IQ? EQ? 4-Q?

What Every HR Professional Should Know
About Hiring Assessments

By: **Whitney Martin**, MS
Measurement Strategist
ProActive Consulting

Research on the utility of using valid selection systems leaves little doubt that getting the right people into the right organizations and the right jobs can make a big difference. Popular business publications have delivered a similar message regarding selection — in the best seller Good to Great, for example, Collins (2001) wrote about the importance of "getting the right people on the bus."

— **yet most practitioners still aren't aware of some of the most important findings from selection research.** (Rynes, Giluk, & Brown, 2007, p. 1001)

About the Author

As a measurement strategist, Whitney's passion and expertise lies in the field of surveys and assessments. A self-professed "data nerd," Whitney has a Master's Degree in the area of Human Resources Measurement and Evaluation and conducted extensive research on the predictive validity of various hiring assessment strategies.

Whitney specializes in delivering *data-supported insights* into job candidates, employees, teams, leaders, customers, and organizations. Using valid and reliable assessment and survey tools, she provides practical, actionable information to leaders that informs business decisions and has a measurable impact on the organization's bottom line.

She resides in Louisville, KY with her husband and two daughters.

Ten years ago, researchers discovered something that should have opened eyes, raised red flags, and rocked the HR community at large. It was found that practitioners were largely unaware of best practices surrounding one of the linchpin functions of HR — the ability to effectively screen for and hire strong employees. In an economy where companies are trying to streamline and cut costs, HR has a tremendous opportunity to have a measurable financial impact on the bottom line and further secure their seat at the strategic table. However, it seems that few have seized the opportunity to familiarize themselves with prevailing research and embrace recommended hiring strategies.

In 2002, Rynes, Colbert, & Brown conducted a study to *determine the extent to which the beliefs of HR professionals are consistent with established research findings* regarding the effectiveness of various HR practices. They surveyed 1,000 SHRM members — HR Managers, Directors, VPs, and SVPs with an average of 14 years' experience. The questionnaire consisted of 35 True-False items that pertained to all the dimensions on the PHR exam (excluding Safety and Labor Relations which are not as generalist-relevant).

The Result? The area of greatest disconnect was in *Staffing* — especially hiring assessments, where *fewer than 50%* of respondents were familiar with prevailing research findings. The relevant questions, including the desired answer and percentage of practitioners who got that item correct, are listed in the following table.

Survey Items from Rynes et al. (2002) Study	Answer	% correct
There are 4 basic personality dimensions, like in MBTI	False	49%
Being intelligent is a disadvantage in low-skilled jobs	False	42%
There's little difference between personality assessments in their ability to predict job performance	False	42%
Integrity tests are not very effective in practice because so many people lie on them	False	32%
Integrity tests have high degrees of adverse impact	False	31%
Conscientiousness is a better predictor of job performance than intelligence	False	18%
Companies that screen for values have better performance than those that screen for intelligence	False	16%

The Research-Practice Gap

Several studies have endeavored to explore why selection research findings have seemingly failed to transfer to HR practitioners. Among the more prevalent causes of this phenomenon are the following:

- ***HR professionals are so overburdened*** that they do not have time to read the latest research.

- ***Research journals are too technically complex***, especially in the field of selection research, which deals heavily in the domain of validity, reliability, sampling procedures, etc.

- ***What academics find interesting is not what practitioners find interesting***. Several studies have analyzed the differences in topics represented in scholarly vs. practitioner literature and found a big disconnect. HR practitioners express a strong interest in topics that are relevant to the day-to-day nature of their jobs, whereas researchers tend to be more interested in construct refinement and theory building.

- ***Practitioners hear and read conflicting things***, even from sources they trust, which creates a general sense of confusion and uncertainty in this area of selection practices. As an example, *HR Magazine* is widely documented as being the publication most frequently read by HR practitioners. The content of *HR Magazine* has been analyzed to determine what is being published regarding hiring assessments. Not only do articles on this important topic tend to be infrequent (only 10 articles appeared between 2000 and 2011 on the use of personality or intelligence tests in hiring), but they prove to represent a mix of research – based and non-research-based (or even research-contradictory) claims, leaving it up to readers to draw their own conclusions about what content is sound advice.

- ***The idea of implementing a new screening measure is daunting***, to say the least. Even if HR managers were confident about what constructs they *should* be measuring, they would still need to decide what tool(s) to measure with, adjust HR processes accordingly, assess administrative demands, get organizational buy in, cost justify, weigh pros and cons, consider potential legal exposure, etc., all while maintaining their current job. It is certainly understandable why many practitioners have opted to continue using methods they're comfortable and familiar with rather than consider options that may be more effective but that would require a significant up-front investment of time and resources.

Implementing a new screening measure can be a daunting task, but may be well worth the investment.

Why Are Effective Selection Methods Important?

If it were the case that all job candidates would perform equally well in a given position, there would be no need for selection devices of any kind. However, because of the wide variation in worker performance levels across job types, it is critical for organizations to understand what differences among individuals systematically affect job performance so that the candidates with the greatest probability of success can be selected. Gallup did a meta-analysis of 25 years' worth of data and concluded: "*People don't change that much…great managers know there is a limit to how much remolding they can do to someone.*" This is why it is crucial to know how to select the *right* people in the first place.

Yet, this critical HR function — effective hiring, which has a huge bottom line impact — is the area where HR practitioners have the most uncertainty around best practices and prevailing research.

What ARE the Most Predictive Screening Measures?

Organizations that choose to rely on less predictive selection methods are unnecessarily creating a competitive disadvantage for themselves. Extensive research has been done on the predictive validity of different hiring methods and measures. The table below reports the relative validity of some of the most commonly used selection practices (sorted from least effective to most effective):

Graphology (Handwriting Analysis)	.02
Unstructured Interviews	.14
Reference Checks	.26
Personality Tests	.20-.31
Integrity Tests	.41
Cognitive Ability Tests	.51
Multi-Measure Tests (i.e.: Cognitive Ability + Personality + Interests)	.71

The assessment approaches with the highest predictive validity are Integrity Tests and General Mental Ability Tests. Multi-Measure tools that combine several measures may yield the highest predictive power.

Source: Based on numerous studies, including *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, 262-274.*

This means that if your hiring process relies primarily on interviews, reference checks, and even personality tests, you are choosing to use a process that is significantly less effective than it could be if more effective measures were incorporated. The following sections provide an overview of research findings relative to the four kinds of assessment instruments outlined in the above table.

Personality Tests

Personality Tests are perhaps the assessment type most familiar to HR professionals. There is consensus among researchers that there are five basic personality traits or factors. In other words, if you take all the personality traits or constructs you can think of and do a statistical factor analysis, they are going to sort themselves into one of these five categories — *Emotional Stability, Extraversion, Openness to Experience, Agreeableness,* and *Conscientiousness*.

Researchers do not agree, however, on whether a given personality trait could be considered "universally desirable" or whether its attractiveness depends on the context (the type of job, company culture, etc.). For example, some studies have found *Conscientiousness* to be universally desirable, or a universal predictor of job performance (i.e.: conscientiousness would be an asset in any position). Others (including this author) ascribe to a contextual or "fit" model, meaning that no personality trait is inherently *good* or *bad*, but its desirability depends on the job (for example, while highly conscientious people may excel in self-discipline and attention to detail, they may also be considered perfectionists who lack the necessary degree of flexibility to adapt to rapidly changing circumstances).

Four Quadrant ("4-Q") Personality Assessments

The most widely used type of personality tests are *Four Quadrant* tests (any personality assessment where the results classify you as some combination of four different options labeled as letters, numbers, colors, animals, etc.). These types of assessments originated in 440 BC when Empedocles noticed that he could group people's behavior into four categories which he labeled earth, wind, fire, and air. In 444 BC, Hippocrates made the same observation, but labeled the categories blood, water, black bile, and yellow bile. Since then, hundreds of iterations of these tools have been developed, all essentially based on the same premise and theory. Some basic attributes of these types of tools are outlined below:

- They are most accurately described as measuring *Style* (i.e.: tendencies and preferences to go about completing a task in a certain manner)

- They most often consist of a list of adjectives from which respondents are asked to select the words that most/least accurately describe them. Therefore, they are fairly *transparent* by nature, and one is unable to determine how accurately the assessment results portray the test taker.

- They measure *State* (as opposed to *Trait*), meaning that they measure one's style *within a context* (you might have a different style at home vs. at work, or as an employee vs. as a manager). This


EARTH	WIND
FIRE	AIR

4-Quadrant "Style" tools are well suited for many organizational applications, but have several shortcomings when used for employee selection.

means that they have limited test-retest reliability because they are designed to be sensitive to the test taker's context or *state*.

- They are Ipsative measures (as opposed to Normative). Ipsative means you are measuring something in relation to yourself (for example, I prefer data entry more than I prefer parties. A normative tool, by contrast, would tell you whether **I** prefer data entry *more or less than **you*** prefer data entry). An Ipsative measure is confined to an individual, whereas a normative tool allows comparisons between individuals.

Because of the above attributes of 4-Q assessments, they have certain strengths and weaknesses, and therefore certain applications to which they are better suited. They are ideal for self-discovery, team building, coaching, enhancing communication, and numerous other developmental applications. However, due to limited predictive validity and test-retest reliability, the lack of norming, the lack of an internal consistency (*lie detector*) measure, etc., they are not ideal for *high stakes* situations such as hiring. The following table compares the attributes and ideal uses of 4-Q tools vs. Multi-Measure tools (described in a later section).
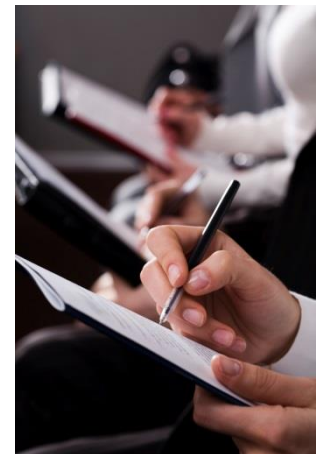
Due to factors such as limited predictive validity and low test-retest reliability, 4-Q assessments are not ideal for high stakes situations (such as hiring).

4-Quadrant Personality Assessments	"Multi-Measure" Assessments
Measure *Style* (tendencies and preferences)	Take more of a *whole person* approach
Measure *State* (within a context)	Measure more stable *Traits*
Ipsative	Normative
Best Uses: Self-awareness, Team Dynamics, Communication, Coaching	*Best Uses:* Hiring, Placement, Promotion, On-Boarding, Talent Pool Management, Self-awareness, Team Dynamics, Enhancing Manager/Employee Relations

As a point of clarification, there *are* personality assessments on the market that measure stable traits, that are normative, that have *lie detector* scales and high reliability and validity. However, the above section focused on the strengths, weaknesses, and best uses of 4-quadrant personality tools specifically due to their widespread use (and perhaps overuse or misuse in some cases). Regardless, personality constructs are not the most predictive measure available, and some of the most popular personality tests really should not be used in the hiring process due to low reliability and validity, among other factors.

Emotional Intelligence

There has been a surge of interest among organizations in Emotional Intelligence (commonly referred to as **EQ** for Emotional Intelligence Quotient). As of September, 2008, there were at least 57 EQ consulting firms, 90 EQ training and assessment organizations, 30 EQ certification programs, and five EQ "universities." However, there is actually very little empirical research that demonstrates that EQ is related to important organizational outcomes. There are two distinctly different models being used in the marketplace, both purportedly measuring the construct of Emotional Intelligence—one based on personality and one based on General Mental Ability. Many studies have concluded that EQ is most likely a repackaged version of other measures (combinations of personality factors or intelligence factors) that have been around for a long time. Emotional Intelligence is yet another area where there appears to be a gap between what practitioners believe (as evidenced by the widespread use of EQ measures) and what science has been able to support conclusively with data.

Integrity Tests

Integrity is a general term that typically encompasses traits such as honesty, dependability, trustworthiness, reliability, work ethic, substance abuse propensity, etc. The main goal of integrity tests is to measure the likelihood of theft or other counterproductive workplace behaviors, though they have also been shown to positively predict overall job performance.

There was a surge of interest in paper-and-pencil integrity testing after the Employee Polygraph Protection Act of 1988 prohibited employers from using pre-employment polygraphs in most industries. There are no legal restrictions on the use of integrity tests in the United States (except for in Massachusetts and Rhode Island), and they have been found not to discriminate or cause adverse impact against any protected group as defined by Title VII of the Civil Rights Act.

An *Integrity Test* is an appropriate selection tool when an organization wants to address issues with 90-day turnover, drug use, theft, absenteeism, work ethic, reliability, etc. These tools are also easy to tie to metrics such as turnover, worker's compensation claims, etc. in order to assess the impact of the tool and determine return-on-investment.

General Mental Ability (GMA)

Thousands of studies over nine decades support the conclusion that GMA is absolutely the best predictor of job performance. In fact, it is common to see the effectiveness of other measures (i.e.: personality, integrity, etc.) reported in terms of what *incremental validity,* or additional predictive power, they provide above and beyond GMA. In other words, the attitude is: "*Obviously you're measuring mental ability since it's the most effective predictor of job performance… now what's*
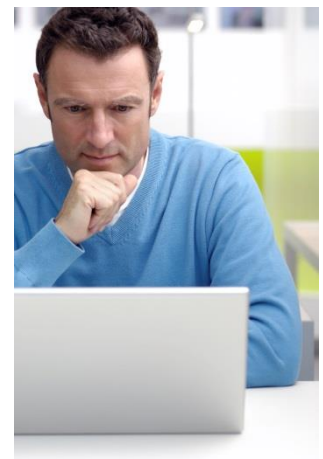
Thousands of studies over nine decades support the conclusion that GMA is the single best predictor of job performance.

the benefit of measuring other things in addition to that?" One reason that GMA is an effective predictor of job performance is that higher GMA allows a person to acquire job knowledge faster, which translates into better performance, both initially and throughout an employee's tenure in a position.

It is likely that some HR practitioners have shied away from using cognitive ability measures due to concerns over adverse impact. This is a valid concern considering that some types of intelligence are affected by factors such as socio-economic status and educational advantages. However, it is interesting to note that unstructured interviews have also been found to result in adverse impact against women, older applicants, and persons with disabilities, yet this screening method is still widely used and has only a small fraction of the predictive power of GMA. Like any screening method, the key is to show job-relevance which can be accomplished through a concurrent validation study (which will be discussed in the *Your Challenge* section).

Multi-Measure Tools

Multi-Measure tools are designed to measure multiple constructs within one instrument. If you take GMA and add the incremental validity gained by also assessing personality and interests, as an example, you'd get a predictive validity coefficient of .71. Typically, these types of tools are designed to measure more stable characteristics that do not tend to change over time (short of drastic circumstances like battling a life-threatening illness, going to prison, the death of a child, etc.). Measuring stable *traits* as opposed to *states* is preferable in hiring because the results are more likely to be an accurate reflection of the candidate's capabilities and approach over time.

Multi-Measure tools also tend to be Normative in nature, which means you can compare one person's scores against someone else's (i.e.: two job candidates) to determine which candidate possess *more of* a desirable trait. Extensive research goes into the norming process so that the statistical probabilities of certain patterns of answers or the expected prevalence of certain traits within a population can be determined. Because of this process, these tools are less transparent in nature than a 4-Q assessment, are more difficult to fake, and have built-in distortion or *lie detection* measures. Many of these tools are also job-fit driven, meaning the desired level of any given trait depends on the job or position for which the individual is being considered.

It may be tempting to think that a Multi-Measure tool would automatically be the "best" tool to use in any hiring context due to the fact that it has the highest predictive validity, or the best chance of helping you systematically hire the best employees most often. However, it is important to first consider the goals of the organization to

*Measuring stable **traits** as opposed to **states** is preferable in hiring because the results are more likely to be an accurate reflection of the candidate's capabilities and approach over time.*

determine what assessment approach will be most appropriate. This process will be further discussed in the *Your Challenge* section.

Criteria for Selecting a Selection Assessment

Attempting to sort through the thousands of assessment tools on the market is an extremely daunting task. However, there are some resources and guidelines that can make it easier. Once you have identified what you are trying to accomplish (see *Your Challenge* section for more details) and what type of measure will help achieve your goals, there are some criteria that any selection tool you use should meet or exceed. The Department of Labor has published an excellent resource titled "Testing and Assessment: An Employers' Guide to Good Practices" that is available through their website. Among the most foundational and most important things to look for in a pre-employment assessment are reliability and validity.
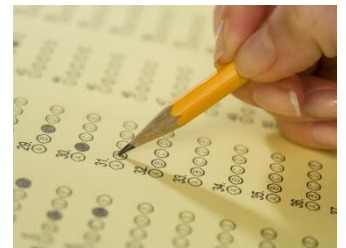
Reliability

Reliability refers to the consistency with which a tool measures something. There are several kinds of reliability coefficients that a test publisher should report. The most widely used measure is internal consistency, or Cronbach's Alpha. Another important reliability measure, especially in pre-employment testing, is Test-Retest Reliability. Essentially, this means if someone takes the test and then re-takes it an hour/day/week/month/year later, how consistent will the results be? As discussed earlier, 4-Q *style* assessments have low test-reliability by nature because they are sensitive to emotion, mood, and circumstance. Reliability is expressed as a decimal ranging from 0 to 1, with larger numbers representing higher reliability. The Department of Labor (DOL) recommends practitioners use instruments with a minimum reliability of .7. Well-developed tools designed to measure more stable traits should exceed this standard.

Reliability and Validity are among the most important criteria to look for when selecting a hiring assessment. Test publishers should be forthcoming with data (most often found in a technical manual) to show how rigorous they were in developing and validating their instrument.

Criterion-Related Validity

If an instrument has criterion-related validity, that means that the assessment results are statistically correlated with some outcome of interest. For example, as test scores go up, job performance goes up. Or as test scores go up, turnover goes down. In other words, you should be able to demonstrate that how someone scores on the test correlates with some outcome (job performance, turnover, sales volume, etc.). Criterion-Related validity allows integration with organizational metrics, allows you to calculate return-on-investment, etc. Not all tools are designed to function this way. For example, 4-Q tools are designed to *describe* how a person will go about doing a job (their *style*), not to predict *whether or not* they will be effective at it.

Predictive Validity

Validity coefficients are also expressed as a decimal between 0 and 1 with larger numbers indicating that the tool has a greater degree of

validity for the purpose it is being used. As was described in the beginning of this paper, some assessment approaches have higher Predictive Validity than others. For example, graphology (handwriting analysis) has only a .02 predictive validity because differences in handwriting have been found to be caused by genetic muscular differences in people's hands rather than differences in personality. Therefore, they are not accurate predictors of job performance. The assessment approaches with the highest validity in predicting job performance are Integrity Tests (.41) and General Mental Ability (.51). These numbers can be further elevated by combining measures, as was described in the *Multi-Measure Tools* section.

Additional Tips

First and foremost, in order to determine the reliability and validity numbers associated with a tool, you must be provided a *technical manual* from the test publisher. The publisher should be forthcoming with this information and should be able to provide robust and detailed information about how their tool was developed and validated. You should pay careful attention to how the instrument was validated. For example, many validation studies take place on college campuses, where students may be incentivized to participate and where they may or may not be very similar to the job candidates you will end up assessing with the tool. Knowing how rigorous the company was in validating the tool is critical.

Your Challenge

As those in the HR profession continue to discuss how to become more strategic and more aligned with the business objectives of their organizations, perhaps the area where they can have the greatest impact on the company's bottom line is in *systematically* selecting employees who have what it takes to become top producers for the organization. Taking the following steps will enable HR practitioners to accomplish this through the successful creation of a strategic, evidence-based hiring process that has high predictive validity.
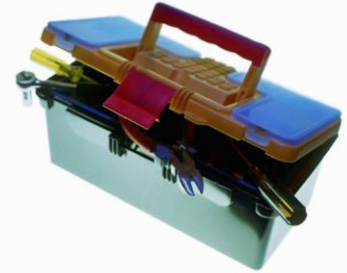
Take steps to create a strategic, evidence-based hiring process that has high predictive validity.

1) **Increase your knowledge** regarding the predictive validity of different hiring methods. Reading this paper is a great step towards achieving this goal. Further knowledge can be gained through college courses, business books, tutoring from academics or consultants, etc. Validity should be among the most important considerations for HR practitioners when choosing a selection tool or method, but research suggests that it actually plays a relatively small role. Ensure that any vendor or consultant you are working with is steering you toward one of the assessment methods with high predictive validity, and that they are enthusiastic about sharing technical manuals and other

supporting documentation that attests to the effectiveness of their tool(s).

2) **Decide what you are trying to accomplish**. Whether you are trying to address turnover, retention, sales volume, customer satisfaction, morale, productivity, theft, absenteeism, safety incidents, or drug use in the workplace, there are different assessment approaches/instruments designed specifically to address these issues or goals. There is little point in using an assessment instrument merely because "it's interesting" or "seems to be pretty accurate" (no matter how cheap or easy it is to use). Know what you are trying to accomplish and know what success looks like. A good assessment tool should make a *measurable* impact.

Whether it is job performance, turnover, customer satisfaction, sales volume, absenteeism, or theft, there are different assessment instruments specifically designed to address each of these (and many more) organizational challenges.

3) **Know how to select a GOOD tool**. Examine the evidence provided by the test publisher. If necessary, have a third party academic or consultant review and critique this information with you. Get the aforementioned Department of Labor guide so you know what to look for, especially related to reliability and validity.

4) **Do a concurrent validation study** of the tool you are using or considering to see whether the test results are directly correlated with the metric or outcome you are interested in (i.e.: job performance, turnover, sales, customer satisfaction, theft, absenteeism, etc.). ***See Case Study below***.

Case Study

One example of what a concurrent validation study might entail is as follows.

I was working with a healthcare organization that was trying to better understand a particular "trouble position." The HR team decided to conduct a job analysis. First, they ranked their 15 incumbents according to objective measures of job performance and determined that 6 were "above average" performers while the remainder were "average" or "below average." They then assessed all 15 incumbents using the Profile XT Assessment, which measures five cognitive abilities, nine behavioral traits, and six occupational interests.

With this data, they were able to create a "performance model" or "benchmark" for the position that was a clear differentiator. When comparing all incumbents' scores against the benchmark (using 85% job match or above as the standard), they would have been able to *correctly identify five out of six of their top performers* and *correctly screen out eight out of nine of their average or below average performers*. This concurrent validation study allowed them to measure the effectiveness of the tool in differentiating between candidates that would likely become top performers for their organization and ones that would have been poor or mediocre performers.

Conclusion

All assessment approaches are not created equal. Knowing which types of assessments will be most effective in accomplishing the specific objectives you have identified for your organization will enable you to select a tool with a measurable impact on the bottom line.

Summary

- The assessment approaches with the highest predictive validity are Integrity Tests and General Mental Ability Tests. Multi-Measure tools that combine several measures may yield the highest predictive power.

- 4-Quadrant "Style" tools are well suited for many organizational applications, but have several shortcomings when used for employee selection.

- Reliability and Validity are among the most important criteria to look for when selecting a hiring assessment. Test publishers should be forthcoming with data (most often found in a technical manual) to show how rigorous they were in developing and validating their instrument.

- It is critical to first identify what you are trying to accomplish. Whether it is job performance, turnover, customer satisfaction, sales volume, absenteeism, or theft, there are different assessment instrument specifically designed to address each of these (and many more) organizational challenges.

- Know what success looks like and how to measure the impact that an assessment tool is having. With this information, you can quantify the impact that your expertise in selecting a hiring tool has on the organization's bottom line!

Whitney Martin partners with companies to identify the most effective assessment instruments for specific organizational objectives. Contact Whitney for assistance in selecting tools and processes that will have a measurable impact on the organization's bottom line.

Primary References

Rynes, S. L., Colbert, A. E., & Brown, K. G. (2002). HR professionals' beliefs about effective human resource practices: Correspondence between research and practice. *Human Resource Management*, *41*, 149-174.

Rynes, S. L., Giluk, T. L., & Brown, K. G. (2007). The very separate worlds of academic and practitioner periodicals in human resource management: Implications for evidence-based management. *Academy of Management Journal*, *50*, 987-1008.

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*, 262-274.

*Additional 80+ references used in the initial research study on which this article is based are available by emailing a request to **whitney@consultproactive.com**.*

Acknowledgement