically the score or scores obtained on an assessment tool and some measure of job performance. The stronger the correlation, the more closely associated with each other the two measures are said to be, as illustrated in the two diagrams below.

The diagrams show the distribution of the scores of group of 30 people who completed an assessment. The graph on the left below shows that there is a lower association between scores from the assessment tool and job performance, whereas the one on the right shows a much stronger association. A strong association is indicated by the majority of data points falling on or close to a 'line of best fit'. The importance of this association is that it more accurately allows us to identify candidates likely to perform well in the job.



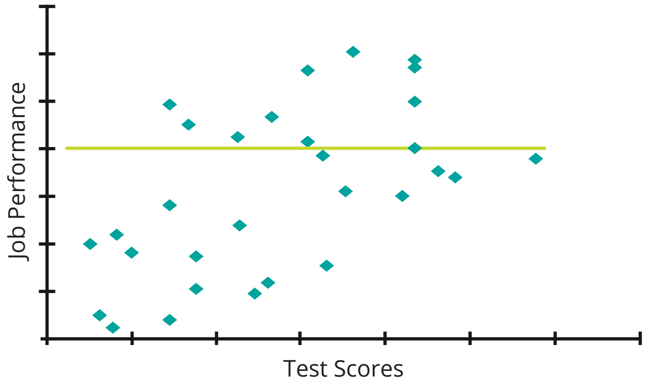


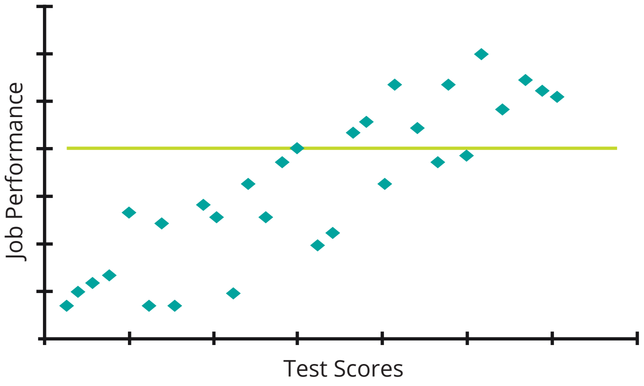
The Concept of Validity

In the context of recruitment, validity concerns the appropriateness of a specific assessment tool for a defined purpose. Tools with higher validity therefore enable recruiters to more accurately identify those candidates who are likely to go on to perform well in the job. In doing so they aid substantially in the management of risk that is an inevitable part of each hiring decision.

Having strong validity means that an assessment tool is able to predict job performance with a good degree of accuracy. By knowing this, it is possible to use assessment scores to identify those candidates most likely to perform well in the job. Consider the diagrams below which plot the scores of two different assessments against job performance for 30 candidates. The horizontal green line shows the ideal level of job performance, meaning we want to select those candidates whose performance will turn out to be above the green line.

In the diagram on the left, which shows a weak association between assessment scores and job performance - or low validity, it is difficult to know what score to use to identify good performers. No matter how high you set your pass mark or cut score on the assessment, illustrated on the horizontal axis, you will identify some people who go on to perform below the level expected. Conversely, low pass marks will also identify some good performers. The diagram on the right shows a stronger association between the assessment scores and subsequent job performance. This makes it much easier to set a pass mark that identifies where the majority of candidates will perform at the required level.





The key message is: having strong validity makes it easier to manage the risk associated with recruitment decisions, as assessment scores more accurately predict performance. It is worth noting that performance is affected by a number of factors and there is (as far as we know) no absolute measure or test that predicts success in the job with 100% certainty, You can, however, use a range of relevant and valid assessments to greatly increase the chances of predicting likely performance and reduce the risk of hiring the wrong person.

What Does Science Tell Us?

In practice, recruiters do not know how employees will go on to perform in a job. This makes selecting valid assessment tools and applying results from them appropriately at the point of hiring critical.

So, we know the principles behind validity and why it is important, but which assessment tools are most valid in terms of predicting likely performance in the job itself? There is extensive research on the use of different assessment tools and techniques. More recently this research has been synthesised to allow comparison of the effectiveness of different methods. One of the main ways this has been done is through a technique called meta-analysis, which is a statistical technique that draws together the findings. The results are presented as correlation co-efficients. A score of 1 indicates that the assessment method predicts performance perfectly. The higher the correlation coefficient, the more predictive of performance the assessment method is. A recent paper by Frank Schmidt and colleagues[1] summarises the effectiveness of 31 different assessment methods for the prediction of job performance and 16 for the prediction of performance in job-related training.

The key finding from this research was that tests of general mental ability (GMA) were identified as having the highest validity of any selection method. They were also seen to be the best predictor of performance on job-related training. The extensive research underpinning the use of GMA means a high level of confidence can be placed in these findings.

What is meta-analysis ?

For many years organisations and academic researchers have been keen to understand the association between assessments made as part of the recruitment process and subsequent aspects of job performance. Many thousands of studies exploring this link have been conducted, but with inconsistent results. There are many challenges in conducting this ‘real world’ research including being able to follow people through from recruitment to a time when job performance can be meaningfully measured, obtaining accurate measures of job performance and small sample sizes.

The last of these points is a particular issue when assessing the validity of recruitment processes. For research results to be robust, it is important that sufficient people are included. However, many organisations have only a limited number of people in each role, so collecting adequate samples takes a long time, if it is possible at all. Meta-analysis is a statistical technique that allows the findings from many individual studies to be combined. In doing so, issues such as measurement errors and small sample sizes are allowed for, meaning more robust estimates of the associations between different assessment methods and job performance can be established.

Meta-analyses have consistently identified that some types of assessment are far more predictive of job performance than others. In the absence of good local validation for a specific job role, these findings provide recruiters with a sound basis for identifying assessments that are likely to be the most effective.[v]

General mental Ability (GMA) Tests

As Schmidt and colleagues observe, when an employer uses GMA to select employees who will have a high level of performance on the job, that employer is also selecting those who will learn the most from job training programs and will acquire job knowledge faster from experience on the job. **This last point touches on the primary reason why candidates higher in GMA perform better in the job; they acquire job knowledge more rapidly and more deeply than those lower in GMA.**

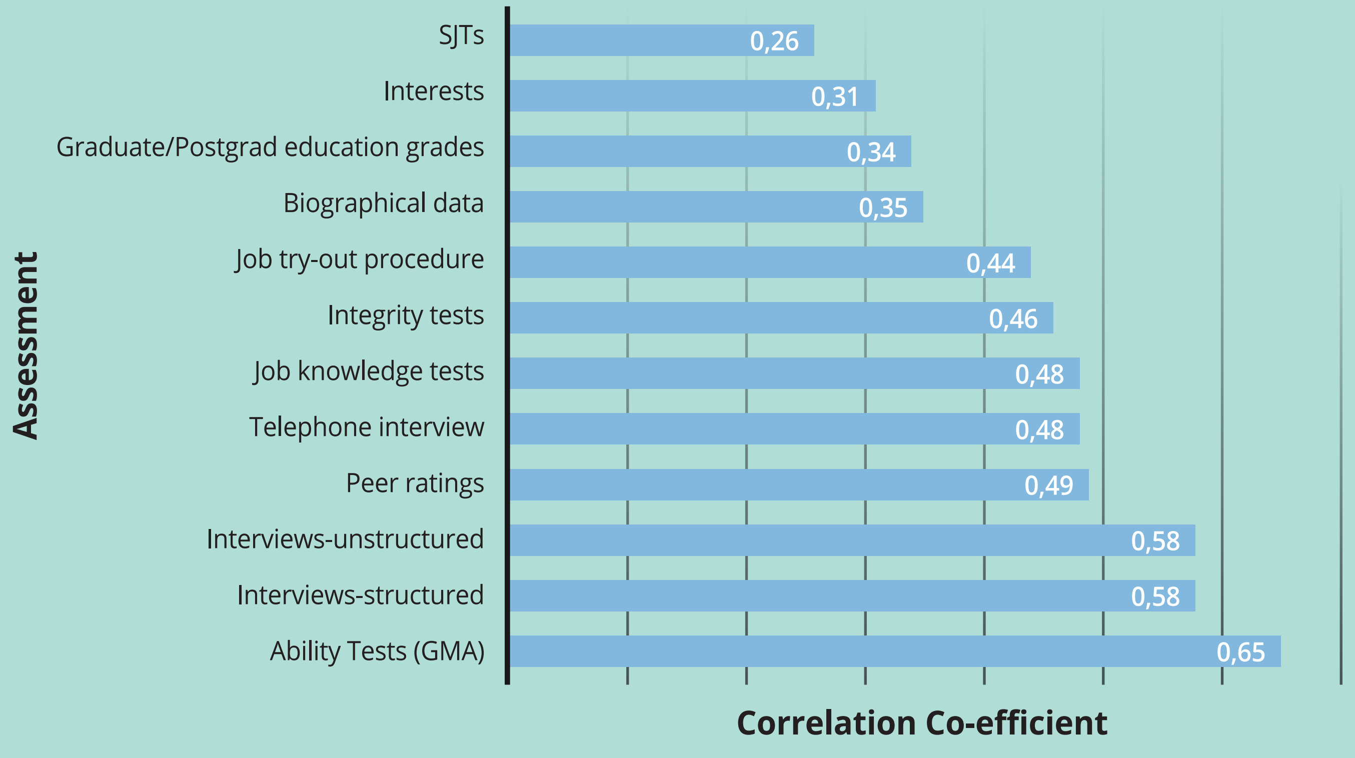
Many tests have been constructed to assess aspects of mental ability, and there are numerous examples where these have been developed primarily for use in recruitment settings. Examples include the Watson Glaser Critical Thinking Appraisal and Raven's Progressive Matrices. As versions of these tests can be delivered unsupervised online or administered to groups of applicants at a time under supervised conditions, they have a high level of utility in recruitment settings. They are relatively easy to use and interpret, do not take long to complete and are relatively cost-effective.

Another notable finding relating to GMA is the wide range of job roles for which it has validity. Typically these tests have been used in areas such as graduate recruitment, where jobs are assumed to place considerable cognitive demands on the job holders. Tests of GMA are known to predict effectively for graduate-level roles, but they also have considerable predictive validity for virtually any role, including unskilled roles. This ability of GMA to predict performance tends to increase as the cognitive demands of jobs increase. Highest validities are seen for professional and managerial roles, where GMA is able to account for over 50 per cent of the variability in job performance on average.

Going Deeper Into Validity

There are many recruitment tools apart from GMA and most selection processes use combinations of these. So how do other measures compare to the effectiveness of GMA ?

Job Predictiveness Correlation Co-efficients vs. Job Performance



It would be rare to be offered a job without some form of interview. Research supports the benefits of interview, as after tests of GMA they are one of the most effective selection methods. Interestingly, there is little evidence to suggest that structured interviews are more effective than unstructured interviews, however structured phone-based interviews do tend to come out as slightly less valid.

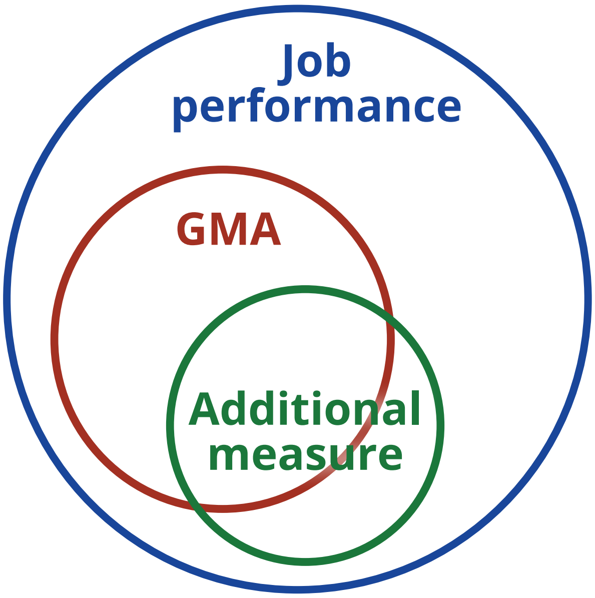
Other methods that show at least modest predictive validity in Schmidt and colleagues' research include integrity tests, biographical data, assessment centres, grade point average (i.e. academic results at graduate level), peer ratings, work sample tests, job tryout procedures, behavioural consistency method and job knowledge tests. However, not all of these methods can be readily applied in all recruitment settings. For example, job tryout procedures cannot easily be used in high-risk jobs or jobs where considerable training is required and peer ratings can only be used with internal candidates.

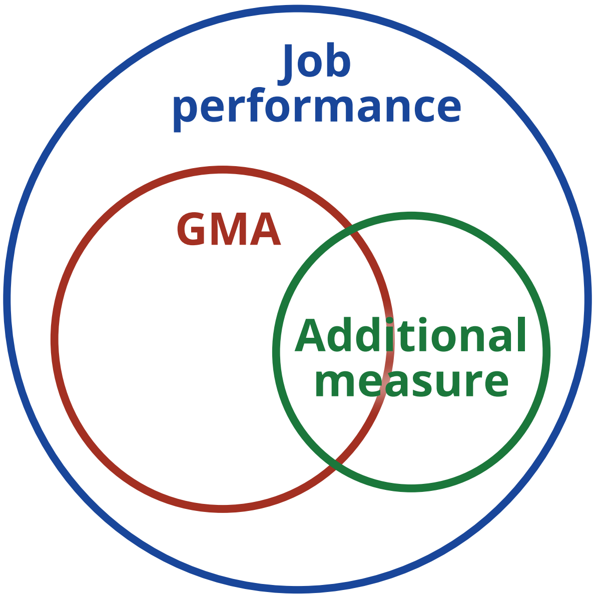
Even when other assessment methods can be used, it does not necessarily mean they will enable the risk of hiring decisions to be better managed. To determine this, we need to understand how additional measures provide 'incremental validity'. Whilst measures may be useful on their own, after the effects of GMA have been allowed for their impact may be much less. Schmidt and colleagues consider GMA to be so fundamental in the prediction of job performance, they evaluate the effectiveness of all other methods after allowing for GMA.

Incremental Validity

Tests of GMA are the single best predictor of job performance but many other methods can also be effective. Given this it might be tempting to use a number of high-validity assessment methods to reduce hiring risk. Unfortunately the situation is not that simple. The challenge is to understand how each method uniquely adds to our understanding of job performance - what is referred to as incremental validity.

Let's assume that the large blue circles below represent job performance. Assessments are used to understand about likely job performance, and we know that GMA accounts for about 50 per cent of this under optimal conditions. This is illustrated by the brown 'GMA' circle overlapping job performance. Now let's see what happens when we include an additional measure, shown by the green circle.





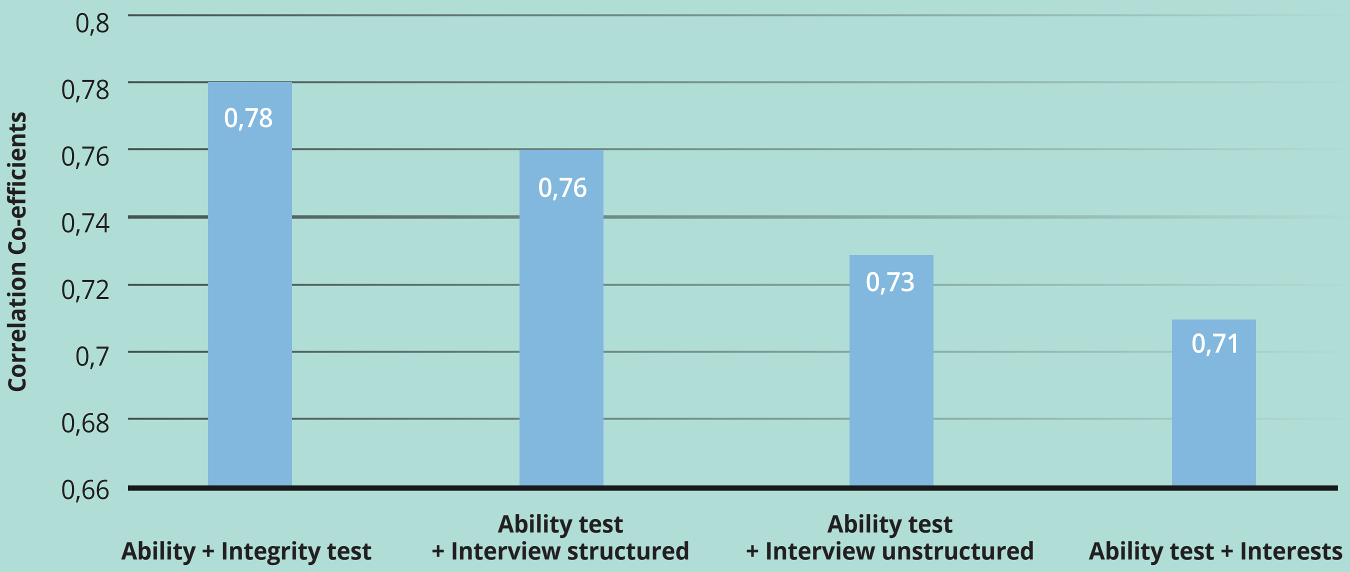
This green measure has a similar relationship with job performance to GMA, but in the case on the left it adds little to the overall understanding of job performance.

This is due to the degree of overlap, or correlation, between GMA and the additional measure being considerable. In the case on the right, the additional measure overlaps with GMA far less. In this case, using the additional measure alongside GMA does add substantially to our overall understanding of job performance.

The extent to which using the additional measure provides additional information on job performance is referred to as 'incremental validity'. Due to the large effect of GMA, it makes sense to understand what the incremental validity of other measures are, once GMA has been allowed for.

Once the predictive ability of GMA is allowed for, the pattern of results changes somewhat. The highest combined predictive validity comes from the use of GMA and an integrity test. This is closely followed by a combination of GMA and a structured interview and then GMA and an unstructured interview. The use of integrity tests adds approximately 20 per cent to the predictive ability of GMA alone, whereas structured and unstructured interviews add 18 and 13 per cent respectively. The only other measure to add 10 per cent or more to the predictive power of GMA are assessments of career interests.

Predictiveness of Combined Assessment Methods

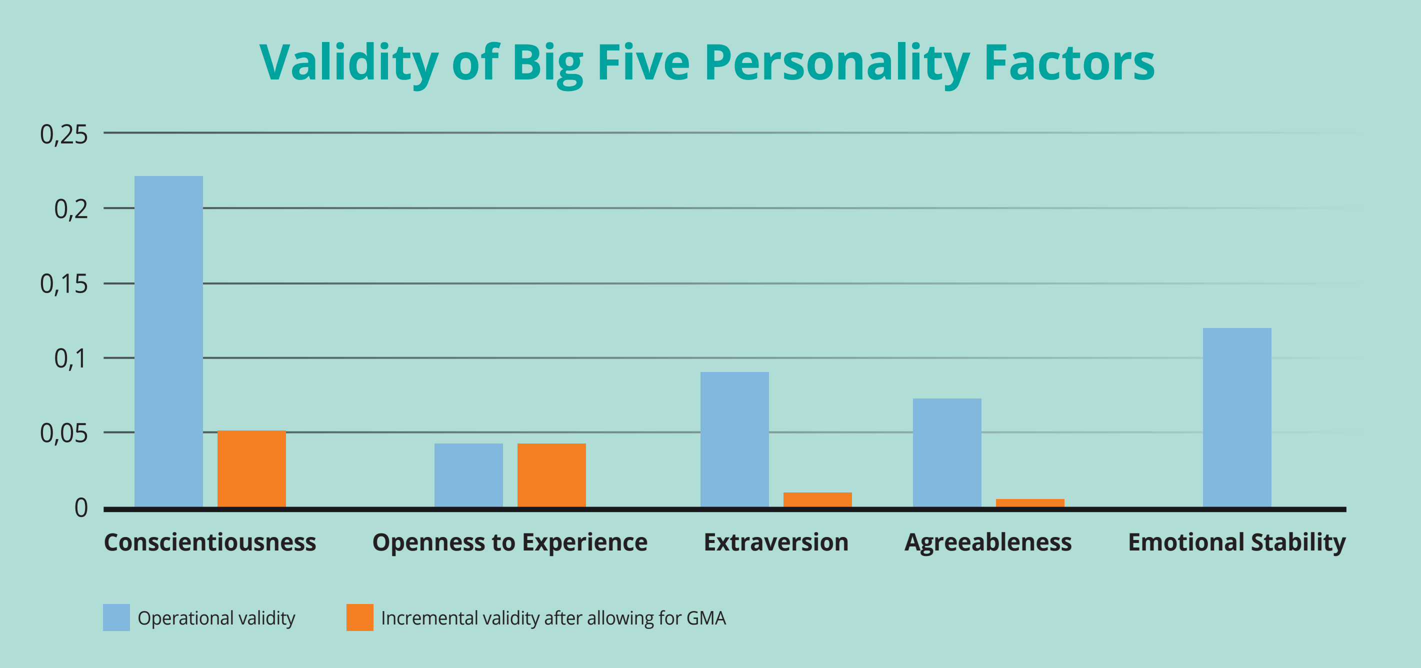


Integrity Tests

As integrity tests add the most incremental validity to GMA, it is worth looking at these in more detail. Two types of integrity tests can be identified: those that ask directly about dishonesty ('overt' or 'direct' assessments), and those that are more personality-based ('indirect' assessments). Direct tests of integrity may ask about attitudes to dishonesty, beliefs about the prevalence of dishonesty and, in some cases, ask respondents directly about past dishonest behaviour. In contrast, indirect assessments attempt to identify aspects of a person's character that may underlie dishonest behaviours. Though personality-based assessments may be more distal to actual behaviour, they have the advantage of being less transparent and therefore open to deliberate distortion of responses. Both have been associated with behavioural outcomes in the workplace but direct measures, despite their apparent openness to faking, have been found to have slightly higher validity than indirect measures[2].

Personality Measures

Finally, as personality measures are widely used in recruitment, and have already been referred to in relation to integrity, it is worth considering the validity of such instruments in relation to job performance. As shown in the graph below, although the Big Five factors of Conscientiousness and Emotional Stability are moderately predictive, the effects of most personality factors are almost zero once GMA has been allowed for. The exceptions to this are Conscientiousness, which as noted above is a significant part of many integrity measures, and to a slightly lesser extent, Openness to Experience.



The Role of Job Analysis

Predictive validity research is a good guide of how helpful different types of measure may be in the recruitment process, but deciding exactly what combination of assessments should be used relies on a clear understanding of job requirements. Without this, recruiters risk blindly applying assessments on the basis of data without a broader consideration of how they relate to specific job requirements.

Job analysis details the requirements and tasks associated with a specific role, and informs the job description and person specification. A clear person specification identifies what aspects of candidates need to be assessed and also provides a structure to the assessment process. A structured recruitment process has clearly defined stages with each stage providing successive information on the candidates. On the basis of this information, candidates may be rejected or progressed to the next stage.

The task for recruiters is therefore to identify appropriate measures and when they are best placed in the recruitment process. For example, if a job requires specific qualifications or for the job holder to have a licence to practise or similar, this should be assessed very early on. There's no point conducting other assessments only to find later on in the recruitment process that a candidate does not have the required credentials.

Qualifications are an example of essential requirements, with the exception of where organisations are willing to support candidates in working towards these. A general principle is to assess essential requirements before those that are desirable. Such essential requirements may include qualifications or eligibility to work, but may also involve assessments such as tests of GMA where minimum requirements for these have been established. As many tests of GMA are available online for unsupervised administration, this means that they can be used very early on in the selection process.

Technology allows an increasing number of measures to be administered to candidates remotely, including tests of GMA, personality assessments, work samples and even interviews. Some of the most valid assessment methods can therefore be used without having any direct contact with candidates. Though this may be efficient in terms of processing applicants, organisations concerned with candidate attraction may not choose this approach. Instead they want to have early personal contact with candidates as a way of engaging them with the organisation. However, tests of GMA are generally well-perceived by candidates[3,] despite the claims made by developers of more game-like assessments.

In summary, there is no single best way to design a recruitment process. Each process should be tailored to the specific needs of the organisations, but key considerations should be: 1) what does the job analysis tell us is needed for successful performance, 2) how do we assess the essential and desirable characteristics for the role in the most valid ways, 3) how are we going to structure the recruitment process and use information gathered to successively filter candidates and 4) how and when do we want to engage candidates with the organisation through the recruitment process.

Group Differences and Bias

There can be an assumption that all assessments should measure equally for all groups. Any group differences that are seen must therefore indicate that the assessment tool is in some way biased or is not measuring fairly for all groups.

Unlike physical attributes such as length or weight that can be directly measured, we do not have direct access to the constructs measured by psychometric assessments. Instead, assessments ask respondents to solve problems, report on their behaviour, make judgements or similar, and from the information they provide we make an inference about the degree to which the respondents possess different constructs. In this process it is quite possible that the assessments are biased in some way - historically there are certainly plenty of examples of this - but modern test development techniques have done much to manage this source of bias.

An alternative interpretation of group difference data is that it reflects real differences between the groups being assessed. There may be a tendency to believe that all people are equal and differences must be the result of biased measurement. However, differences exist in the physical world, so why not the psychological? For example, just because a measuring tape shows that men are taller on average than women, does not mean that it is biased. This difference is accepted as it is visible and we can directly measure height, unlike many psychological characteristics that have to be inferred from other observations.

Despite best practice in test development, some group differences remain. An authoritative review of this area was conducted under the guidance of the American Psychological Association[4]. It concluded that whilst sex differences in performance on tests of GMA was generally seen to be small, this was not the case for ethnic group differences. Whilst various reasons had been offered to account for these differences, no adequate reason for them could be identified.

Diversity and Adverse Impact

Diversity is a key issue for many organisations, as they strive to address the uneven representation of different groups in their workforces. Many aspects of the recruitment process have been cited in diversity debates at various times. The judgements made by recruiters are one of the areas that have received much attention, with research highlighting numerous forms of bias that may affect recruitment decisions based on information from application forms, resumes, interviews or similar techniques.

Psychometric assessments have also long been cited in diversity discussions. The nature of the data produced from such assessments means that the extent to which they provide an objective assessment of different groups is relatively easy to study. As with validity research, many studies have been conducted into group differences in psychometric assessment scores and synthesis of these results has not always been straightforward. Whilst, there is general agreement that tests of cognitive ability show some differences, especially when results are compared across different ethnic groups, they also predict job performance equally for different groups[5]. This means that two people scoring the same on a test of GMA will show approximately the same level of job performance regardless of their group membership[6].

Adverse impact is where a substantially smaller proportion of people from a minority or 'protected' group are successful compared to the proportion from the majority group who are successful. In the US, protected classes include race, gender, those aged 40 and above, religion and disability among others. The 'four-fifths' or '80%' rule is commonly used to determine whether adverse impact has occurred. Adverse impact is not restricted to recruitment; it can be introduced into other activities such as training and development, job redundancies and even performance appraisals. The four fifths rule has legal standing in the United States but not in many other countries, though it is a useful starting point in any territory for monitoring selection activities or other employee decisions (e.g. promotion, allocation of development opportunities etc.).

Addressing Adverse Impact

The ability of GMA tests to predict job performance means their use is defensible from a legal perspective in most situations, but the results obtained by different groups can have the effect of reducing diversity in the workforce. This is clearly not desirable as organisations wish to ensure that their workforce is rich and diverse in make-up and experiences.

One practical impact of group differences is that adverse impact tends to become greater as cut scores are set higher. For GMA tests, in particular, setting a low cut score - for example no higher than the 25th or 30th percentile may help to minimise adverse impact.

Adverse impact can be a complex area and we advise that organisations consult psychologists and other experts in this field to find out more about the ways to moderate and minimise it. Psychometric test publishers strongly advise that GMA tests are not the sole source of assessment data upon which to base a hiring decision. When used with other relevant assessments, however, their high ability to predict job performance can be of great value to recruiters.

Other forms of assessment such as interviews, integrity tests and situational judgement tests tend to show far less evidence of group differences. Though they may be less predictive than tests of GMA, employers increasingly see them as desirable as they maintain diversity in the applicant pool. Such assessments may be used as an alternative to GMA or in combination with it. When used in combination, GMA may be applied at a separate point from other assessments or alongside them, where results from all assessments can be combined to produce a weighted result which will have less adverse impact. Such strategies can have a modest benefit in reducing the impact of GMA on diversity, but do not offer a complete solution[7].

Predicting for More than Just Performance

For much of the history of research into selection methods, the key outcome has been job performance or productivity. More recently employers have shown interest in additional areas that have not always been thoroughly understood.

Employee engagement has become a key area of interest for many employers. Employees who are engaged are energised by their work and committed to it and the organisation. Engagement is also associated with a range of positive workplace behaviours and reduced turnover. Gallup has been tracking employee engagement in the US, typically finding that only about one-third of the workforce reports being engaged with their jobs. Worldwide, this figure is much lower[8]. Disengaged employees are not only likely to have poorer performance but they can also have a negative effect on colleagues, meaning productivity issues can spread beyond the individual hire.

Organisations increasingly seek to understand more about candidates' potential fit to the job role, a key determinant of engagement. In this context, fit is not about having the qualifications, skills or knowledge to do the job, or even the specific behaviours someone typically displays, but whether the person's values, attitudes and needs are aligned with those of the organisation. Most often this is assessed at an organisational level (e.g. "does this person share the values of the wider organisation?"), but they may also be at a more macro level, such as the work unit or specific team.

This focus on fit represents a shift from looking primarily at the 'can do' to considering the 'will do'. Schmidt and colleagues' work concludes that person-organisation fit measures have an overall low predictive validity, but what exactly constitutes 'fit' and what outcomes it affects are unclear. Whilst fit may be weakly related to performance, it does show stronger relationships with turnover intentions and positive work attitudes[9]. Fit is most often assessed by looking at the match between an individual's values and those of the organisation, but questions remain about how best to understand values at an organisational level and how to relate these to individual candidate's needs and expectations.

Values is another area that has become of increasing interest to organisations. As well as being involved in fit, organisations seek to be more 'values conscious' and look at individual's strengths. Prominent strengths-based approaches focus on values[10], arguing that individual strengths result from the identification and application of values. In recruitment contexts, strengths-based approaches tend to focus on each individual applicant's unique pattern of strengths and how these might be utilised within the role and organisation more broadly. These tend not to replace more established selection methods but are used in addition to them. Organisations cite a number of reasons for using strengths-based approaches including finding better fit, creating organisational advantage and increasing diversity. It is notable that values and strength-based approaches were not included in Schmidt and colleagues' research, presumably due to research in these areas still developing.

Limitations of What We Know

Much of this paper has been based around Schmidt and colleagues' recent paper on the validity of different assessment methods. This research is probably the most comprehensive of its kind, but it is not without limitations. It is important that assessment users are aware of these limitations if they are to design assessment processes with due considerations as to the limits of our scientific knowledge.

A significant limitation of all meta-analytic research is that the details of specific studies - such as the organisation it was conducted in, job roles assessed for, sample size etc - are often lost. Whilst the best meta-analytic studies may evaluate the quality of the research they use and weight results accordingly, this is not always the case. Whilst local validation, where a link is established between assessments and job performance in a specific organisation, is always preferable, opportunities for this are restricted to organisations employing and recruiting relatively large number of people into similar roles. It also requires that employees are given sufficient time to establish themselves in the role, meaning assessment of job performance may not take place until 12 months or so after recruitment.

A further consideration relates to how assessment results were applied and at what point in the selection process. It has already been noted that recruitment is a process involving various stages of information gathering and decision-making. This process is not adequately modelled in research that takes a more two-dimensional approach to validity; a snapshot of the association between scores on a measurement instrument and job performance. Data may be collected early in the recruitment process or later on, possibly after some candidates have already been rejected. In yet other cases data may be collected on people who are already working in a role, meaning no data are available on those already rejected. Decisions about when to use different assessment methods are critical in managing risk, but research often fails to give sufficient detail to adequately inform these decisions.

Finally, validity is measured against 'job performance'. What constitutes effective job performance varies considerably between roles and organisation, and defining performance in a way that allows it to be readily and accurately measured is a challenge. Often job performance relies simply on managerial ratings, which can be influenced by multiple sources of bias. As discussed above, other outcomes of interest to employers are yet to be explored in the same depth as 'job performance'.

Though it is important to be mindful of these limitations, research consistently shows that the sensitive application of valid measures substantially reduces risks associated with recruitment and results in considerable gains in workforce productivity. Recruiters can be confident in generalising from validity research for the majority of job roles, but need to apply this sensitively and with due regard for the specific situation to gain maximum returns.

References

1 Schmidt, F. L. (2016) The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical implications of 100 Years. Retrieved on 16 October 2017 from: www.researchgate.net/publication/309203898

2 Van Iddekinge, C. H., Roth, P. L., Raymark, P. H. and Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. Journal of Applied Psychology, 97, 499-530.

3 Anderson, N., Salgado, J. F. and Hülsheger, U. R. (2010). Applicant reactions in selection: Comprehensive meta- analysis into reaction generalization versus situational specificity. International Journal of Selection and Assessment, 18, 291-304.

4 Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R.; Sternberg, R. J. and Urbina, S. (1996). Intelligence: Knowns and Unknowns. American Psychologist, 51, 77–101.

5 Schmidt, F. L. and Hunter, J. E. (1981). Employment testing: Old theories and new research findings. American Psychologist, 36, 1128–1137.

6 Rothstein, H . R and McDaniel, M. A. (1992). Differential validity by sex in employment settings. Journal of Business and Psychology, 7, 45-62.

7 Sackett, P. R., Schmitt, N., Ellingson, J. E. and Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. American Psychologist, 56, 302-318.

8 Gallup (2013). State of the Global Workplace: Employee Engagement Insights for Business Leaders Worldwide. Washington DC, Gallup.

9 Arthur, W. Jr., Bell, S. T., Villado, A. J. and Doverspike, D. (2006). The use of person-organization fit in employment decision making: An assessment of its criterion-related validity. Journal of Applied Psychology, 91, 786-801.

10 Peterson, C. and Seligman, M. E. P. (2004). Character strengths and virtues: A handbook and classification. New York: Oxford University Press.

Pearson | TalentLens