

2006

---



# State-of-the-art Measurement in Human Resource Assessment

James Houran, President, 20 20 Skills™ Employee Assessment  
Rense Lange, President, Integrated Knowledge Systems

## **HVS INTERNATIONAL NEW YORK**

372 Willis Ave.

Mineola, NY 11501

+1 516.248.8828 (ph)

+1 516.742.3059 (fax)

---

**Abstract.** – The use of assessments in employee selection - and corporate spending on these services - continues to rise. Traditionally, HR assessments are constructed and validated using Classical Test Theory (CTT). Unfortunately, CTT neither provides state-of-the-art measurement nor guarantees to meet legal requirements, since it does not yield interval-level and bias free results. Accordingly, the use of such assessments can lead to improper candidate selection. Most of the technical challenges to CTT can be addressed using modern, i.e., Item Response Theory based, approaches. We illustrate the power and potential of modern test approaches in a hypothetical testing example derived from HVS International’s 20 20 Skills™ Assessment v2.0. We show that modern test approaches provide superior measurement over CTT and that modern approaches capture crucial and highly specific candidate information that is otherwise unattainable. Thus, true state-of-the-art measurement not only reliably and validly measures trait levels of core competencies, but it also permits HR professionals to prepare evidence-based action plans for behavioral interviews and consultations with professional references.

---

## Introduction

“Google” the phrase “human resource assessment” and approximately 14,000,000 entries appear. Automated and standardized employee screening and selection is apparently a widespread practice considering that there are no less than 2,000 businesses which offer Human Resource (HR) assessment instruments for offline or online application. HR assessments provide cost effective and time efficient tools that can be used throughout the recruiting process to identify interviewees with skill sets required for specific positions and who have a personality profile that is compatible with a particular corporate culture. Furthermore, assessment instruments can be administered periodically to existing employees to serve as an ongoing training tool to improve the performance and retention of the employees and in the setting of their performance goals.

With the potential value that HR assessments hold for businesses, it is crucial that HR professionals become more discerning about the scientific validity of the assessment instrument on the market. For instance, Houran (2004) recently noted that some claims of scientific validity for online tests and measurements are difficult to evaluate because psychometric data are either not collected or rarely made available for public scrutiny – including peer review in the scientific literature. Finn and Banach (2000) similarly discussed the difficulties of ascertaining the credentials and identity of service providers, accessing accurate information, reliance on untested methods, difficulties in online assessment, and the lack of standards and regulation regarding online testing practices. This opinion was also echoed and expanded in a high profile article recently published in the *American Psychologist* (Naglieri, Drasgow, Schmit, Handler, Prifitera, Margolis, & Velasquez, 2004).

The present paper outlines the concept of *state-of-the-art measurement* in human resource assessment. It is our position that the scientific validity of any assessment begins with the fundamental issue of quality of measurement. This issue is first emphasized in a technical comparison of the traditional Classical Test Theory (CTT) approach in industrial-organizational psychology versus the features and benefits of modern test and measurement theories as embodied in Item Response Theory (IRT). Next, we illustrate the power and potential of IRT in HR assessment by showing how this approach can extract valuable information on employment candidates that CTT approaches inherently miss. Finally, we discuss the issue of measurement quality in relation to other issues that assessment vendors typically use to differentiate and promote their products. It is shown that measurement quality must be addressed first and foremost of any other issue and that the superior, state-of-the-art measurement offered by IRT has tremendous value for tackling a variety of other HR and business related research questions.

## Background

Kline (1986, 2000) noted that researchers traditionally construct and validate assessment instruments using factor analysis and reliance on the KR-20 or Coefficient Alpha as the major index of the scale's quality. Often, and despite early warnings against this practice (Comrey, 1978), item-level factor analysis is used to identify distinct "clusters" or subscales of items. Next, the (sometimes weighted) scores on the items within such clusters are added and these sums are then assumed to yield valid indices of different latent traits like Service Orientation or Leadership.

Unfortunately, a CTT raw-score approach essentially treats all test items within a cluster as equivalent, thereby ignoring the possibility that some items are more diagnostic of being high (or low) on the particular construct than are other items (Bond & Fox, 2001). Not only is this very questionable practice, but – as will be shown later – it also causes the loss of potentially very useful information. Most importantly, by focusing on the number of "points" respondent earn on the test it is easy to lose sight of the fact that it is their *trait levels* that we are really after. Unfortunately, summed scores (even weighted ones) do *not* provide linear (i.e., interval-level) measures of the underlying latent traits – and this is *especially* true for the groups of greatest potential interest (i.e., high scoring achievers). Moreover, standard CTT approaches typically do not recognize that some items may be biased such that respondents with *identical trait levels* may receive systematically *different* scores. This might be the case, for instance, when women (or younger test takers) endorse some questions more (or less) often than do men (or older test takers) with *equal* trait levels.

Thus, while HR assessment providers almost universally use traditional CTT-based scaling, and present their tests as satisfying the quality standards set forth by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement, 2002) and the *Uniform Guidelines for Employee Selection Procedures* (Equal Employment Opportunity Commission, Department of Labor, Department of Justice, & the Civil Service

Commission, 1978), there is in fact no guarantee that CTT-based approaches indeed yield valid measures of test takers' trait levels.

### **IRT and Rasch Scaling: the State-of-the-Art in Assessment**

Embretson (1999) noted that although most textbooks continue to emphasize CTT approaches, professionals in the tests and measurement field have rejected these approaches in favor of methods derived from modern test theory – notably Item Response Theory (IRT) and its variants like Rasch scaling (Rasch, 1960/1980; cf. Wright & Stone, 1979; Bond & Fox, 2001). In virtually all instances of large-scale testing where results have real-world implications for those being tested, the use of CTT has been abandoned. The quality of HR assessment services, or for that matter any other type of online testing, is only as good as the quality of their testing methodology. We expect therefore that it is only a matter of time before the advantages of IRT based methods begin to be reflected in the practices of all serious online testing services, just as they have been adopted in other professional testing domains.

In this context, it seems appropriate to introduce the differences between CTT and IRT by citing some of the “rules” provided by Embretson (1995; 1999, p. 12):

1. The standard error of measurement differs between persons with different response patterns but generalizes across populations.
2. Shorter tests can be more reliable than longer tests.
3. Comparing test forms across multiple forms is optimal when test difficulty levels vary across persons.
4. Unbiased estimates of item properties may be obtained from unrepresentative samples.

In other words, the notion that all test scores are equally reliable has been abandoned in favor of local (i.e., score-specific) standard errors of estimate (*SE*). Thus, no longer is there a single index of reliability. Also, contrary to common wisdom, longer tests are not necessarily “better,” as – depending on the variation in the trait levels of the test takers – many questions are almost guaranteed to be redundant in the trait level they actually measure. Rather, by using items that best address test takers' different trait levels (i.e., by purposely using *non-parallel* forms) greater measurement precision can be obtained. In the extreme case, items are *selected* specifically to optimize reliability (minimize *SE*). When this is done in an interactive, computerized fashion one speaks of Computer Adaptive Testing, or CAT (Wainer, 2000). The savings achieved in the number of items needed when using CAT methods typically approaches 50% (Lange, in press). Further, given an appropriate item-pool, *smaller SEs* can be achieved using CAT than with fixed-length tests. Even greater savings may obtain when the main objective is to classify respondents into a small number of mutually exclusive categories (Eggen, 2004; Lange, in press).

Rasch (1960/1980) scaling, which is probably the most widely used IRT related method, is especially useful in providing extensive indices of model fit for items as well as persons. It also covers a wide variety of data types including binary items, rating scales, Poisson counts, percentages, and paired comparisons (Linacre, 2004). While item and person fit is certainly important, misfit is not a sufficient reason for rejecting the Rasch model. Rather, it should be understood that Rasch scaling provides a measurement model that specifies the conditions under which observations indeed constitute trait measures. Accordingly, misfit should be construed as a property of the data, rather than the model. As Bond and Fox (2001) explained, “the goal is to create abstractions that transcend the raw data, just as in the physical sciences, so that inferences can be made about constructs rather than mere descriptions about raw data” (p. 3). Researchers are then in a position to formulate initial theories, validate the consequences of theories on real data, refine theories in light of empirical data, and follow up with revised experimentation in a dialectic process that forms the essence of scientific discovery. Conversely, misfit can be exploited in advanced areas of research where theory is sufficiently powerful to predict particular deviations from the Rasch model (Bond, 1995; Lange, Greyson, & Houran, 2004; Lange, Jerabek et al., 2004; Larsen, Lange, & Hughes, 2006).

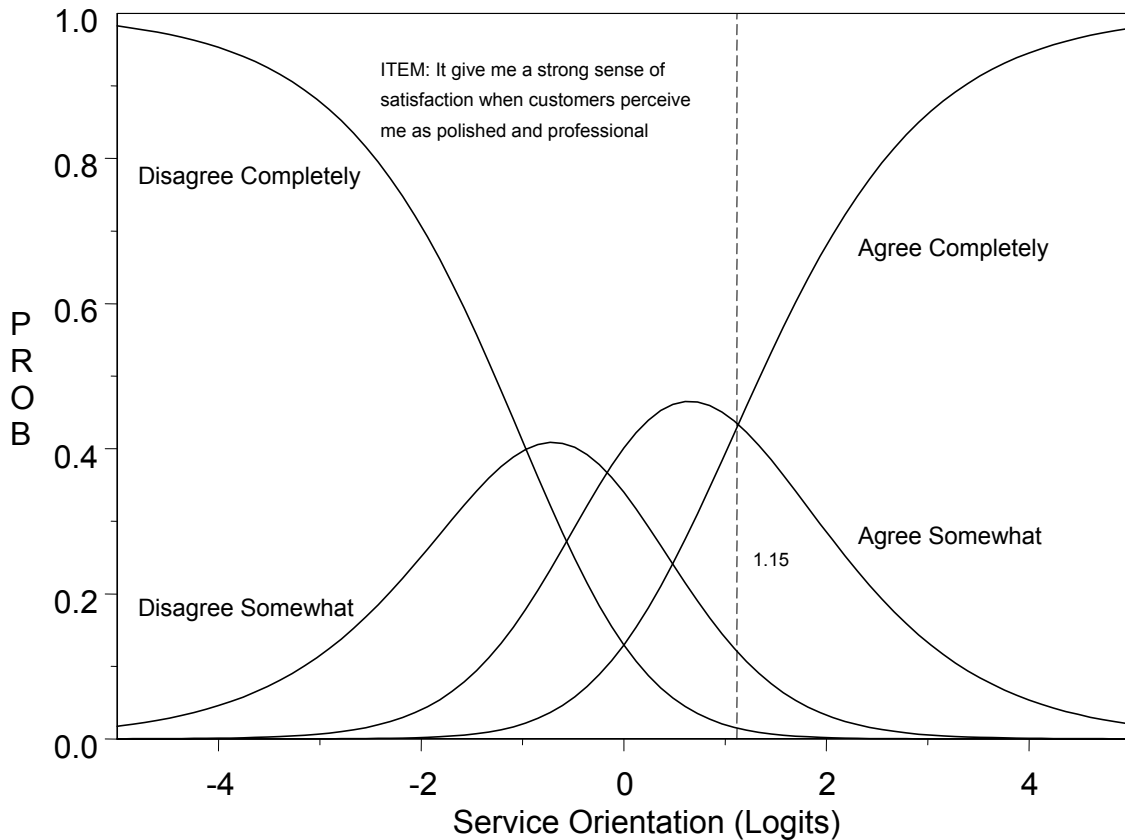
Lack of model fit is most problematic if people receive systematically different scores based *not* on corresponding trait differences, but rather on a group specific interpretation or understanding of the questions being asked. Such biases may reflect culture-related differences in expressing one’s feelings, opinions, or symptoms (Lange, Thalbourne, Houran, & Lester, 2002), or they may be related to respondents’ age or gender (Lange, Thalbourne, Houran, & Storm, 2000). Of course, it is only to be expected that tests translated into other languages may not yield measures equivalent to those obtained in the base language (van de Vijver, & Poortinga, 1997). This is especially pertinent in HR contexts where a global organization assumes that an employee assessment measures all of its cross-cultural applicants equally well. Again, IRT methods are helpful in cases where relatively few items are affected by the translation, as it may be possible to recalibrate just these items.

Similar issues play a role when paper-and-pencil tests are adapted for online use. For instance, web versions almost certainly will have different layouts and response contingencies than do paper-and-pencil tests. As well, when older tests are used, the meanings of its wordings may have changed over time (e.g., consider how the vernacular versions of “cool” and “hot” have fluctuated over the last decades). Finally, differences in the method of administration (i.e., offline vs. online) may systematically affect respondents’ reactions to the questions. It is mandatory, therefore, that web adaptations of paper-and-pencil instruments be re-administered and recalibrated based online. Moreover, the exact same item format and layout should be used during pilot testing and operational use. Finally, online tests should *not* rely on norms that were established by paper-and-pencil methods (Naglieri et al., 2004).

Along with a desire to avoid legal issues related to unsubstantiated claims of reliability, validity, and efficacy, there are good practical reasons for switching from CTT to IRT, especially since online testing typically produces sufficiently large datasets to expand the

approach into many new directions. For instance, the flexibility afforded by web-based item administration easily allows for the introduction and calibration of new questions or question types, *without* losing continuity and without the need to re-compute baselines or previously established cutoff scores. Finally, as is discussed in detail below, the availability of a variety of person fit statistics allows outlying (i.e., aberrant) respondents and responses to be identified (Wright & Stone, 1979).

**Figure 1:** Probabilities of category responses given the level of Service Orientation.



### Learning from Test Takers' Distinctive Response Patterns with IRT

IRT methods explicitly model test takers' answers as a probabilistic process involving these respondents' own trait levels, and the trait levels implied by the questions. Such questions may be binary, i.e., requiring True/False type answers, or rating scales with ordered categories like *Disagree Completely*, *Disagree Somewhat*, *Agree Somewhat*, and *Agree Completely*. Figure 1 illustrates the behavior of such a rating scale by plotting the probability of observing one four ordered responses along the Y-axis *given* a person's trait level (X-axis).<sup>1</sup> This example is based on one of the Service Orientation items on

<sup>1</sup> Curves such as shown in Figure 1 are governed by a number of parameters that must be estimated from empirical data. Specialized software exists for this purpose, examples of which can be found at web pages <http://www.rasch.org/look.htm> and <http://www.assess.com/firmSoftCat.htm>.

HVS International's 20 20 Skills™ assessment v2.0 (“*It gives me a strong sense of satisfaction when customers perceive me as polished and professional*”). It can be seen that test takers with the lowest Service Orientation trait levels most likely answer *Disagree Completely*. However, with increasing Service Orientation trait levels *Disagree Somewhat*, and then *Agree Somewhat* and *Agree Completely* become the most likely answers.

Although each of these four possible ratings may occur at any trait level of Service Orientation, Figure 1 shows that their probabilities may be exceedingly small. For instance, for a Service Orientation trait level of 1.15 the answer *Disagree Completely* will occur rarely as it has just a 2% chance of being selected, whereas *Agree Somewhat*, and *Agree Completely* each occur 43% of the time. If we assume that the ordinal ratings are scored 0, 1, 2, and 3, respectively, it can be derived (cf. Wright & Masters, 1982) that the average rating for respondents with Service Orientation trait levels of 1.15 is 2.26 and the standard deviation of such ratings should be about 0.54.<sup>2</sup> Knowing such conditional (i.e., trait level specific) mean and standard deviations allows us to quantify the extent to which a person's actual responses *deviate* from what is to be expected by chance alone. In particular, observing the rating *Disagree Completely* (with value 0) given by someone at 1.15 yields a z-score of  $(0 - 2.26) / 0.54 = -4.18$  which clearly ( $p < .001$ ) constitutes a statistical aberration. For this person, the z-scores for the answers *Disagree Somewhat*, *Agree Somewhat*, and *Disagree Completely* ratings are -2.33, -0.48, and 1.37, respectively.

Note that the above does not simply tell us that a particular rating is low or high in some population or sample of interest. Rather, it identifies low or high ratings relative to what is to be expected *given a particular individual's trait level*. Thus, the same rating might be too low or too high, depending on the test taker's trait level. For instance, the above showed that if someone with a Service Orientation trait level of 1.15 gives the rating *Disagree Completely* to the statement “*It gives me a strong sense of satisfaction when customers perceive me as polished and professional*” we may conclude that this rating is unexpectedly low – whereas the rating *Agree Completely* would be appropriate. By contrast, for someone with a Service Orientation trait level of, say, -2, the rating *Disagree Completely* is to be expected – but, the rating *Agree Completely* would be considered as unexpectedly high.

Of course, as is the case for most HR assessments, individuals' scores can also be compared to particular norm groups and populations. However, by repeating the above analysis for every answer in a test taker's record, we can now also derive an in-depth profile of an individual's particular views on issues related to the variable under consideration. This information can be used in two major ways. *Firstly*, it serves to differentiate between respondents with very similar scores on a particular trait or variable. For instance, whereas before we might have to conclude that applicants A and B are indistinguishable because they have the same total scores, we can now investigate their idiosyncratic differences in very great detail. *Secondly*, the finding of outlying answers provides vital information to HR professionals and managers that can be used effectively during interviews or follow-

---

<sup>2</sup> Note that these values are *conditional upon the trait level*. Hence, they bear no relation to the mean and standard deviations or the shape of the distribution of the trait level in the population.

up with professional references of a candidate.

Based on earlier research, these two uses have been implemented in the feedback reports of the 20 20 Skills™ assessment v2.0, and some detailed examples are provided in the following section. We note here that 20 20 Skills™ is a web-based assessment specifically designed for identifying peak performers in the service-hospitality industry by evaluating candidates on ten, research-based core competencies (cf. Lange & Houran, 2006): Applied & Personal Problem Solving, Creativity, Ethical Awareness, Group Process & Team Building, Leadership, Loyalty, Personal Efficacy & Motivation, Sense of Humor, Sensitivity to Diversity, and Service Orientation. It is also used for internal benchmarking, establishing specific performance goals for coaching and training, and documenting aspects of job performance for periodic evaluations.

### **20 20 Skills™ Assessment Feedback: the Example of Service Orientation**

The graphical report in Figure 2 indicates the overall score (90) of a hypothetical test taker (“Maria S.”) on the “Service Orientation” subscale by an arrow, together with an indication of the reliability of this score in terms of its standard error of estimate (yellow band). As is indicated on the right, this respondent’s percentile rank can immediately be derived relative to a number of client selectable populations (e.g., the score of 90 falls inside the top 10% of the scores obtained in the test calibration sample).

The report additionally shows how the candidate scored in IRT terms on each individual test item or question that comprises the Service Orientation subscale<sup>3</sup>. The sample candidate scored high overall, thereby suggesting a high trait level of Service Orientation. However, item level analysis also reveals three instances in which the candidate responded too weakly (denoted by “-”) or too strongly (denoted by “+” or “+ +”) to specific test items given her overall trait level of Service Orientation. Test items that the candidate endorsed within a normal range are denoted as “√.” Thus, these three instances are of potential concern and require further fact finding to determine their meaning and significance. To that end, the feedback report for the 20 20 Skills™ assessment v2.0 automatically generates sets of questions and issues to address with candidates in subsequent behavioral interviews or to explore with the candidate’s professional references (see Figure 3). The qualitative interpretations of the item level analyses depicted in Figure 2 were created in consultation with an expert panel consisting of experienced HR professionals, as well as clinical and social psychologists.

---

<sup>3</sup> Due to space restrictions, Figure 2 only provides a *representative* sample of items from the actual Service Orientation subscale on the 20 20 Skills™ assessment v2.0.

**Figure 2:** Graph from the 20 20 Skills™ assessment v2.0 report showing item level IRT analysis for Service Orientation test items.

20 20 Skills v2.0™  
**REPORT**

Assessment Completed 3/13/2006  
Mandarin Oriental Hotels  
**Maria S. (#4947546)**

Middle Level Position

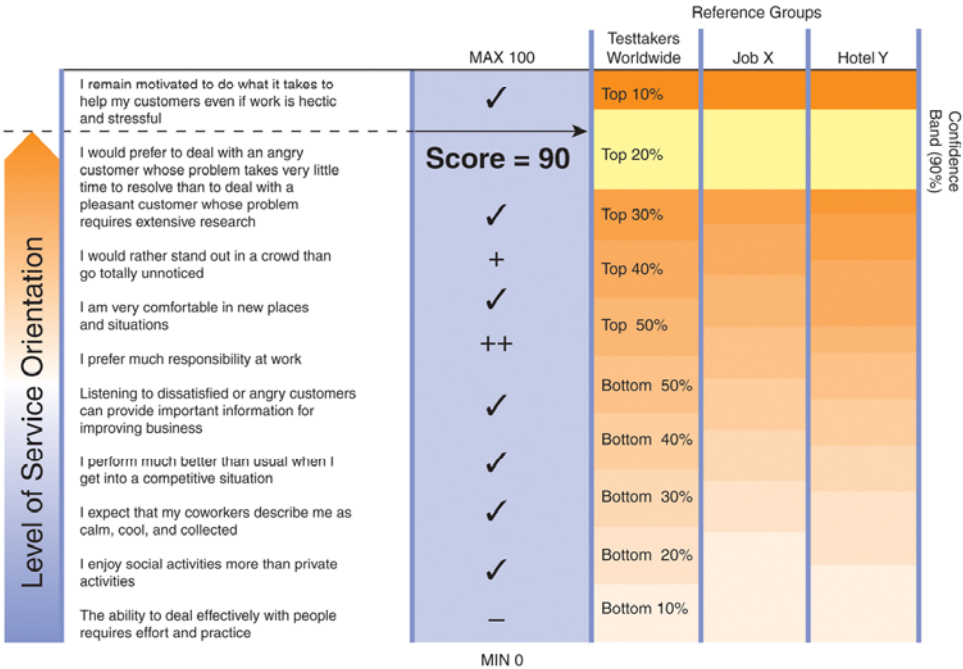
**Dimension 10:  
Service Orientation**

**Synopsis**

Service Orientation refers to a set of attitudes and behaviors that affect the quality of the interaction between the staff of an organization and its customers.

- As is indicated by the horizontal yellow confidence band, Maria's score of 90 (out of 100) on this competency indicates a high level of Service Orientation.
- Customers experience high Service-Oriented employees as cooperative, responsive, personable, and considerate. High service-oriented employees have a strong emotional IQ and are good at interacting with others. As a result, they can be reliable ambassadors for representing your organization to customers, for fulfilling consumer needs and expectations, and for producing a positive tone and mood with clients and even coworkers.
- Maria is at a competence level that meets or exceeds the median standards of the selected reference groups.

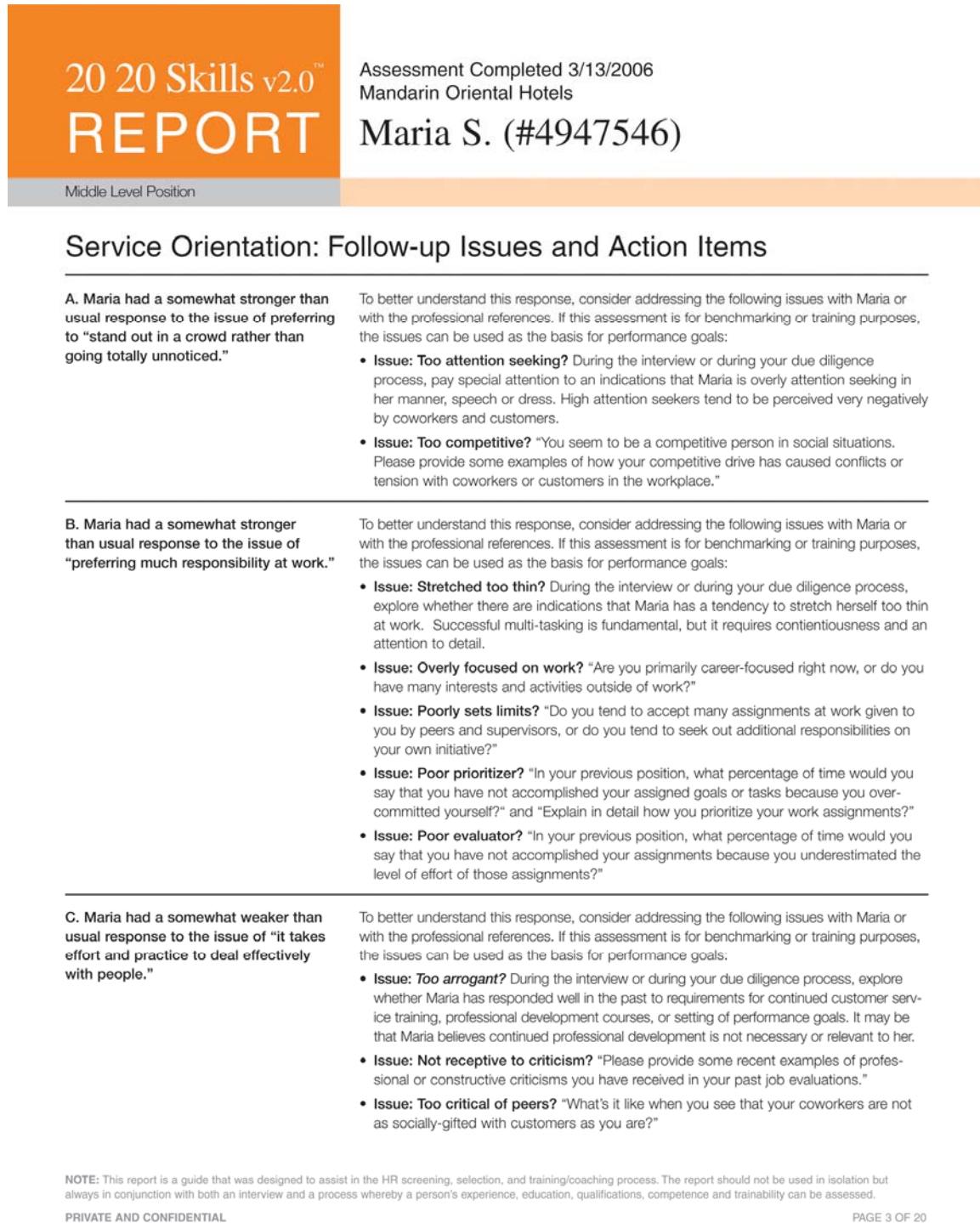
**Distinctive Response Patterns**



**LEGEND:** (—) Extremely Weak Response, (—) Weak Response, (✓) Normal Response (+) Too Strong Response, (++) Definitely Too Strong Response

**NOTE:** This report is a guide that was designed to assist in the HR screening, selection, and training/coaching process. The report should not be used in isolation but always in conjunction with both an interview and a process whereby a person's experience, education, qualifications, competence and trainability can be assessed.

**Figure 3:** Qualitative interpretation from the 20 20 Skills™ assessment v2.0 report recommending how HR professionals can apply the results of the item level IRT analysis.



We note that in Figure 2 the items are approximately positioned along the Y-axis where non-endorsement (*Disagree Completely* and *Disagree Somewhat*) changes to endorsement (*Agree Somewhat* and *Agree Completely*). This essentially yields a qualitative interpretation of the scores as it can easily be seen which items are likely to be endorsed by the test takers (those below the arrow) and which ones are not (those above the arrow). In fact, by considering the test item hierarchy as a whole one gains an immediate picture of what attitudes and behaviors are associated with Low, Medium, and High levels of Service Orientation. For instance, in Figure 2, candidates who are low in Service Orientation would endorse the statement “*I expect that my coworkers describe me as calm, cool and collected*” but not “*I prefer much responsibility at work.*” However, this last statement is endorsed by those with Medium and High levels of Service Orientation. Only those with the highest levels of Service Orientation would endorse almost all of the items shown in Figure 2.

This qualitative picture together with the IRT fit analysis will be extremely valuable to HR professionals because it identifies performance goals and training issues that need to be addressed in order to increase the candidate’s trait level of Service Orientation (or any other trait or competency). For instance, Figure 2 shows that the sample candidate (“Maria S.”) has the potential to be even stronger in her Service Orientation skills if subsequent employment training focused on the broad issue of “*I remain motivated to do what it takes to help my customers even if work is hectic and stressful.*” This training might encompass stress management, setting priorities and learning strong organization skills.

## Discussion

Given the increased competition among the 2,000 plus businesses that offer HR assessment, vendors are increasingly pressured to differentiate themselves in their marketing and sales materials. We have found that assessment vendors differentiate or promote their products in three main ways: stating their number of years in business in the field of industrial-organizational psychology, listing the academic credentials of their staff, and referencing a reportedly impressive body of supporting research and literature for their assessments. All of these broad marketing approaches aim to instill customer confidence in the scientific validity of the vendor’s products. Other vendors provide more detailed points of consideration when HR professionals consider competing products. Table 1 summarizes these issues around a predictable set of questions.

Although all of the questions in the first column of this table deserve consideration by HR professionals when evaluating competing assessments, column two indicates that each issue is effectively meaningless in the absence of a high quality of measurement. Indeed, as we emphasized earlier in the paper, quality of measurement is the foundation to the reliability and validity of any assessment instrument and research study. Of course, this means that the benefits of superior measurement provided by IRT have HR and business applications far beyond the construction of employee assessments. Indeed, modern test approaches offer state-of-the-art measurement for any number of qualitative and quantitative research questions.

**Table 1:** Common assessment issues noted by vendors and complementary points to these issues (adapted from Hogan Assessment Systems, Inc.).

<b>Common assessment issues raised in vendors' sales materials</b>	<b>Counter issues HR professionals should consider in tandem</b>
The vendor should be a member of the American Psychological Association, Society of Industrial/Organizational Psychology, or other professional organization that mandates ethical and statistical guidelines for creating assessments.	It is beneficial for vendors to hold memberships in professional organizations. However, the ethical and statistical guidelines can vary in quality and scope across professional organizations, and even among divisions within the same professional organization. Thus, while professional memberships and adherence to a basic set of guidelines is beneficial, this itself is no indication, much less a guarantee, of quality of measurement in an assessment.
The tests should preferably be reviewed in <i>Buros' Mental Measurement Yearbook</i> ?	Peer-review, be it with an academic journal or a review in <i>Buros' Yearbook</i> , is a standard approach for attempting to grade the general quality of a given piece of research. However, peer reviewers vary in background knowledge, expertise and practical knowledge. Moreover, the prevalence of CTT approaches in the social sciences implies that peer review by itself is no guarantee that a given assessment meets the more stringent standards of quality of measurement as defined by modern test theories. HR professionals need to know the precise credentials of the reviewers to help determine this latter issue.
Each test should be supported by a test manual that is organized according to the <i>Standards for Educational and Psychological Testing</i> or the <i>Uniform Guidelines on Employee Selection Procedures</i> ?	Standardized organization of test manuals allows HR professionals to better compare the background information of several competing assessments, but neither the existence of a test manual per se nor its organization speak to the quality of measurement of an assessment.
The vendor should supply technical reports containing validity studies using the tests in real organizations.	Technical reports are a necessary and standard practice. However, such reports are no guarantee of the quality of measurement of a given assessment. Furthermore, a validity study using traditional CTT approaches would not be considered competent in terms of modern test theory and analytics.
There should be a standardized validation process that is followed before the vendor implements a selection test in an organization.	Standardized validation processes are another necessary and standard practice. Yet, validations based in CTT approaches do not guarantee quality of measurement, and hence do not support the reliability and validity of an assessment.
The cutoff scores established for selection purposes should be rigorously obtained.	CTT approaches do not guarantee the reliability or validity of cutoff scores for selection purposes. The validity of cutoff scores is inherently grounded in the quality of measurement of the assessment.
The effectiveness of the tests recommended by a vendor should be systematically evaluated.	The effectiveness of any test should be carefully evaluated. However, CTT approaches for the systematic evaluation of performance of assessments do not guarantee reliable and valid results. The quality of outcome studies is inherently grounded in the quality of the assessment.
The vendor should maintain a research archive that can be accessed to confirm the results of validity studies.	Maintenance of research archives of raw data is an ethical and useful practice. But, CTT-based confirmation studies do not guarantee the reliability or validity of an assessment. The reliability and validity of outcome studies are only as good as the quality of measurement of assessments used for dependent and independent variables.
The vendor should have a policy for supporting customers in the event of a legal challenge to the use of a test.	It is a sound business and academic practice for vendor's to have clear policies for supporting customers in the event of a legal challenge to the use of a test. However, providing materials such as Technical Manuals or summaries of validity studies for use in legal proceedings does not guarantee the reliability or validity of that supporting material. The reliability and validity of test construction and validity studies are only as good as the quality of measurement used in research.

For instance, criterion validity, content validity and construct validity studies are the accepted standards for supporting the use of specific assessments. Criterion-related validity studies consist of empirical data demonstrating that the assessment is predictive of important elements of job performance. Content validity studies consist of data showing that the content of an assessment is representative of important aspects of performance on the job for which the candidates are to be evaluated. Finally, construct validity studies consist of data showing that the procedure measures the degree to which candidates have identifiable characteristics which have been determined to be important in successful performance in the job for which the candidates are to be evaluated. Unfortunately, assessments constructed with CTT approaches do not guarantee that the assessment meets the technical quality of modern test approaches which yield interval-level measures free of response biases related to extraneous variables, such as a respondent's age, gender, and ethnicity. It is crucial to control for such biases because statistical theory (Stout, 1987) and computer simulations (Lange, Irwin, & Houran, 2000) alike demonstrate that response biases can lead to spurious factor structures of constructs, significant distortions in scores, and consequently erroneous reliability and validity findings.

Finally, IRT based assessments and analytics can help overcome the statistical limitations noted above in other types of HR and business research, such as ROI and adverse impact studies, benchmarking studies, employee performance evaluations, and customer and employee satisfaction surveys. Thus, it is through careful research designs utilizing IRT approaches that evidence for the efficacy of employee assessments will be convincingly documented, as well as yield new and provocative insights that will inform current thinking and a host of scientific models with industrial-organization psychology that will ultimately take employee assessment to the next level.

### ***About the authors***

*James Houran holds a Ph.D. in Psychology and recently joined HVS to head the 20 20 Skills assessment business. Prior to joining HVS New York, James worked for a private online company developing proprietary tests and measurements and also researched personality and peak experiences for six years as an Instructor of Clinical Psychiatry at the SIU School of Medicine.*

*Rense Lange holds a Ph.D. in Psychology and a Masters' in Computer Science, both from the University of Illinois at Urbana-Champaign. He is one the world's foremost expert in tests and measurement and applied Item Response Theory and Rasch scaling, and Computer Adaptive Testing (CAT) in particular. In addition to serving on the faculty of the University of Illinois, the Southern Illinois University School of Medicine, and Central Michigan University, Rense has worked for ten years as the lead psychometrician at the Illinois State Board of Education and he is the Founder and President of Integrated Knowledge Systems, Inc.*

*The authors' individual and collaborative research programs have received worldwide recognition. Each has over 100 refereed journal publications, books, chapters, and conference presentations. Their research has also been featured in general media outlets like Rolling Stone, New Scientist, Psychology Today, Wilson Quarterly, USA Today, The Today Show, BBC, A&E, and the Discovery Channel.*

For more information, please contact the authors at:

James Houran  
[jhouran@2020skills.com](mailto:jhouran@2020skills.com)  
 516.248.8828 x 264

Rense Lange  
[renselange@earthlink.net](mailto:renselange@earthlink.net)  
 217.502.4589

HVS International – New York  
 372 Willis Avenue  
 Mineloa, NY 11501

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement (2002). *Standards for educational and psychological testing*. Washington, DC: Author.
- Bond, T. G. (1995). Piaget and measurement II: empirical validation of the Piagetian model. *Archives de Psychologie*, 63,155-185.
- Comrey, A. L. (1978). Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology*, 46, 648-659.
- Bond, T. G., & Fox, C.M. (2001). *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.
- Embretson, S. E. (1995). The new rules of measurement. *Psychological Assessment*, 8, 341-349.
- Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S.E. Embretson and S. L. Hershberger (Eds.), *The new rules of measurement: what every psychologist and educator should know* (pp. 1-16). Mahwah, NJ: Lawrence Erlbaum.
- Equal Employment Opportunity Commission, Department of Labor, Department of Justice, & the Civil Service Commission (1978). *Uniform guidelines on employee selection procedures (1978)*. Washington, DC: Author. <http://www.uniformguidelines.com/index.html>
- Finn, J., & Banach, M. (2000). Victimization online: the down side of seeking human services for women on the Internet. *Cyberpsychology and Behavior*, 3, 243-254.
- Houran, J. (2004). *Ethics in cross-cultural compatibility testing in Europe: an opportunity for industry growth*. Paper presented at the Internet Dating / Online Social Networking Industry Association Inaugural Meeting, Nice, France, July 15-16, 2004.
- Kline, P. (1986). *A handbook of test construction*. London: Methuen.
- Kline, P. (2000). *Handbook of psychological testing* (2<sup>nd</sup> ed.). London: Routledge.
- Lange, R. (in press). Binary items and beyond: a simulation of computer adaptive testing using the Rasch partial credit model. In E. Smith & R. Smith (Eds.), *Rasch measurement: advanced and specialized applications*. Maple Grove, MI: JAM Press. (Also to appear in a future issue of the *Journal of Applied Measurement*).

- Lange, R., & Houran, J. (2006, April). *Perceived importance of employees' traits and abilities for performance in hospitality jobs*. Paper presented at the 2006 International Objective Measurement Workshop. Berkeley, CA.
- Lange, R., Greyson, B., & Houran, J. (2004). A Rasch scaling validation of a 'core' near-death experience. *British Journal of Psychology*, *95*, 161-177.
- Lange, R., Jerabeck, I., & Houran, J. (2004). *Building blocks for satisfaction in long-term romantic relationships: evidence for the complementarity hypothesis for romantic compatibility*. Paper presented at the Adult Development Symposium Society for Research in Adult Development Preconference, AERA, San Diego, CA, August 11.
- Lange, R., Irwin, H. J., & Houran, J. (2000). Top-down purification of Tobacyk's Revised Paranormal Belief Scale. *Personality and Individual Differences*, *29*, 131-156.
- Lange, R., Thalbourne, M. A., Houran, J., & Lester, D. (2002). Depressive response sets due to gender and culture-based differential item functioning. *Personality and Individual Differences*, *33*, 937-954.
- Lange, R., Thalbourne, M. A., Houran, J., & Storm, L. (2000). The Revised Transliminality Scale: reliability and validity data from a Rasch top-down purification procedure. *Consciousness and Cognition*, *9*, 591-617.
- Larsen, D., Lange, R., & Hughes, L. (2006, April). *Restructuring the University of Pennsylvania Smell Identification Test to Measure Olfactory Ability in Pre-Clinical Neurodegenerative Disorders*. Paper presented at the Ninth International Alzheimer's Conference. Geneva, Switzerland.
- Linacre, J. M. (2004). *Facets Rasch measurement computer program*. Chicago, IL: Win-steps.com.
- Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: new problems, old issues. *American Psychologist*, *59*, 150-162.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, *55*, 293-326.
- van de Vijver, F.J.R., & Poortinga, Y.H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*, 29-37.
- Wainer, H. (2000). *Computerized adaptive testing: a primer*. Mahwah, NJ: Lawrence Erlbaum.
- Wright, B.D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.