

EVOLUTION OF THE ACCOMPLISHMENTS SURVEY INTERVIEW (ASI)TM SYSTEM

Richard Joines, President
Management & Personnel Systems, Inc.

www.mps-corp.com

This paper discusses the background of our Accomplishments Survey Interview (ASI)TM system. The ASI is a pure manifestation of the ***Behavioral Consistency Method*** in the context of a ***Structured Interview*** process. To fully understand the ASI, it is necessary to briefly review these selection concepts.

The Behavioral Consistency Method

The ASI grew out of a Training & Experience (T&E) evaluation system developed by the U.S. Office of Personnel Management (OPM) during the 1970's. The work at OPM resulted in an internal publication that reported research on a new method of conducting examinations for professional and supervisory jobs (Schmidt, Caplan, Bemis, Decuir, Dunn & Atone, ***The Behavioral Consistency Method of Unassembled Examining***, U.S. Office of Personnel Management, Washington, D.C., 1979).

The new unassembled examining system was based on what is known as the behavioral consistency principle, which holds that the best predictor of future performance is past performance. As the authors stated in their 1979 OPM report:

“Past behaviors (not past exposures) are the best predictors of future behaviors, and the more similar the past behaviors are to the future behaviors, the better they should be as predictors (page 7).”

At the time of the OPM report, candidates for professional and supervisory jobs did not have to take a written test or interview. Instead, candidates were evaluated based on the information contained on their application form, which was sometimes supplemented by a second and more job-related application. In reviewing the candidate's application, points were awarded based on the candidate's education and experience. More education and more experience related to the target job led to more points being awarded.

The OPM team viewed this approach as “credentialism.” They argued that people with the same education and experience (e.g., same amount of experience in managing budgets of the same size) would not necessarily be equally qualified, that one person could have demonstrated a significant pattern of achievement while the other had only a mediocre track record. To the OPM team, the relevant question regarding budget experience would be which candidate demonstrated skill in managing budgets, not which candidate had the most experience.

The OPM team's view was that credential-based systems had limited validity. They reasoned that validity would be greater if the system gave credit for what people had actually done. They further reasoned that a person's past accomplishments would serve as valid indicators of past success, and that those with a strong track record would be more likely than those with a mediocre track record to be successful in the future. In short, they argued that the behavioral consistency principle would hold true, and this would result in a system with high validity.

The OPM team proceeded to devise a completely new system that required candidates to report their accomplishments. They developed a specific reporting format that required all candidates to report the same information, quoted below (page 51):

1. What the problem or objective was.
2. What you actually did and when (approximate date).
3. What the outcome or result was.
4. The estimated percentage of this achievement which is directly attributable to you and the estimated percentage which is due to the efforts of other people. If you do not give an estimate, you will be claiming total credit for the achievement.
5. The name, address, and telephone number of someone who can verify the information.

The OPM system for collecting candidate information about accomplishments has been reworded by some, including MPS, but since the approach is in the public domain, no company should claim it as their own (hint: be on the lookout, and you'll see this happen). At MPS, we take no credit for the simple, yet truly ingenious way of having candidates structure the information about their achievements. We use this approach, but in so doing, extend our thanks to the OPM team that devised this system.

A very significant advantage of the OPM team's approach of focusing on prior accomplishments is its content validity. As the OPM team stated in their report:

“The appropriate validity basis for scores derived using this procedure is content validity. The behaviors sampled in the achievements are content valid because they sample the kinds of achievements required in performance of the job (p. 13).”

After designing the system, research was undertaken to compare it with two commonly used federal T&E systems. The result was very favorable for the behavioral consistency method. It was found that candidates could be more reliably rated (i.e., be assigned similar ratings by different Personnel Staffing Specialists), and importantly, the team convincingly argued that the new approach would likely have much higher validity than the other existing federal T&E procedures.

Since the publication of the OPM report, there has been a great deal of research in support of the validity of the behavioral consistency method. It is not an overstatement to say that this method has gained universal acceptance within the profession of Industrial/Organizational Psychology. The research in support of its validity is overwhelming.

The leader of the OPM team, Dr. Frank Schmidt, was already a major figure in the field of I/O Psychology at the time the OPM report was issued. Dr. Schmidt has made many significant contributions with regard to employment testing as well as statistical methods for understanding research design and results.

One of the areas in which Dr. Schmidt is best known is using meta-analysis techniques to estimate the true validity of different types of tests. In this arena, he and his publishing partner, Dr. John Hunter, are the leaders in the field, having established many of the techniques that have become accepted within the I/O Psychology profession.

Dr. Schmidt and Dr. Hunter collaborated in 1998 to publish a very significant article for the testing profession (Schmidt, Frank L. & Hunter, John E., *The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings*, Psychological Bulletin, 1998, vol 124, No. 2, 262-274). In this article, the authors reported the validity of various well-known, commonly used testing methods.

Of particular interest was their finding with regard to the validity of the behavioral consistency method. The meta-analysis results suggested a whopping validity of .45. This compares very favorably with the best of the best available employee selection methods. Interestingly, they found that the accumulated research gives a completely different picture of the validity of the more traditional T&E point methods, which came in with a low validity (.11).

As experts in testing know, a validity of .45 is great, while a validity coefficient of .11 is minimal and typically not of much practical value. In short, all of the accumulated research firmly establishes the validity of the behavioral consistency method, and the wisdom of the OPM team in devising this new approach.

The Structured Interview

During the 1970's and 1980's, research on the interview was discouraging. Many I/O Psychologists were ready to conclude that the interview, as a selection technique, was of little or no value. The interview was thought to be fraught with various types of interviewer rating errors, which resulted in low reliability and low validity. Study after study found new pitfalls with the interview. The picture that emerged was that there were just too many problems to be overcome.

While all of this was going on, some very important changes occurred in the way interviews were typically conducted. In the traditional unstructured interview, the interviewer is free to ask any question as long as it's legal. The interviewer is also free to probe any area that comes up or that seems like a fruitful line of inquiry.

Upon a moment's reflection, it should be apparent that the interviewer is the key element here, not the system. One interviewer might be very, very good while another might be very poor. In short, the unstructured interview places a premium on interviewer training, ability and skill.

Research results identified many problems and many interviewer rating errors, all of which detracted from the reliability of the interview process. In addition, many employers paid the price of permitting this type of freedom because interviewers would often err in asking questions that violated Equal Employment Opportunity or Age Discrimination laws.

Primarily for these reasons, government agencies embraced the “structured” interview, and moreover, a "panel" approach to interviewing. In the structured interview, all questions had to be determined in advance and all of these questions had to be asked of all the candidates. Thus, there was "structure" to the process.

To determine which questions to use, Human Resources professionals turned to position descriptions to identify which knowledges and skills should be the subject of the interview questions. Frequently, they would get input from managers of the target job, or even conduct a formal job analysis. The need to predetermine the questions led to interviews that were more directly job-related and content valid.

Let's look briefly at what might be expected in research results from: (1) imposing interview structure by requiring that the same questions be asked of all candidates; (2) using a panel of three interviewers rather than a single interviewer; and (3) focusing in a conscious and planned way on developing a set of interview questions that measured the knowledges and skills needed to do the job; in other words, asking job-related questions in the interview.

First, let's ask what effect imposing the same set of questions would likely have on research results. This is actually a pretty straightforward, common sense proposition. If different interviewers interview the same candidates using an unstructured interview system, they wind up with different information because they asked different questions. When they rate candidates, they do so on the basis of this different information, and it therefore should not be a surprise to learn that the ratings tend to vary quite a lot. In other words, the consistency (or reliability) of the ratings is often a problem.

On the other hand, when the interviewers all ask the same questions, they have essentially the same information upon which to base their ratings, and not surprisingly, their ratings are more consistent. This is another way of saying that the reliability of their ratings are much higher.

Second, what about using a panel vs. a single interviewer? From testing theory and statistics, we know that if three equally reliable interviewers combine their ratings to arrive at a rating for the candidate, the score will have much higher reliability. This is actually a statistical fact. Those who aren't put to sleep by statistics can look up the Spearman-Brown prophecy formula, which is the formula used to estimate reliability using multiple raters. Using three raters is akin to tripling a test in length, which increases the reliability of the test.

Now, here's the important point. If ratings made by individual raters are low in reliability, then combining their ratings can have a huge, positive impact on the reliability of the process. This

happens to be another statistical fact. Basically, when the process has a lot of imprecision, adding a second rater helps eliminate error and reliability is higher. When dealing with a weak system, statistical analysis proves that two heads are better than one! Adding a third rater further improves reliability where the system is imprecise.

Third, a planned and conscious effort by Human Resources professionals to focus interview questions on job-related tasks, duties or Knowledges, Skills and Abilities (KSA's), can be expected to increase the validity of the process. Even though the process may be supported and defended on the basis of content validity, the goal of every Human Resources professional is to produce a ranked list that gets it right, with candidates ranked from the candidate who is most likely to be successful down through the candidate who has the least chance of being successful. The best way to achieve this is to ask questions about topics that are job-related.

In 1994, a major article on interviewing was published, and this article caught a lot of I/O Psychologists and so-called interviewing experts off-guard. The article was: *Hunter and Hunter (1984) Revisited: Interview Validity for Entry-Level Jobs* (Huffcutt, Allen I., and Arthur Jr., Winfred, Journal of Applied Psychology, 1994, vol. 79, No. 2, 184-190). The article was astounding because it breathed new life into the interview. In the process, it also revealed the value of the structured approach to interviewing.

Since 1984, many in the I/O profession accepted the interview research results given in the Hunter and Hunter article as the death knell for the interview. These results were rather dismal, to say the least. They put the validity of the interview in the dog house, down at .14, which was about the same as a credential-based T&E system.

The Hunter and Hunter article had mental ability tests at a validity of .53. A validity of .53 is terrific, while a validity of .14 is low. It may be of some practical value under some circumstances, but not much.

In their 1994 article, Huffcutt and Arthur studied a much larger number of interviews than Hunter and Hunter, and they did so by distinguishing interviews by their degree of structure. They proceeded to reach a phenomenal conclusion:

"Interviews, particularly when structured, can reach levels of validity that are comparable to those of mental ability tests (page 184)."

What a dramatic turnaround for the interview! The interview had vaulted from a virtual deathbed of selection methods to the elite realm of tests of general mental ability.

The interview's resurrection was perhaps not quite as dramatic as just portrayed. Some people had noticed a trend of more favorable results that had been building for some years. Three significant studies since Hunter and Hunter's 1984 paper had found support for the validity of the interview (Wiesner and Cronshaw, 1988; McDaniel, Whetzel, Schmidt, and Maurer, 1991; Marchese

and Muchinsky, 1993), but the Huffcutt and Arthur article made sense of everything. It was clear: if you structure the interview, you get much better reliability and validity.

The Huffcutt and Arthur article also addressed another important issue. They asked whether it was better to structure the interview as much as possible, or whether a point could be reached where additional structure didn't yield additional payoffs.

Some government agencies had reasoned that if some structure was good, complete (100%) structure would be best. This led to the development of interviews that required interview panel members to read the same questions for all candidates, then be silent!

In the 100% structured interview, oral panel members were not allowed to ask a candidate anything except the main questions to be asked of all candidates. Oral panel members were not allowed to ask a candidate to repeat an answer, even if they didn't clearly hear the candidate's answer. Also, oral panel members could not ask any kind of clarification or follow-up question, even if they thought they needed to do so in order to understand the candidate.

In this extreme form of structure, oral panel members had their hands completely tied. The idea of this interview was that 100% structure would produce higher reliability than any interview with less than 100% structure. This was viewed as a worthwhile goal by many. The question becomes a little tricky. It goes like this: "Why would anyone want to do anything that would lower reliability?" Strangely enough, there are good reasons to do just that!

What the advocates of the 100% structured interview failed to understand was that increasing reliability was not a desirable goal if it meant detracting from validity. Consider the following situation.

Suppose three oral panel members "think" they hear an answer which they consider negative. They aren't allowed to seek any clarification to verify what they think they heard. Given their situation and what they think they heard, they all assign a rating of 1 on a 7 point rating scale for the interview rating factor to which the question pertained. Since the ratings of all three panel members are the same, the reliability of the ratings is perfect on this rating factor. The question is whether perfect reliability equals perfect validity.

Assume now that the oral panel members were permitted to ask a clarification question, and all of the oral panel members realized that the candidate had said something which was actually very positive. Finally, suppose that the ratings now assigned by the three panel members are: 5, 6, 6. There no longer is perfect reliability because the ratings are not all the same.

Now here's the point. If the ratings by the panel members are now more accurate and a truer reflection of the candidate's ability to be successful on the job, the interview has higher validity! Thus, the ratings may be somewhat less reliable, but they are more valid.

This leads us to the crux of the matter. Which is better, reliability or validity? The answer to this question is simple. Validity is our ultimate goal. Validity trumps reliability every single time. We can live with less reliability if it gives us greater validity. Remember, our goal is to serve the best interests of the organization. That is achieved by producing a list that has the candidates correctly ranked based on their probability of being successful on the job. Correctly ranking the candidates refers to validity, not reliability. It's not a debatable issue. We want reliability but at the moment it interferes with maximizing validity, we don't want it any more. Case closed.

The answer provided by the Huffcutt and Arthur research was that increasing structure beyond a certain point didn't result in any meaningful improvement in interview validity. Absent any research support for complete structure, there really is no reason to require oral panel members to sit in silence when they need to ask a clarification or follow-up question that will give them information needed to make a more valid rating.

Finally, we must add that this is a common sense proposition. This exact scenario was not investigated by Huffcutt and Arthur because, even though they had 114 research studies to sort through, this was still not a large enough number to investigate every possible condition.

The ASI is a Marriage of Two Methods

If you marry the essential features of the Behavioral Consistency Method and take the best of what we know about Structured Interviewing, you get the ASI. The ASI is a logical outgrowth of these two selection methods.

Think of the ASI this way. On the one hand, we learned from the Behavioral Consistency Method that what really counts when evaluating someone's preparation for a job is the person's track record of accomplishments in job-related performance dimensions. Next, we learned from interviewing research that the structured interview yields much higher reliability and validity and can be just about as good as any other existing employee selection method.

The ASI marries these two selection systems by taking the information available about candidates in the Behavioral Consistency Method and putting it into a structured interview context. What the marriage means is that highly job-related information can be put at the disposal of the interviewers, and using a structured interview approach, the interviewers can probe the information and develop a clear picture of which candidates have a track record of meaningful achievements, and which do not. This leads to a highly valid system.