At Your Service: Blending Science and Practice to Develop a Customer Service Assessment Tool

Therese Macan  
University of Missouri - St. Louis

Lee Konczak  
Washington University

James Breaugh  
University of Missouri – St. Louis

When developing and validating a new assessment measure for employee screening and selection purposes, Human Resource (HR) professionals are faced with many decision points, challenges, and unique features. Given the lack of systematic direction in the literature on how to handle these challenges, we believe it is important to share approaches, decisions made, and lessons learned when blending science and practice in these types of situations. This paper draws on a real-life case and describes steps taken to achieve buy-in across multiple locations in developing a customer service-oriented assessment for use with part-time employees.

Human Resource (HR) professionals and consultants who have been involved in developing or implementing a selection system for an organization can attest to the challenges and pressures that often necessitate compromises among theory, best practices and organizational realities. Cascio and Aguinis (2005) noted that a key criterion for HR practitioners in selection is “organizational acceptability,” a concept relatively neglected by researchers. Perhaps this disparity provides some explanation why there are innumerable resources available that detail the psychometric steps necessary in designing and developing tests and assessments for use in organizations (e.g., Nunnally & Bernstein, 1994), but little information that discusses the contextual factors that can dramatically affect the decisions made and ultimate success of the newly designed selection techniques in organizations (Adler, Macan, Konczak, Muchinsky, Grubb & Hurd, 2008; Muchinsky, 2004). Exploring these acceptance and implementation issues may help bridge the gap between science and practice. Researchers may have an opportunity to learn from practitioners about the process of implementation.

Numerous researchers have lamented the gap between science and practice in personnel selection practices (Hollenbeck, DeRue, & Guzzo, 2004; Klehe, 2004; Macan & Highhouse, 1994; Ryan & Tippins, 2004; Rynes, Colbert & Brown, 2002). These contributions have highlighted the need to bridge this gap between state-of-the-art I/O psychology practices which are supported by research and organizations’ failures to adopt such practices, offering ideas for why this disparity may exist (Cascio, 2008;
Guion, 1998; Johns, 1993). Our paper takes a different approach and describes specific factors and strategies utilized to balance the achievement of high quality selection procedures given organizational constraints. Specifically, we detail a project involving the development and validation of a customer service-oriented paper-and-pencil assessment. Our aim is to share approaches others might consider incorporating in such a process to achieve a high quality product that is more likely to be used as intended and designed. We present five unique aspects or practical challenges to this project in which we blended science and practice in developing and validating an assessment. These challenges included:

1. Cost Effectiveness and Administrative Efficiency
2. Gaining Stakeholder Buy-In
3. Diversity of the Applicant Pool
4. Balancing Statistical and Practical Significance
5. Gathering Multiple Forms of Validity Evidence

We describe each of these challenges below with corresponding strategies employed, decisions made, and/or lessons learned. Where appropriate, we also indicate areas needing more research to aid in dealing with the dilemmas that arise between practice and theory that researchers should explore.

**Challenge #1 – Cost Effectiveness and Administrative Efficiency**

Recent research indicates that customers who hold positive perceptions of an organization’s service quality are more likely to remain customers and more likely to spread the word to expand the client base (Schneider, Ehrhart, Mayer, Saltz & Niles-Jolly, 2005; Schneider, White, & Paul, 1998). Many companies, especially in the entertainment/recreation industry, have discovered that achieving excellent service quality distinguishes them from their competitors (Zeithaml, Parasuraman & Berry, 1990). From an employee staffing perspective, the need to have good measures for hiring individuals who can provide excellent customer service is critical.

The host organization for this study sought an assessment device that: (a) predicted customer-service behavior, (b) had minimal adverse impact, (c) could be administered in a short period of time, (d) could be scored easily and quickly, (e) was suitable for applicants who did not need to have a high level of reading ability, and (f) could be administered in a paper-and-pencil format. The organization’s first decision point was whether it should buy an existing measure or develop one. Typically, it makes more sense for organizations to purchase a vendor’s existing assessment based on the measure’s rigorous development, validity evidence, and track record of withstanding legal challenges.

The organization explored using established instruments but each had its own limitations for this organization, rendering this approach not feasible. One major limitation was cost. This organization screens approximately 50,000 people a year. Existing scales typically cost $10-$15 per applicant meaning the use of such an
instrument would cost the company $500,000 to $750,000 a year. With these recurring annual fees, the organization decided it was more cost effective to design its own assessment device. An added benefit of this strategy is that the company would own the pool of items generated and thus have the ability to develop some variation of the test for another part of the company. A second limitation of many of the existing tests sold by vendors is that they are administered via a computer and the organization did not have the capability to handle such a large computer administration in the hiring time frame.

Other limitations of existing measures were that they were not appropriate for this organization’s applicant sample, not specific to the particular customer service delivery of this organization, and deemed not suitable given the jobs. The organization’s decision to develop its own assessment measure, therefore, addressed both scientific and practical concerns. From a scientific perspective, we know that different market segments require different approaches to service delivery, suggesting that human resource practices should ideally be developed specifically for their particular market service needs (Schneider et al., 1998). Schneider et al., however, lamented the lack of attention to this important area by HR professionals. By conducting a thorough job analyses and developing an assessment measure based on these particular elements, the organization could be confident its assessment instrument was strategically tapping the particular service quality elements necessary in its market. From a practical perspective, by developing its own scale, the organization could also ensure that the items written were face valid and perceived as job relevant, further garnering buy-in not only from applicants but from various organizational decision-makers (Gilliland, 1994; Gilliland & Hale, 2005; Ryan & Ployhart, 2000).

In sum, although there are many excellent selection instruments offered by vendors, we describe a situation in which an organization decided to develop its own assessment measure. Nonetheless, many of the strategies used to blend science and practice embraced during this process could certainly be instrumental to stakeholder buy-in when deciding among which already developed assessment tools to choose from vendors. The key seems to be understanding stakeholder needs and devising strategies to meet those needs while maintaining a solid grounding in test development and validation principles.

**Challenge #2 – Gaining Stakeholder Buy-In**

Muchinsky (2004) stressed the need for both science and practice to be considered when developing instruments. An exclusive focus on psychometric issues without being cognizant of organizational issues can result in an extremely sound instrument that the organization does not use or uses in ways not intended. Adler et al. (2008) emphasized the importance of explicitly attending to the change management issues when implementing new selection systems including how to involve key stakeholders. One major challenge in this project was to ensure buy-in by all relevant parties so that the instrument developed would be used and used properly. This challenge was further complicated in that this organization had multiple locations across the U.S. that already tended to “do things their own way” and viewed themselves as separate entities. For a
number of reasons (including legal), the corporate office wanted more standardization across the locations. Given this backdrop, it was crucial that we obtained buy-in from all the locations throughout all phases of this project. By sharing our experience, we hope to illustrate the various steps others might take to achieve buy-in from different parties in an organization. One step we took at the outset was to develop an Advisory Committee that we called upon at various stages in the process to review aspects of the project. This committee was comprised of the various decision-makers in the organization who had direct contact with key decision-makers at the various locations.

Consistently through every step of the job analyses, item generation, item refinement, and instrument validation, we involved individuals from each of the locations. Specifically, we included HR personnel, supervisors, incumbents and even applicants. As an example of this involvement, in conducting the job analyses, our first step was to learn more about the jobs through observations and focus groups. We conducted site visits and focus group interviews with job incumbents, front line supervisors, management and senior management. To further ensure inclusiveness, when in-person interviews could not be arranged, we conducted focus group interviews with job incumbents by means of conference calls. We followed these approaches with a job analysis survey not only to provide corroborating evidence (i.e., science) but also to include more individuals into the process and give voice to more incumbents, supervisors and HR personnel (i.e., practice). The left-hand side of Figure 1 provides a summary of the job analysis process we utilized. The goals of the job analysis were to: 1) secure appropriate information to document job requirements; 2) establish a foundation for selection tool development; and 3) involve appropriate employee groups in the process to ensure comprehensive data and buy-in to the project.

In developing the job analysis survey, we sought expert advice from employees in the HR departments at the various locations. Throughout scale development and validation, we shared our results and incorporated suggestions from the Advisory Committee. These same procedures were used in developing the scale items and their subsequent validation.

In addition, as shown in the middle section of Figure 1, we included three separate phases in our item refinement step not only to ensure a quality set of items (i.e., science) but to obtain buy-in from relevant individuals at every phase (i.e., practice). Specifically, after generating an initial set of items based on the results from the job analysis, these items were reviewed by organizational personnel. Based on their feedback, we deleted or revised items.

We conducted a second phase of item refinement to ensure that the items were: (1) easy to read, (2) understandable, (3) perceived to be fair, and (4) viewed as relevant for the positions. To achieve these goals, subject matter experts (i.e., employees of the organization who were familiar with the jobs in question) rated each of the items with regard to readability, clarity, fairness, and job relevance both in terms of job content and job skills. These raters were also asked to make any comments about the items and
suggest possible rewordings. We also presented the items to the Advisory Committee for their review and revised a few items based on their feedback.

In the final and third phase of item refinement, we administered the items to actual employees and applicants. The instructions informed respondents that their participation was voluntary. Applicants were also told that their responses would not affect hiring decisions. All respondents were asked to respond honestly and to provide any comments about the items. No names were gathered but respondents did provide age, race and gender information. Their responses to the items were analyzed to examine: (1) the degree of response variation per item and (2) whether any item showed adverse impact on ethnicity, age and gender.

Although this level of inclusiveness added time to the assessment development process, we undertook this extensive process to achieve buy-in from all constituents. In fact, we intentionally included larger sample sizes than dictated as necessary by theory to ensure buy-in. Although we cannot say for certain that this level of involvement was necessary, we chose to err on the side of more inclusion rather than less. Clearly, research is needed that evaluates the level of buy-in of those sampled as well as those not included to better assess what strategies are most effective in achieving buy-in.

**Challenge #3 – Diversity of the Applicant Pool**

One interesting element of this project was that we were developing the assessment measure for temporary, part-time employees. Typically organizations do not place much emphasis on developing selection processes for part-time individuals and may even recruit and select these individuals using no or little assessment (Landy, 2007). The part-time nature of the jobs resulted in a unique element with regard to the types of employees applying for these part-time positions. Specifically, there was a large age range within the applicants for these positions, spanning high school age students to retirees. Therefore, in developing the instrument we had to be sensitive to ensure that interview questions, job analysis surveys and, ultimately, the assessment items were understandable and relevant to a wide range of individuals (i.e., some younger job applicants had no work experience; others were retired with extensive job experience), free of jargon specific to a particular age group, and applicable to anyone from 15 to 80+.

In fact, we took extra steps throughout the process to ensure we were sensitive to these age issues. That is, we ensured that our sampling plan stressed that incumbents and applicants be representative of this wide age range. For the concurrent validation study, the age of respondents ranged from 15 to 86, with an average age of 33.4 years (SD = 18.84). Of course, individuals also were sampled so that both genders (59% female) and all ethnicities were represented. With regard to ethnicity, 44% were Caucasian, 28% Hispanic, 14% African American, 4% Asian/Pacific Islander, 2% other, and 1% Native American/Alaskan Native (7% did not report their ethnicity) for the validation sample. We consistently collected demographics on all samples of respondents and reviewed the numbers throughout all the stages of the process. We even conducted additional job analyses focus groups to ensure representation from all age groups. In addition, we went
Figure 1. Assessment Instrument Development Process

Step One: Job Analysis
- Phase 1
  1. Review Documents
  2. Create Focus Group Interviews
  3. Job Observations & Site Visits
  4. Focus Groups
    a. On-site
    b. Conference calls

Step Two: Item Generation and Refinement
- Phase 1: Initial Item Refinement
  SME review
  Advisory Committee Review

- Phase 2: Initial Item Refinement
  Items reviewed by relevant parties

- Phase 3: Further Item Refinement
  Items administered to:
  - Employees
  - Applicants

Step Three: Concurrent Validation Study
- Phase 1
  1. Developed Instrument & Similar Vendor Instrument administered to Employees
  2. Supervisors provided job performance ratings

Final Instrument Developed
CONVERGENT EVIDENCE with other measure and STRONG BUY-IN
outside the organization and had high school students review the initial assessment items to ensure comprehension. Finally, we computed a readability test on the final assessment. The final assessment received a score of 5.4 on the Flesch-Kincaid Grade level test, meaning the items were understandable to an average student in the 5th grade, consistent with job requirements.

It was a challenge creating an assessment instrument for such a wide age range of individuals (e.g., you could not ask about experiences on a prior job given this was the first job for some applicants). One lesson learned from trying to create such a scale is to seek as much review as possible of the items from as many sources as possible. For example, in administering the initial pool of items, we not only asked incumbents to respond to the items, we also asked them to give us feedback and suggest revisions to the wording of the items by including a writing space after each item. Throughout this project, we also paid close attention not only to the items but also to the instructions we gave. We made sure we had a wide range of individuals review the instructions to any survey, assessment, etc. to ensure understanding by a diverse sample.

**Challenge #4 – Balancing Statistical and Practical Significance**

A number of statistically-related decisions had to be made during the project that required balancing science and practice. For example, in conducting the job analysis survey, incumbents and supervisors made both frequency and importance ratings for each behavior/trait (e.g., “Be a Team Player”). Given that a specific employee behavior or attribute may be important even though it does not frequently occur, we decided to focus on the importance ratings in evaluating whether a given behavior or attribute deserved attention in the development of the instrument (most correlations between these two ratings were quite sizable ranging from $r = .45$ to $r = .82$).

With these same job analysis data not only did we have to decide what evaluation standard to use for judging whether a given employee behavior or attribute was important, we had to decide how to determine if there were meaningful differences across departments and/or locations. While conventional scientific practice might suggest we look at the results of statistical significance tests, with a large sample ($n = 396$), even small differences can reach a $p < .05$ level. For example, collapsing across locations, we found a statistical difference among the three major departments (e.g., retail, food/beverage, attractions) ($F = 5.10, p < .01$) for the importance rating for the item “Provide Service to Guest”. While the need to consider effect sizes and tests of significance is good science (Cohen, 1994; Shea, 1996; Thompson & Snyder, 1997), we think it is also prudent to examine the mean values. In our case, the mean values were 2.87, 2.97, and 2.96, on a 3-point scale (1 = “not important”, 2 = “somewhat important”, 3 = “very important”). Although the first importance rating for providing customer service might have been less than the other two, clearly such service still was rated as being extremely important. Thus, from a practice perspective, given the purpose of the job analysis process was to determine whether the instrument to be developed could legitimately be used across various departments and locations, we decided it was not
appropriate to focus on such statistical group differences as long as a competency was rated as being important in an absolute sense in all groups.

In summary, using our decision rules above, we found that two of the 21 behaviors/attributes (i.e., “Sell Things” and “Math Skills”) did not receive an average importance rating of 2.50 or higher on the three-point importance scale (1 = “not important”, 2 = “somewhat important”, 3 = “very important”). With regard to the item, “Sell Things”, positions in the one department received lower ratings (mean = 1.69) than did positions in the other two (mean = 2.29) and (mean = 2.61). A similar pattern of differences was found for the item “Math Skills”; positions in the one department received a mean score of 1.86, positions in the other two received a mean score of 2.66 and a mean score of 2.81. Thus, the results of the job analysis survey strongly supported the development of an instrument that assessed the 19 behaviors rated to be important across the major departments and locations. The remaining two behaviors were not assessed because they were not applicable to all three departments. Instead, we suggested that a separate assessment be developed for these skills relevant to the specific departments in which they were important.

Challenge #5 – Gathering Multiple Forms of Validity Evidence

We conducted a concurrent validity study to assess the properties (e.g., validity, adverse impact) of the instrument we developed. In conducting the concurrent validation study, it was important to gather data from a cross-section of employees at each of the locations and their supervisors. Participants were told that the organization was creating a questionnaire for hiring part-time employees. They were asked to help with development by responding honestly to the items. Although they were asked at the end to provide their name for data tracking purposes, respondents were assured their responses would be kept strictly confidential, be used for research purposes only, and not affect their employment status in any way. The unique aspect to this validation was that respondents were also asked to complete an existing instrument purchased through a vendor. While textbooks on scale development recommend convergent validity be examined, in practice this is rarely done (Spector, 1992). Instead, scores on the assessment instrument being developed or examined are typically only correlated with performance measures.

We also gathered job performance data from supervisors who were asked to rate the employee’s job performance honestly as accurate ratings were necessary to help understand how to increase the effectiveness of the scale. They were assured their ratings would be kept strictly confidential and used for research purposes only. No employment decisions would be made on the basis of the appraisal ratings. In addition, no individual ratings would be shared with Senior Management. Supervisors made ratings on 11 different performance dimensions taken from an existing performance appraisal form (e.g., “Guest/Customer Relations (internal and external)”; “Ability to work effectively with others”) and one overall job performance rating.

Out of the 88 “draft” items completed by employees in this sample, twenty-seven were selected to comprise the assessment. Items were chosen based on (1) their
correlation with performance, (2) low adverse impact, and (3) content coverage. We assessed content coverage by reviewing the individual items and making sure the various categories previously identified through the job analysis were being tapped. In terms of predicting job performance, we examined two performance measures. The first measure (Overall Performance Rating) was a one-item overall final rating made by a supervisor. The second measure (Summary Performance Measure) reflected an index in which the 11 dimensional ratings of performance made by supervisors were summed. The correlation between these two performance measures was \( r = .89 \). Given this high correlation, it is not surprising that the assessment instrument predicted both performance measures \( (r = .29 \) and \( r = .27 \) respectively (see Table 1).

We believe these correlations (none of the correlations reported are corrected for unreliability or range restriction) are particularly impressive if they are compared to the validity coefficient for the measure previously purchased from a vendor \( (r = .25) \). The validity coefficients are especially impressive given the vendor measure had been subjected to considerable more development, is much longer than the measure we developed, and involved the use of multiple response alternatives. (Given the administration of the measure was on-site and computer administration was not feasible, our items only involved dichotomous response options in order to be easily hand-scored.)

**Table 1**

*Validity Evidence for the Assessment Instrument Developed*

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Device</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assessment Instrument Developed</td>
</tr>
<tr>
<td>Overall Performance Rating (final rating on performance appraisal instrument)</td>
<td>.29**</td>
</tr>
<tr>
<td>Summary Performance Measure (scores on 11 items on appraisal measure were summed)</td>
<td>.27**</td>
</tr>
<tr>
<td>Vendor Instrument</td>
<td>.39**</td>
</tr>
</tbody>
</table>

*\( p < .05, **p < .01 \)
Our assessment score correlated ($r = .39; p < .01$) with scores on the vendor’s instrument. Thus, we found that the new measure, while correlated, is not redundant with the vendor measure. By including the previously used measure in our validation stage, therefore, we were able to compute not only criterion-related validity evidence, but we were also able to assess this convergent validity evidence. In addition, the vendor measure provided us with baseline information that allowed us to better compare the validity evidence we obtained. That is, we found that our new measure predicted the same performance as well as the previously used assessment.

**Conclusion**

We believe the assessment device developed meets the objectives we set out to achieve. That is, the assessment measure predicted supervisors’ ratings of employee performance for part-time positions at a respectable level with minimal to no adverse impact. It is quite short, consisting of 27 items, and can be administered in about 10 minutes while requiring less than 5 minutes to score by hand. In this way, the new selection device was a cost effective alternative that also could be administered efficiently (Challenge #1). In addition, the assessment instrument was developed with the diverse ages, races and genders of applicants in mind. It is also easy to read, with the readability index indicating that the items are at a 5th grade reading level. Thus, the assessment is applicable to a diverse pool of applicants (Challenge #3).

Of most unique importance though is that a large number of individuals in the organization participated in the development of this instrument as a means to gain stakeholder buy-in (Challenge #2). In fact, we intentionally over-sampled employees of the organizations throughout the various phases of the project as a means of increasing buy-in and ensuring that the specific customer service elements necessary for this market were captured. We made decisions based on the data gathered but also balanced statistically significant concerns with the practical aspects of the outcomes (Challenge #4). We examined both concurrent and convergent validity, obtaining multiple forms of validity evidence (Challenge #5) as means to help the organization gauge the value of the approach taken. Thus, the assessment device developed should allow the organization to make good hiring decisions for a diverse sample of potential employees (science) and be used across the organization as intended (practice). We believe we accomplished this by focusing on customer needs and requirements, involving key stakeholders early and often throughout the development and validation process and adhering to the underlying scientific principles of sound test construction and validation to achieve a healthy balance of science and practice.

Bridging the gap between science and practice within selection testing though offers a vast opportunity for future research efforts. Our study highlights the need for researchers to think beyond psychometric and legal defensibility. We need to understand the organizational context, concerns and politics that would encourage organizations to accept and use sound selection tests. In fact, we hope our descriptive case study sparks numerous research ideas that aim to provide evidence for using the various approaches we incorporated. We echo our colleagues’ call for prescriptive studies that lead to the
best strategies for fostering acceptance and implementation of the innovative scientific findings with regard to various selection assessment measures (Lievens & Chapman, 2009; Sackett & Lievens, 2008). Moreover, we need to think long term and assess whether any buy-in advantage disappears with time. That is, how do we solidify the use of sound selection practices when current employees are gone and new people who were not involved in the development of the instrument are using the assessment? In sum, researchers and practitioners need to partner more (as we did in this study) and share results. In this way, we can strive to find the most effective approaches to overcoming organizational resistance to innovation and “sell” state-of-the-art selection practices to practitioners (Johns, 1993; Lievens & Chapman, 2009).

References


**Author Note**

Please address all correspondence to:

Therese Macan  
University of Missouri - St. Louis  
Department of Psychology  
One University Boulevard  
St. Louis, MO 63121-4499  
Therese.macan@umsl.edu  
(314) 516-5416 office  
(314) 516-5392 fax

Other Authors Information:

Lee J. Konczak  
Washington University  
Olin Business School  
Campus Box 1133  
One Brookings Drive  
St. Louis, Missouri 63130  
konczak@wustl.edu  
(314) 935-5042

James A. Breaugh  
University of Missouri – St. Louis  
College of Business  
One University Boulevard  
St. Louis, Missouri 63121  
jbreaugh@umsl.edu  
(314) 516-6287