

# 2006

---



## Bias and the Bottom Line: How *Wrong* Surveys Take Your Business in the *Wrong* Direction

Paper Presented at the 17<sup>th</sup> Annual HICAP: Hotel Investment Conference – Asia Pacific, Oct. 11-13<sup>th</sup>, InterContinental Hong Kong, Hong Kong

Rense Lange, President, Integrated Knowledge Systems, Inc.  
Gene A. Ference, President, HVS-The Ference Group  
James Houran, President, 20 20 Skills™ Employee Assessment

### **HVS INTERNATIONAL NEW YORK**

372 Willis Ave.  
Mineola, NY 11501  
+1 516.248.8828 (ph)  
+1 516.742.3059 (fax)

**November 2006**

---

New York San Francisco Boulder Denver Miami Dallas Chicago Washington, D.C. Weston, CT Phoenix Mt. Lakes, NJ  
Vancouver Toronto London Madrid New Delhi Singapore Hong Kong Sydney São Paulo Buenos Aires Newport, RI

---

**Abstract.** – Many assessment instruments used in the service industry have inadequate psychometric properties because they neither rely on modern test theory nor do they take into account item bias. Yet, these issues are critical when global businesses require cross-cultural assessment. Using actual employee-satisfaction data from a major hotel chain, we show that poor measurement and lack of bias testing predictably leads to spurious results that can negatively affect a company’s bottom line. In contrast, customization of assessments using modern analytics and bias testing yields powerful insights from data that are actionable, evidence-based and can provide a substantial return on investment.

---

The *Standards for Educational and Psychological Testing*<sup>1</sup> defines assessment as “any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects, or programs” (p. 172). Common examples in the service industry include employee screening and selection tests, company climate surveys and employee or guest satisfaction surveys. Hiring and training professionals know very well that assessments must meet professional testing standards and comply with legal requirements so that they are safeguarding against subgroups of test takers receiving unfair differences in assessment scores<sup>3</sup>.

We noted previously that satisfying testing standards and legal requirements is difficult because of biases that are inherent in the questions being asked<sup>3</sup>. For instance, in the context of job satisfaction such biases occur when *equally satisfied individuals* give systematically different answers to the same questions. It is crucial to control for response biases because statistical theory<sup>6</sup> and computer simulations alike<sup>4</sup> show that such biases can lead to spurious factor structures, significant distortions in scores and consequently erroneous research results. Such biases can also introduce unfair and illusory differences in scores across sub-groups of respondents (e.g., women vs. men, older vs. younger employees, and between those answering paper-and-pencil vs. web-based versions of a test or questionnaire), which invite substantial legal concerns. These issues are compounded when companies have *global* business units, in which case there are also cross-cultural factors with which to contend<sup>5</sup>. In particular, when questionnaires are translated, or when respondents are not native speakers of the language in which the questionnaire is written, biases almost certainly will creep in.<sup>a</sup>

Aside from psychometric and legal issues, there are pragmatic reasons to address cross-cultural factors in assessments. In particular, **organizations need results that are correct and that contribute positively to the company’s bottom line, and accurate questionnaire data also provide a powerful tool for helping set new directions.**

---

<sup>a</sup> The literature distinguishes several different types of bias, including response bias, item bias, test bias, and bias due to the systematic over- or under- prediction based on group membership. However, response biases are relevant only through their biasing effects on item endorsement, and item bias will eventually manifest itself as test bias as well as the systematic over- or under- prediction based on group membership. For this reason our terminology here does not consistently attempt to distinguish between these bias types.

However, biased assessment can significantly and negatively affect your bottom line by giving you misleading data. Acting on *wrong data* will take your business in the *wrong direction*. We therefore pose the following question: Are your current tests and questionnaires costing you money because they provide incorrect information due to responses biases related to respondents' age, gender or culture? Experience indicates that unless such biases have been studied and corrected for each instrument you are using, the answer is almost certainly *yes*.

Assessment vendors sometimes suggest that bias plays no role in their case because they have different norms for different subgroups, or they may point to an absence of group differences. However, such statements are beside the point – and, worse, they may be erroneous<sup>2</sup>: the *Standards for Educational and Psychological Testing* requires that those with those with equal trait levels should receive equal scores. If, say, US employees as a group are in fact more satisfied with their job than are Asian employees, then US employees *should* receive higher scores than Asian employees. However, US employees' higher scores should reflect their actual job satisfaction only. In other words, the presence or absence of group differences has absolutely no bearing on the presence or absence of bias<sup>2</sup>.

As should be clear from the above, we believe that selecting your assessment vendor requires the same – if not greater – due diligence as does the selection of employment candidates. Study your vendors' white papers, case studies, online FAQs or test manuals carefully to find out *if* and exactly *how* these address bias. If these materials do not clearly address the issue of response biases, then demand that information from the vendor. You are likely to discover that most vendors do not take this crucial issue into account – yet, it is the core foundation of the internal validity of all assessments. This article illustrates how response biases are detected through the use of state-of-the-art statistics grounded in modern test theory<sup>2,3</sup>, as well as what can happen if you instead choose not to identify and correct for biases in organizational assessments. Our approach is based on a form of modern test theory called “Item Response Theory (IRT),” which is the statistical gold standard used in the construction of such well known tests as the GRE, LSAT, and MCAT.

## **The Present Research**

The data for this project stem from a large-scale study on employee satisfaction completed by HVS – The Ference Group for a large international hotel chain. A total of 6806 employees completed the proprietary *Ference Satisfaction Scale (FSS)*. These employees were selected from three hotels (each) in the East Asia, South America, Europe, the United States of America, as well as Egypt and the Arab Peninsula. This study was proprietary and no age and gender information is available.

The FSS was developed by the Ference Group and consists of 97 questions that are to be rated on a 4-point category Likert-type ordinal dimension: Strongly Disagree,” “Tend to

Disagree,” “Tend to Agree” and “Strongly Agree.” In some cases items were worded negatively (i.e., agreement indicated *dissatisfaction*) and the ratings were reversed in such cases. Some examples of FSS questions are given at the end of this paper.

A series of analyses indicate that the FSS has many properties that are typically reported as evidence of psychometric soundness. Within the framework of classical test theory, item level factor analysis indicates that nearly half of the variance (47.2%) is explained by the first principal component. Unidimensionality is further supported by the finding that the largest eigenvalue (45.7) far exceeds the next highest ones (3.7, 2.2, 1.8, ...). Finally, internal consistency is nearly perfect, as coefficient alpha reaches 0.99.

For reasons outlined in earlier papers<sup>3</sup>, we use IRT based techniques – and Rasch scaling in particular – as our basic analytic framework (for an independent introduction, see Embretson<sup>2</sup>). These analyses similarly provided strong evidence of overall psychometric quality: Only five items of the FSS misfit the Rasch rating scale model and principal component analysis of items’ residuals indicate that the Rasch dimension explains 78% of the variance. The first residual factor explains just 6% of the variance, and the loading on the first residual factor of just 2 of the 97 items exceeds 0.5.<sup>b</sup> Finally, the Rasch reliability coefficient of the differences between respondents is 0.97.<sup>c</sup>

### *Analysis of Cultural Bias*

The above indicates that the FSS’ measurement properties surpass those of similar instruments dealing with satisfaction. However, additional analyses are needed to verify that the FSS yields unbiased scores as well. Regardless of the measurement framework adapted, this entails that items within a factor are essentially interchangeable and that items’ measurement properties should not vary systematically with extraneous properties of the persons being tested. For instance, when a particular subgroup is found to score higher than another, this should be because this group is indeed more satisfied. But when the scoring differences result from one group not understanding the assessment questions, such differences are meaningless – or worse misleading.

Unfortunately, this is exactly what we found. About 40% of the FSS items have significantly different qualitative properties depending on respondents’ geographic locations. In particular, *when equally satisfied employees are studied*, there are significant differences in the patterns of item endorsement ( $\chi^2_{485} = 6698.7, p < .001$ ) between hotels in Asia, South America, the United States of America, Europe, and Arabic countries. Note that it is *not* required that employees should express different overall levels of satisfaction in these five locations. But, the problem is that when respondents with the *same levels of satisfaction* are considered, respondents give systematically different endorsements to the various items depending on where they work and live.

---

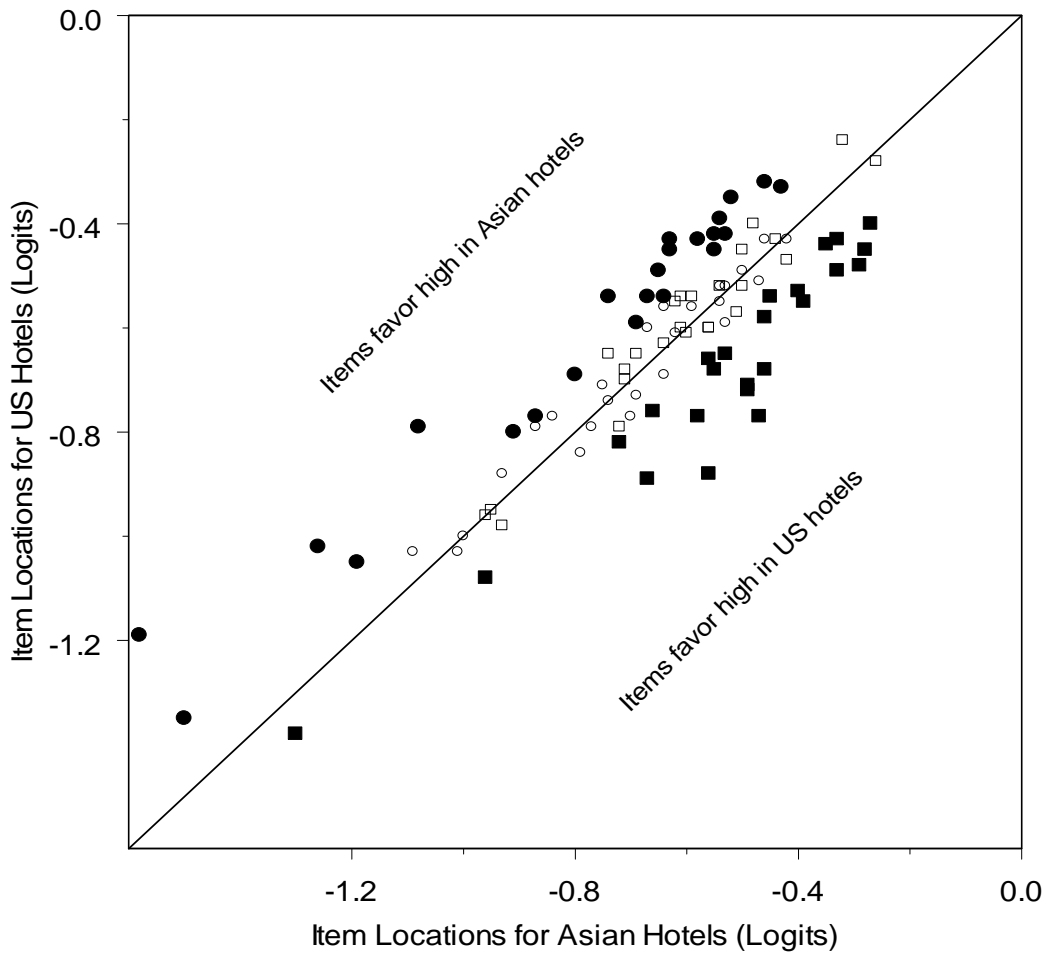
<sup>b</sup> Practice indicates that only absolute loadings greater than 0.5 are cause for serious concern.

<sup>c</sup> Because the Rasch model makes more realistic assumptions about data, Rasch reliability indices tend to be slightly lower than those obtained via Classical Test Theory.

The cultural dependence of satisfaction has two major implications. *Firstly*, it exposes a threat to *construct validity*. That is, two employees A and B with the same satisfaction “score” might have arrived at this score for quite different reasons. Note that this is analogous to the situation in academic contexts where one student obtains a high score by being good in mathematics whereas another student obtains the same high score by being good at reading tasks. While their scores may well be the same, the students certainly are different. *Secondly*, in extreme cases items’ cultural (or other) dependency may introduce *bias* into (i.e., distort) the entire measurement process. For instance, despite their similar scores, employees A and B might in fact not be equally satisfied at all! Thus, employee A might be very happy and motivated at his or her present place of employment, whereas B is ready to quit his or her current job almost immediately.

Although our data set covers hotels in five clearly different geographic locations, in line with the 2006 HICAP Conference, in the remainder we limit our analyses to just two of these – i.e., Asian-based hotels versus US-based hotels.

**Figure 1: FSS Item Locations for Asian vs. US employees**



### *Asia Versus the US*

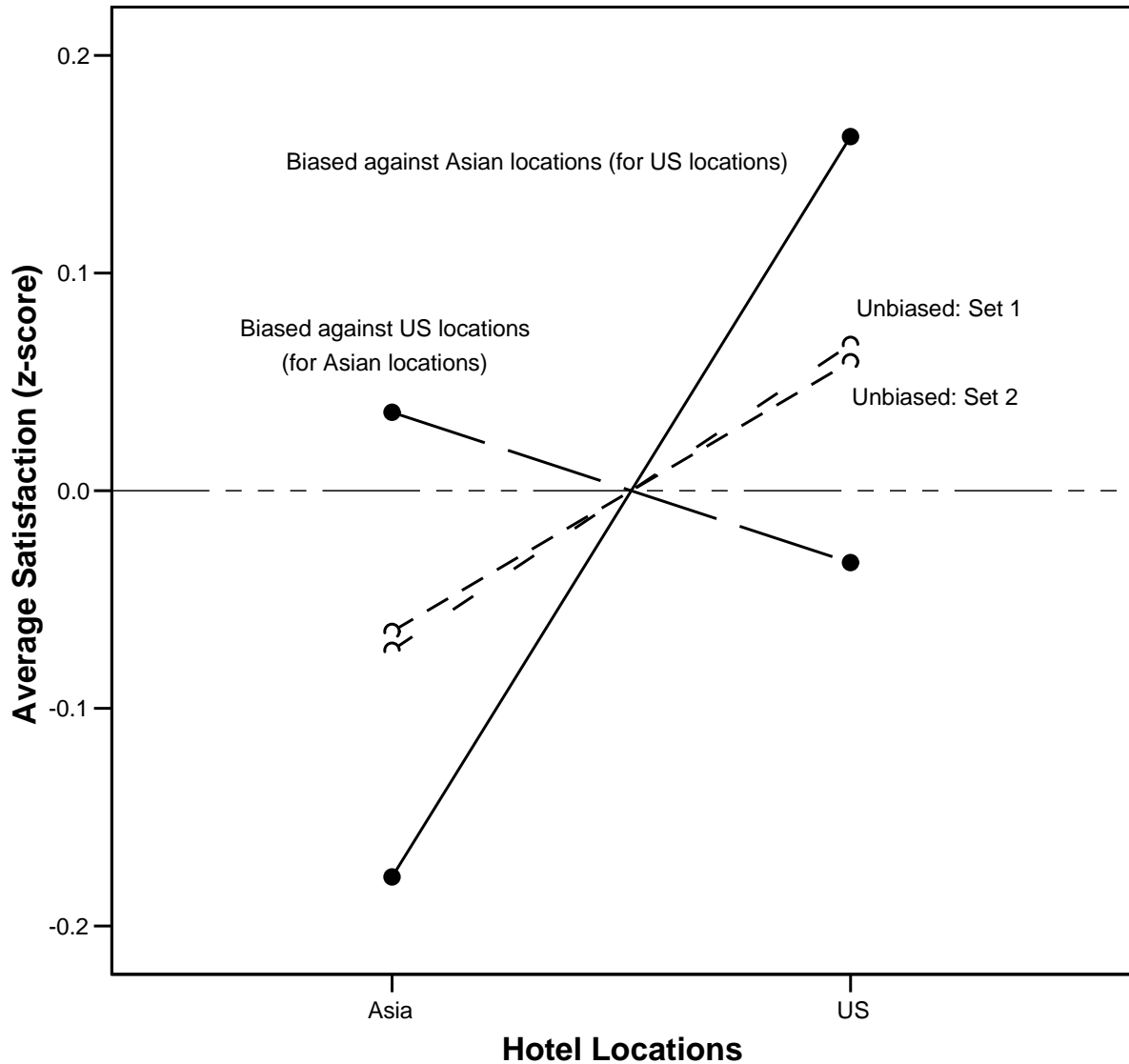
We already noted that when *respondents' overall satisfaction levels are held constant* items' endorsement levels should not vary between Asian and the US hotel personnel. The Rasch rating scale model quantifies endorsement in a log-odds (or, *Logit*) measure and the preceding thus implies that except for random error a plot of items' Asian vs. the US Logit measures should fall along a straight line. As is illustrated in Figure 1, when overall satisfaction is held constant the endorsement levels of the FSS items in Asia (X-axis) and the US (Y-axis) are indeed positively correlated ( $r = 0.87$ ).

This correlation is imperfect however, and we distinguish three different groups of items:

1. Items that receive **disproportionately high endorsement in Asian hotels** (i.e., relative to their endorsement by *equally-satisfied* US personnel). A total of 23 such items were identified, and they are represented by solid circles and they appear in Figure 1 above the equality line  $Y = X$ .
2. Items that receive **higher endorsement in US hotels** (i.e., relative to the endorsement given by *equally-satisfied* Asian personnel). A total of 24 such items were identified, and they are represented by solid squares and they appear in Figure 1 below the equality line  $Y = X$ .
3. The remaining 50 items received **nearly identical endorsement** (i.e., holding satisfaction constant), and these unbiased items are shown by the open markers near the equality line  $Y = X$ . To illustrate our earlier assertion that (unbiased) items within a factor should be interchangeable, these items were randomly divided into two equal size subsets (Set 1 and Set 2) as is denoted by the open circles and squares.

Without the type of analyses performed here, it is impossible to know whether the items being used in a satisfaction survey (or, for that matter, any other type of assessment) are indeed unbiased or not. It is very possible, for instance, that biased questionnaires are created without the test authors being aware of this. This might occur for example when a particular item response format is favored that is unfamiliar to those from certain cultures, when complex statements are favored over simple ones (thereby favoring greater [or at least different] understanding by native speakers), or when questions refer to "real-life" examples that are less likely to occur in certain cultures.

**Figure 2: Average Satisfaction of Asian vs. US Employees Measured by Four Item Sets**



For this reason, we computed respondents’ satisfaction based on each of the four item sets as identified in Figure 1 and the respective averages (after standardizing to z-scores) are shown in Figure 2. At least for the hotels under study, the two unbiased item sets (i.e., Set 1 and Set 2 with open markers) indicate that US personnel showed greater overall satisfaction than did Asian personnel ( $p < .001$ ). We can be quite confident of this finding because we first established that the items are unbiased. Set 1 and Set 2 yield nearly identical means within Asian and US hotels (i.e., the two dotted lines in the center overlap almost completely). As is required, this demonstrates that unbiased items are essentially interchangeable as they produce highly similar results.<sup>d</sup>

<sup>d</sup> The qualifier “essentially” is added to stress the fact that some technical measurement properties of a scale may change when items are exchanged.

We stress again that the fact that US personnel scored higher than Asian personnel does *not* constitute evidence for *or against* the presence of bias. However, *this difference varies with the items being used*, as is reflected by a powerful interaction effect ( $F_{3,8010} = 226.40, p < .001$ ). In other words, the slopes of the lines in Figure 2 differ systematically across the four item sets, indicating that the difference between the Asian and US means depends on which items are being used. In other words, the requirement that within a single factor items should be essentially interchangeable is clearly violated.

Notice that the difference between US and Asia becomes highly exaggerated when the items shown as a solid square in Figure 1 are used, as these *over-estimate* the satisfaction of US personnel (equivalently: these items *under-estimate* the satisfaction of Asian personnel). By contrast, the satisfaction of US personnel appears far lower when items favoring satisfaction in Asian personnel are used. In fact, this decrease is so dramatic that US personnel now appears to be *less* satisfied than Asian personnel, and the effect is strong enough to reach statistical significance ( $z = 3.43, p < .01$ ). Recall that Figure 2 expresses respondents' satisfaction in a  $z$ -score metric. It can thus be seen that the total biasing effect between Asian and US employees is nearly half of one standard deviation – a powerful effect indeed.

The crucial point therefore is that if, for some reason, one selected only the items shown as solid dots or as solid squares in Figure 1, *demonstrably wrong and highly misleading results will occur*.

**Table 1: Correlations Among Satisfaction Scores Derived Via Biased and Unbiased Items from the FSS**

Item Subset				
a. Items that exaggerate US satisfaction	1			
b. Unbiased Item Set 1	0.91	1		
c. Unbiased Item Set 2	0.92	0.93	1	
d. Items that exaggerate Asian satisfaction	0.88	0.92	0.91	1
Item Subset	a	b	c	d

### ***Correlations***

We noted earlier that the biasing effects reported here require stronger methods than the correlation based analyses provided by classical test theory. To illustrate this statement, Table 1 shows the correlations between the satisfaction measures obtained via each of the four item sets. It can be seen that all four measures are highly correlated (Median  $r = .91$ , all  $p < .001$ ), but the correlations are highly similar (range: 0.88 to 0.94). Despite their high correlation, we already know that two of these measures are in fact show powerful bias. Conversely, it thus follows that correlational methods – which form the basis of Classical Test Theory – are *not* suited to identify measurement bias.

## Secret for Successful Assessment: Customization Using Modern Analytics

“Off-the-shelf” assessments and “one-size-fits-all” survey solutions are readily available and often marketed by well-known vendors. The unfortunate fact is that such products also introduce major risks. Specifically, the issue of bias appears to be rarely, if ever, taken into account by vendors, and we strongly suspect that poor research results caused by biases are widespread in the industry. Assessments that are “customized” to particular business units and employees are an improvement. However, these solutions decrease the generalizability and comparability of results across the entire organization. **Thus, the service industry is likely not receiving the maximum benefits that surveys and assessments can offer; in fact, there may well be a negative effect on the bottom line.**

On the other hand, this article has demonstrated with real service industry data the power of customization with modern analytical techniques (i.e., IRT mathematics). Although space limitations prevent us from doing so here, IRT also allows one to correct for measurement bias without appreciable loss of generalizability.<sup>5</sup> While IRT approaches to hospitality research can be more expensive up-front than are traditional approaches, the superior quality of its results protect against making mistakes that in the end are far more costly yet. Advanced organizations rightfully view proper assessment as an important investment for growing and enhancing businesses, and we believe that the concepts and knowledge presented here will help you protect that investment.

In addition to protecting companies against making erroneous decisions as was described above, bias analyses will also identify the specific factors that determine employees’ job satisfaction given organizations’ idiosyncratic circumstances. Specifically, we found that for the hotel chain under study here, some factors contributed equally (i.e., across Asia and the US) to increasing employee satisfaction, whereas other issues differed across cultures. Since the detailed findings are obviously proprietary, we illustrate this point with a few high-level examples as these have clear implications for the likely success of this organization’s attempts to improve employee satisfaction.

Our analyses indicate that the following are the primary issues that Managing Directors should address in *both cultures* so as to increase employee satisfaction at this particular hotel chain:

### FACTORS THAT ENHANCE EMPLOYEE SATISFACTION EQUALLY IN ASIA AND THE US

“Respect in relationship between management and other personnel”

“My department head encourages people to make suggestions”

“I feel that the distribution of work is fair among employees”

However, several other issues are perceived as less significant by the Asian employees relative to the company's US employees. Thus, pursuing these should yield greater return on investment in its US branches than in its Asian branches.

**FACTORS THAT INCREASE EMPLOYEE SATISFACTION IN US – BUT NOT IN ASIA**

- “Confidence in fairness of hotel management”
- “New employees receive general hotel orientation”
- “My general manager visits our work place”

Finally, the following issues are not perceived as significant by the employees of the US-based business unit, and pursuing these should yield a greater return on investment in its Asian branches versus its US branches.

**FACTORS THAT INCREASE EMPLOYEE SATISFACTION IN ASIA – BUT NOT IN THE US**

- “Benefits package as good/better than other hotels”
- “My total compensation package is fair for my job”
- “Can talk to manager when I ask to do so”

These examples clearly demonstrate how the use of modern analytics allows us to go beyond the benchmarking of employee characteristics (or any outcome) to address the root causes for the issues being evaluated. Accordingly, the right instruments and analytics give companies targeted and evidence-based recommendations for taking its business in the right direction. While this paper used employee satisfaction as its main example, it should be stressed that IRT approaches apply to almost any type of assessment or survey. Also, many different types of data can be accommodated as IRT applies to True/False answers, estimates of proportions, counts and “partial credit” type data as well.

It is appropriate to close on a note about the concept of *culture* itself. We used the example of East versus West in this paper to make our examples clear and dramatic. However, subtle but significant differences in “company culture” can also occur among different business units of the same company that are located within the same geographic region. Such differences often provide unexpected and unwelcome situations, especially if there are programs and initiatives in place to align missions, visions and job performance across those units. In our experience, companies are eager to change from assessments that merely promise to ones that actually perform.

Likewise, Eastern philosophy notes that there is opportunity inherent in change – *zhuǎnjī* (“turn” + “incipient moment” = “favorable turn; turn for the better”), *liángjī* (“excellent” + “incipient moment” = “opportunity” [!!]), or *hǎo shíjī* (“good” + “time” + “incipient moment” = “favorable opportunity”). When it comes to the service industry, IRT analytics and bias testing allow employee performance evaluations, organizational climate surveys and customer satisfaction surveys *to work together* in a coherent, valid and streamlined fashion. Never before has the industry had an opportunity to fully gauge the alignment of business goals and strategies through assessment and subsequently gain a detailed plan for reaching those goals in more efficient and cost effective ways.

## About the Authors

**Rense Lange** holds a Ph.D. in Psychology and a Masters' in Computer Science. He is one of the world's foremost experts 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 dealt with all aspects of large- and small-scale assessment as the lead psychometrician at the Illinois State Board of Education and he is the Founder and President of Integrated Knowledge Systems, Inc.

**Gene Ference** holds a Ph.D. in Industrial-Organizational Psychology and is one of the highest respected Industrial Psychologists and Management and Organizational Development Specialists in the industry with over 35 years of experience in building peak-performing cultures and developing brand engagement strategies. His work has directly assisted clients in successful applications for the *Malcolm Baldrige National Quality Award*, *Employee of Choice*, *Best Human Resources*, *Employer of the Year* and *Fortune 100 Best Companies to Work For*, as well as the quality of work life and service culture awards of Great Britain, Brazil, The Netherlands, Mexico, Australia and Singapore.

**James Houran** holds a Ph.D. in Psychology and recently joined HVS to head the 20 | 20 Skills™ assessment business. He is a 15-year veteran in research and assessment on peak performance and experiences, with a special focus on online testing. His award-winning work has been profiled by a myriad of media outlets and programs including the Discovery Channel, A&E, BBC, NBC's *Today Show*, *Wilson Quarterly*, *USA Today*, *New Scientist*, *Psychology Today* and *Rolling Stone*.

For information on Best Practice Organizational Assessments & Professional Coaching and Training Workshops, contact:

Gene A. Ference, Ph.D.  
[gference@hvsinternational.com](mailto:gference@hvsinternational.com)  
203.266.6000

For information on the Best Practice 20 | 20 Skills™ assessment system and general IRT analytic applications, contact:

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

## References

- <sup>1</sup>American Educational Research Association, American Psychological Association, & National Council on Measurement (1999/2002). *Standards for educational and psychological testing*. Washington, DC: Author.
- <sup>2</sup>Embretson, S. E., & S. L. Hershberger (Eds.) (1999). *The new rules of measurement: what every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum.
- <sup>3</sup>Houran, J., & Lange, R. (2006). State-of-the-art measurement in human resource assessment. *HVS Journal*. 28<sup>th</sup> Annual NYU Hospitality Industry Investment Conference, New York, NY, June 4-6.
- <sup>4</sup>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.
- <sup>5</sup>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.
- <sup>6</sup>Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, 55, 293-326.