

Rationale and Research Evidence Supporting the Use of Content Validation in Personnel Assessment

Charles F. Sproule
Director, Sproule & Associates

January 2009

A monograph of the
International Personnel Assessment Council



International Personnel Assessment Council
www.ipacweb.org

Table of Contents

About the Author	3
About IPAC	3
Acknowledgements	3
Abstract.....	4
Introduction	5
Definitions.....	6
More about Content Validity	8
Why Use Content-Oriented Validation Strategies?	10
Another Viewpoint – Is Content Validity Really Validity?.....	12
Author Reactions - Content Validity and the Easter Bunny	13
Some Advantages and Disadvantages of Content Validation	16
Court Cases and Content Validation	17
Meta-Analysis Research Demonstrating the Value of Content Validation	17
A Hierarchy of Direct and Indirect Assessment Methods	18
Cumulative Validity Data on Assessment Methods.....	20
Conclusion – Validity of Direct and Indirect Assessment Methods	23
Adverse Impact.....	24
Use of Assessment Methods by Public Organizations	26
Written Job Knowledge Tests.....	28
Ratings of Training and Experience (T&E’s).....	30
Applicant Acceptance and Adverse Impact.....	32
Length of Experience and Validity	32
Specificity of Experience and Validity	33
Interviews	33
Fairness and Adverse Impact – Structured Interviews.....	37
Other Comments – Structured Interviews / Oral Examinations.....	37
Making Use of Other Validation Evidence	38
Summary – Making a Case for Content Validation	39
References.....	41

About the Author, Charles F. Sproule

Charley is currently retired. He spent the past eight years as owner and Director of Sproule & Associates, a personnel assessment consulting services organization. Sproule & Associates developed and conducted personnel assessment training programs for professional organizations, including MAPAC, IPAC, IPMA-HR, and federal, state and local agencies. Charley spent most of his career (37 years) as a personnel assessment manager with the Pennsylvania State Civil Service Commission where he managed the Research, Evaluation, and Test Development Divisions. Other experience included a part-time mobility assignment with the U.S. Civil Service Commission as a Personnel Research Psychologist, and work as a consultant to a variety of federal, state, and local agencies.

He published a variety of articles, chapters, reports, and monographs. During his career he was a member of the American Psychological Association (APA), the Society of Industrial and Organizational Psychology (SIOP), the International Personnel Management Association (IPMA), the International Personnel Assessment Council (IPAC), the Harrisburg Area Chapter of IPMA (HAIPMA), and the Personnel Testing Council of Metropolitan Washington (PTC/MW). Charley served as an IPMA Chapter President, co-founder and President of IPAC, and as the founder and first President of the Mid-Atlantic Personnel Assessment Consortium (MAPAC).

Charley wrote this Monograph as his last contribution to the assessment profession prior to retirement. The monograph is intended to promote the use of content validation. It summarizes information which demonstrates the value of content validation for commonly used assessment procedures. Charley was motivated to complete this publication after attending a PTC/MW session in Washington DC on "Content Validity" and the Easter Bunny" on March 5, 2008, and considered using an alternate title "Content Validity Rocks" for this monograph. Charley can be reached at cfsproule@verizon.net

About IPAC

The International Personnel Assessment Council (IPAC) is a professional association devoted to providing state-of-the-art information related to personnel assessment and selection.

Acknowledgements

A substantial portion of this publication is based upon information taken from a series of IPAC seminars developed for personnel assessment professionals. IPAC authorized use of the seminar materials for this publication. Charley Sproule co-developed these seminars with Nancy Abrams, Bruce Davey, and Jim Johnson. Other contributors are identified within the text of the publication. All of these individuals wrote substantial parts of the materials used in this publication. Their contributions are recognized and appreciated.

Many reviewers helped to improve earlier drafts of this publication. Some of the reviewers included: Nancy Abrams, Wanda Campbell, Tom Ramsay, Jim Frankart, and Bryan Baldwin. Others who asked not to be recognized also contributed ideas, critical comments and suggestions. This assistance was invaluable and is recognized and appreciated. The author also wishes to thank the International Personnel Assessment Council (IPAC) for making this monograph an initial publication of a very important professional group for assessment practitioners.



Abstract

This article makes a case for use of content validation in personnel assessment. It reviews content validation legal requirements, professional standards and principles for best practice. It describes why employers often rely on content validation. Content valid assessments tend to have lower levels of adverse impact and higher applicant acceptance than more general assessment methods.

Research evidence is presented to demonstrate that, across a range of assessment methods, except for general ability tests, direct assessments have higher levels of criterion-related validity than indirect assessment methods. Tests with high content validity are more job-specific and thus are more direct assessments. Research evidence is reviewed which demonstrates that more job-specific assessments have higher levels of criterion-related validity than less job-specific measures within the three most commonly used assessment methods (job knowledge tests, ratings of training and experience, and interviews). A strategy involving use of a variety of validation methods is recommended.

Introduction

Historically content validation has been relied upon by test developers and test users as a professionally acceptable method of demonstrating validity. The *Uniform Guidelines* allow test users to demonstrate validity through content validation studies. However, more recently, the professional *Standards* (1999) identify a variety of sources of validity evidence and imply that content evidence alone may be insufficient. Some academicians state that content validity is not really validity at all.

The purpose of this article is to present reasoning and research results which support the use of content validation in personnel assessment. Definitions of “validity” “validation” and content-oriented validity will be reviewed, as well as why many organizations rely on content validation. An alternative viewpoint, that content validation is not really validation at all, will be presented and discussed.

Meta-analysis validity evidence across a wide range of assessment methods will be reviewed and related to content validity. Findings for direct assessment methods will be compared to indirect methods. Data on the level of criterion-related validity of three commonly used assessment methods will be reviewed. The data presented will demonstrate that assessment methodologies supported by strong content validity evidence have high levels of criterion-related validity. This is true across assessment methods, as well as within three commonly used assessment methods. Meta-analysis data provide support for reliance on content validation. Except for general ability tests, tests which are more direct measures and tests which are more job-specific have higher levels of criterion-related validity.

Much of the information in this article is based on instructional materials from three seminars on personnel assessment developed by the International Personnel Assessment Council (IPAC). The seminar information has been supplemented with information from recent literature. The seminars were practitioner developed and based on best practices of IPAC member agencies and current research. They were initially developed between 1984 and 1993, and were updated between 2001 and 2002. The seminars are on the topics of “Planning Hiring and Promotional Assessments” (Examination Planning), “Training and Experience Ratings” (T&E’s), and “Structured Employment and Promotion Interviews” (Oral Examinations).

The IPAC personnel assessment seminars were developed by committees of IPAC members. Details on the content and development of the seminars are available from the IPAC Training Committee, and are summarized in the Instructor Manual for each seminar. Those who prepared the most recent versions of the seminars include the author, who led the update efforts, Nancy Abrams, Bruce Davey, and James Johnson. All are IPAC past-presidents. Many others contributed to the development of the seminars, including Mike Aamodt and his graduate students (Examination Planning); Ron Ash (T&E’s); and Kris Smith

(Oral Examinations). Other contributors, who included some members of the IPAC Training Committee, are identified in the Instructor Manual for each seminar.

Much of the information in this article is not new. It has been reported elsewhere. However, the information has not been organized and presented together or analyzed from a perspective which demonstrates the value of content validation.

Definitions

The *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) indicate that a variety of types of evidence is needed to establish validity. After studying the *Standards* one could conclude that content validation evidence alone may be insufficient. Following is information on “validity” and “validation” as defined in the *Standards*. Some relevant definitions from the *Uniform Guidelines on Employee Selection Procedures* (1978) and the *Principles for the Validation and Use of Personnel Selection Procedures* (2003) are also included.

The *Standards* define “validity” as: “The degree to which accumulated evidence and theory support specific interpretation of test scores entailed by proposed uses of a test.” “Validation” is defined as: “The process through which the validity of the proposed interpretation of test scores is investigated.” The *Standards* identify a variety of “sources of validity evidence” as outlined below:

Sources of Validity Evidence

Evidence based on:

- Test content
- Response processes
- Internal structure
- Relations to other variables
- Convergent & discriminant data
- Test-criterion relationships
- Validity generalization

Following is some more detail on each of the sources of validity evidence. For more complete information, see the *Standards*. Validity evidence can be based on:

- **Test content**
 - An analysis of the relationship between a test’s content and what it is intended to measure. Test content includes the content specifications as well as guidelines

and procedures for administration and scoring. Evidence based on content can come from expert judgments.

Note: This type of evidence has been called content validity in the past. In contrast to the *Standards*, the federal *Uniform Guidelines on Employee Selection Procedures* (EEOC et al., 1978, Sec. 5, p. 38298) allow test users to "... rely on criterion-related validity studies, content validity studies, or construct validity studies." The *Guidelines* state that content validation is "demonstrated by data showing that the content of a selection procedure is representative of important aspects of performance on the job."

- **Response processes**

- Studies of how individual test takers approach items
- Interrelationships among test parts, and between the test and other variables
- Studies of how observers or judges, who score examinee performance, record and evaluate data

- **Internal structure**

- Whether relationships among test components conform to knowledge of the proposed construct upon which proposed score interpretations are based.
- Studies of differential item functioning for different groups

- **Relations to other variables**

- Does the test or test part relate to other measures to which it would be expected to relate?

- **Convergent & discriminant data**

- Do test scores correlate with other measures of the same construct?
- Do test scores not correlate with measures of different constructs?
- Note: This type of evidence is a significant aspect of what was previously known as construct validation.

- **Test-criterion relationships**

Note: This type of evidence is labeled criterion-related validation in the *Uniform Guidelines* (EEOC et al., 1978). The *Uniform Guidelines* state that such evidence is: "Demonstrated by empirical data showing that the selection procedure is predictive of or significantly correlated with important elements of work behavior."

- **Validity generalization**

- Results of meta-analyses studies
- Synthetic validity studies

- Job Component validity studies
- Transportability studies
- **Consequences of testing**
Evidence about consequences may be directly relevant to validity when the consequences can be traced to a source of invalidity such as construct under-representation or construct-irrelevant components.

More about Content Validity

One section of the *Standards* (standard 14.9, p. 160) indicates that validity evidence based upon test content can serve as “the primary source of validity evidence” when a close link between test content and job content can be demonstrated. This statement implies that some additional validity evidence is needed beyond content validity evidence.

The *Principles for the Validation and Use of Personnel Selection Procedures* (Society of Industrial and Organizational Psychology, 1999, p. 4) support the *Standards* definition of validity “... as a unitary concept with difference sources of evidence contributing to an understanding of the inferences that can be drawn from a selection procedure”. The *Principles* (p. 21) state that: “Evidence for validity based on content typically consists of a demonstration of a strong linkage between the content of the selection procedure and important work behaviors, activities, worker requirements, or outcomes on the job.”

In contrast to the *Standards*, the *Principles* contain statements which indicate that users can rely on test content alone to provide validity evidence (*Principles*, 2003, p. 5). When “planning the validation effort ... the design of the study can take many forms such as single local studies” (*Principles*, 2003, p. 8).

The *Standards* say that evidence based upon “test content” or content validation is only one source of validity evidence. The same is true for criterion-related validation evidence. IPAC personnel assessment seminars advise practitioners to begin with literature review and content validation, and present a variety of types of evidence to support their assessment procedures. Examples of research evidence which supports three commonly used assessment procedures appear later in this article.

Content validity is demonstrated to the extent that the content of the assessment process reflects the important performance domains of the job. It is a validation process that can address the constraints faced by many organizations. It is a practical approach to validation. It prescribes an approach to developing the assessment process based on a study of the job. Validity is thus built into the assessment procedures. Assessment methods based on use of this strategy are usually statistically related to job performance.

The content validation process provides a rigorous scientific structure to help assure that the judgments made to plan and develop tests are appropriate. As Doctor Ebel stated: "Content validity is essential to the basic intrinsic meaning of any measure. The criterion measures needed for criterion-related validity must themselves possess content validity" (Mussio and Smith, p. 9). Thus, content validity is a prerequisite for validation.

In 1982 John E. Hunter of Michigan State University made a presentation to the International Personnel Management Association on "What is the Validity of a Content Valid Test?" He used validity generalization to estimate the criterion related validity of a content valid test. The validity was found to be "very high." Content valid job knowledge tests correlated .78 with work sample test criterion measures and .43 with supervisory rating criterion measures. These are corrected meta-analysis correlations. Hunter stated: "psychologists ... know that the validity of content valid tests is very high" (Hunter, 1982, p.12). More recent meta-analysis results will be presented later in this article.

Content validation methods focus on content relevance and content representation (Stelly and Goldstein, 2007, p. 256). Content relevance is the extent to which the tasks of the test or assessment are relevant to the target domain. Representativeness refers to the extent to which the test items are proportional to the facets of the domain. Content relevance and representativeness are commonly assessed using subject matter expert ratings.

The central element of content-oriented validation strategies is conduct of a job analysis, which collects and analyzes current, accurate, specific data from an adequate and representative sample of incumbents and supervisors. The IPAC job analysis model is based upon the scientific method. Data is collected and analyzed to answer specific questions. The model includes:

- Developing task statements in a specific format to describe the work performed
- Developing operational knowledge, skill and ability statements (KSA's) which describe the requirements needed to perform important job tasks
- Linking KSA's to job tasks
- Rating tasks and KSA's on importance and relationship to successful job performance
- Determining which tasks and KSA's are entry-level, and which are full performance
- Conducting the job analysis as a cooperative effort of assessment specialists and job experts
- Documenting the job analysis methods and findings

A variety of other features are part of the IPAC job analysis model which is described at an operational level in the Examination Planning seminar, and in a three-day MAPAC training course on *Job Analysis for Content Validation* (MAPAC, 2003), and in a one-day IPMA *Job Analysis* training course (2002). The job analysis procedures are based upon a multi-

purpose job analysis model developed by the author for an earlier IPMA Job Analysis training seminar.

Based on the job analysis results, the next step in the content validation process is preparation of a documented examination plan. The examination plan defines and includes:

- What will be assessed
- The type(s) of assessment methods to be used, including the evidence and rationale supporting these decisions
- Linkages of tasks, KSA's, and test parts
- Test weighting based upon the job analysis data
- The appropriate method of use of each assessment
- A plan for test development or test selection and test review
- A plan for the sequencing, standardized administration and objective scoring of the assessments
- A plan for establishing passing scores
- Evaluation of test effectiveness by a study of test reliability, and
- Statistical analysis of the test

Other features of the examination plan are operationalized in the IPAC Examination Planning seminar. The job analysis, examination plan, test development procedures, and steps to conduct the assessments and evaluate their effectiveness are all designed to meet legal and professional requirements. In essence, the model is an applied research methodology designed to answer a series of specific questions. The IPAC seminars present considerable substantive detail on how to establish and document content validity evidence. The above short summary does not describe the complete methodology.

An excellent discussion and comparison of the content validation requirements of the *Uniform Guidelines*, the *Standards* and the *Principles*, as well as guidance on conducting content validation studies in a rigorous scientific manner can be found in Stelly and Goldstein's 2007 chapter on content validation.

Why Use Content-Oriented Validation Strategies?

Many public sector agencies rely on content validity as their primary method of test development and validation. This is because most public sector agencies conduct job-specific testing and make use of job-related assessments, rather than using general ability tests. Litigation and consent decrees related to the adverse impact of general ability tests have moved many public agencies away from use of general ability tests. Also, most public sector agencies do not have the staff resources needed to conduct criterion-related test validation research, sample sizes are usually insufficient for the conduct of criterion-related

validation research, predictor scores are often restricted, and adequate criterion measures are often not available.

Two of the problems mentioned above which often make it too difficult to conduct criterion-related validation studies are sample size and range restriction. Small sample sizes are often due to many job classifications having only a small number of incumbents. To illustrate the problem of range restriction, most merit hiring is done in score order. Only those with high scores are typically hired. When only a few hires occur, this presents both a sample size problem and a severe range restriction problem. Criterion-related validation is not feasible when the sample size is small and the predictor score representation is overly restricted. The scores of those hired do not represent the scores of the applicant group, so we have no way to determine if those with low or moderate scores would be good or poor job performers.

Another problem concerning criterion-related validation is that adequate criterion measures are often not available. Also, readily available criterion measures may not be good measures. The most frequently used criterion measure in criterion-related validation research is supervisory ratings of job performance. A recent article on the relationship between job performance and ratings of job performance (Murphy, 2008, p. 151) states that “most reviews of performance appraisal research (e.g., Landy & Farr 1980, 1983; Murphy & Cleveland, 1991, 1995) suggest that the relationship between job performance and ratings of job performance is likely to be weak or at best uncertain.” This is another reason why many employers rely on content validation.

Content-oriented validation of assessment processes is important for a variety of reasons. One reason is to help achieve the goal of identifying the best available candidates. As illustrated later in this article, content-oriented validation methods typically possess criterion-related validity when studies of that kind are carried out. It is a practical means of meeting legal requirements and professional standards when supplemented by supporting data, and is often the only method feasible.

When using content validation it is more likely that the assessment procedures will be viewed as appropriate and fair by candidates and by hiring officials. It is also easier to explain how and why an assessment procedure is used if questions arise.

Should legal challenges occur validity documentation is central to the employer’s defense. Defending an assessment procedure without such evidence is usually difficult and unsuccessful. As one expert put it many years ago, “If you haven’t documented the steps you took to assure content-oriented validity, and available research evidence on which you relied concerning their use, you have no evidence of validity.”

Another Viewpoint – Is Content Validity Really Validity?

The author wanted to prepare this article for some time and was motivated to complete it by attendance at a half-day workshop conducted on March 5, 2008 by Kevin R. Murphy, PhD, Professor of Psychology and Information Science Technology at Pennsylvania State University. The workshop was sponsored by the Personnel Testing Council of Metropolitan Washington. It was entitled “**Content Validity and the Easter Bunny.**” Following is a sampling of the information presented on slides during the workshop. A paper copy of the slides was provided to workshop participants. Contact Doctor Murphy for more details.

Doctor Murphy has “doubts about content validity”. The following quote is from Guion, R. M. (1978) “Content Validity” in moderation. *Personnel Psychology*, 31, 205-213: “ ...there is no such thing as content validity ...” (p 212).

There is a lack of clear standards for content validation. What defines a linkage? How many linkages are enough?

There is little evidence that assessments of content validity are linked to criterion-related validity (Carrier, M.R., Dalessio, A.T., Brown, S.H., 1990) Correspondence between estimates of content and criterion-related validity values. *Personnel Psychology*, 43, 85-100.

Virtually all methods of content-oriented validation rely on judgments of the match of the link between the test and the job. When the set of plausible predictors shows positive manifold, this link does not affect validity.

When the set of plausible predictors shows positive manifold, a general factor will always emerge. All tests and criteria will correlate positively with this general factor, and therefore with each other. It is positive manifold that matters, not cognitive content. Virtually all types of widely-used predictors are positively correlated with performance and with one another.

Content matching is often irrelevant to criterion-related validity. Data shows that when there is a good match between test content and job content, test batteries show consistent validity; and when there is poor match between test content and job content, test batteries show consistent validity. Data were presented from validation studies on the Armed Services Vocational Aptitude Battery (ASVAB) to demonstrate this.

Content matching and apparent job-relatedness are important for user acceptability. It provides face validity which is one of the most important aspects of validity. Also, content matching is important and is still acceptable under the *Uniform Guidelines*. There is no sign that the *Guidelines* will change soon.

Conclusion: Content matching is important, but not for validity. If you own stock in content validity, sell it.

Author Reactions - Content Validity and the Easter Bunny

Much of Doctor Murphy's logic, which supports general ability tests rather than job-specific content valid tests, is related to Schmidt and Hunter's (1998) conclusion that "The research evidence for the validity of general mental ability measures for predicting job performance is stronger than that for any other measure." Schmidt and Hunter stated that general mental ability tests have the highest validity and the lowest cost, and can be used at all job levels. By job type they report general mental ability validity of .58 for professional and managerial jobs, .51 for mid-complexity jobs, .40 for semi-skilled jobs, and .23 for semi-skilled jobs. The more complex the job, the better general mental ability predicts performance. The Schmidt and Hunter conclusions are based upon thousands of criterion-related validity studies.

The success of general ability tests in litigation has not been as positive as the research data would appear to indicate. Cognitive ability measures are extremely likely to have adverse impact. Cognitive ability has the highest validity but the largest White/Black subgroup difference (Ployhart and Holtz, 2008). However, even though there is typically a one standard deviation difference in test performance between some groups, "... comprehensive surveys have failed and critical analysis of available studies have failed to support the hypothesis that ability tests are less valid for blacks than for whites in predicting occupational or educational performance." (Anastasi, 1988, p. 197). Also see Widgor and Garner, 1982.

Despite the above evidence of the validity of general ability tests, many employers, including the federal government and many state and local police organizations, have stopped using general ability tests because of candidate challenges and adverse impact. Many organizations have moved to other assessment methods, including content valid job-specific ability tests. Group differences on a variety of other test types, especially content valid tests, as outlined later in this article, are lower than that for general ability tests. General mental ability tests often do not appear to be job-related and they are more likely to be viewed less favorably by candidates.

Concerning the content validity aspects of Murphy's presentation, the author agrees that content validity data alone are not enough to fully support an assessment procedure. The *Standards* indicate that a variety of sources of validation evidence is needed. The author agrees that content validity data should be supplemented by other research data, such as that which will be presented later in this article for three commonly used assessment procedures.

The author disagrees with the conclusion that developing content evidence is not a method of validation, and one should sell content validation stock. In my view, content validity is “The Force” in personnel assessment. I “owned” a “content validity sector mutual fund” during my years of public sector assessment work. The “fund” paid off multiple times, when we successfully defended our tests in hearings and in federal district court. The *Standards* identify “test content” as one of the “sources of validity evidence” just as “test-criterion relationships” are identified as one of the sources of validity evidence. Both content evidence and criterion-related evidence can be used to “develop a scientifically sound validity argument.” Both methods rely on expert judgment as an integral part of the validation process. Also, as stated earlier, the *Uniform Guidelines* recognize content validation as a legally acceptable validation model. Also, the *Principles* contain a number of statements which indicate that content validation alone can be relied upon by test users.


As stated earlier, the *Standards for Educational and Psychological Testing* (1999, Standard 14.9, p. 160) state that there are circumstances where validity based upon test content can serve as “the primary source of validity evidence”. This is when “... a close link between test content and job content” can be demonstrated.

Criterion-related validation evidence relies on expert judgment to select an appropriate criterion measure; and the criterion itself is often judgmental (i.e. supervisory ratings). If expert judgments are acceptable as part of the criterion-related validation process, why are they not acceptable to establish content validity evidence? All validation evidence relies on judgments. The best that we can do is to collect as much evidence as feasible using the scientific method, and analyze and present the evidence in a fair and objective manner.

In addition, criterion-related validation is often not feasible. If content validation alone is not sufficient as implied by the *Standards* and stated by Doctor Murphy, there will be a lot of tests without adequate validity evidence.

In a “President’s Note” published on the PTC/MW web site prior to the 3/5/08 conference session described above, Martha Hennen, PTC/MW President provided some comments and observations. She quoted some conditions which Dr. Guion (1977) proposed that would provide “sufficient justification” for use of a measure based upon content validation. These include:

- a) The content domain is rooted in behavior that has a generally accepted meaning,
- b) The content domain is specified with little or no ambiguity,
- c) The content domain so specified is relevant to the purpose of measurement,
- d) Qualified judges must agree that the domain has been adequately sampled, and
- e) The response content must be reliably observed and evaluated.



So, based on the above, there are circumstances where Dr. Guion believes content validation is appropriate. Most job knowledge tests would appear to meet these conditions.

The PTC/MW President also referenced a recent book chapter (Stelly and Goldstein, 2007) which contains a chapter on content validity which “provides some concrete recommendations regarding methods for collecting and establishing content validity evidence.”

Concerning the lack of clear standards for content validation, the IPAC seminars provide a well-defined content validation model.

The IPAC seminars recommend that test developers begin with content-oriented test development, and supplement the content validation evidence with other evidence. As Chief of Research, Chief of Evaluation, and Chief of Test Development for the Pennsylvania State Civil Service Commission (in another life), our organizational approach to validation was to: 1.) consistently do content validation studies, 2.) reference and summarize relevant findings from prior research, and 3.) periodically supplement content validation studies with criterion-related studies. In addition, some test transportability studies were conducted. The criterion-related studies were conducted for job classes with substantial hiring activity. Test analysis and improvement studies, adverse impact analysis studies, and job analysis update studies were conducted periodically. Developmental research and studies of assessment methods and procedures were also conducted periodically. Pennsylvania actively participated in consortia efforts to improve staff training and assessment resources. This model for improving assessment resources and procedures, and building validity evidence is recommended to the reader.

Some Advantages and Disadvantages of Content Validation

Advantages	Disadvantages
The content validation strategy can be readily applied to full-performance level jobs where the work performed remains stable and prior job-specific preparation is required.	Content validation can be difficult to apply to entry-level jobs where there is no job-specific prior preparation required. It can be difficult to apply when assessing broad constructs.
Content validation is easier to understand and apply than more complex validation strategies. It is understandable to candidates and employers. As indicated in the next section, content validation has been found to be acceptable in court.	Some academicians have the view that content validation is not validation at all.
Content validation is an acceptable validation strategy under the <i>Uniform Guidelines</i> . The <i>Principles</i> also indicate that content validity evidence alone may be sufficient.	The <i>Standards</i> require a variety of types of evidence of validity. Content evidence alone may not be sufficient, so it should be supplemented by other types of validity evidence.
Content validation is typically more feasible than other validation strategies. For example: It can be conducted with small samples, whereas larger samples are necessary for criterion-related and construct validation studies. Content validation can be conducted when job performance measures are not readily available.	Content validation is not easy. It is a rigorous scientific method which requires time, resources, expertise, attention to detail, reliance on assessment experts and subject matter experts, and thorough documentation.
There is research evidence, as reported in later sections of this article, that measures with high content validity have high criterion-related validity.	Content validation may not be an appropriate validation strategy when job requirements change frequently, or when job requirements are not well defined.
Job-specific content valid assessments often have lower adverse impact than general ability tests.	More resources are required to develop job-specific tests for each occupation as compared to using general ability tests for a wide range of jobs.

Court Cases and Content Validation

Stelly and Goldstein (2007, p. 278-285) provide a brief history of court cases relevant to content validation, and provide useful guidance based upon the court decisions reviewed. Their review is not an extensive legal analysis, but shows that courts have found content validation to be an acceptable method for demonstrating validity. Some of the cases they reference involve application of content validation to broader KSA's. The five cases cited are:

- Guardians Association of the New York City Police Department, Inc. v. Civil Service Commission of the City of New York (1980),
- Association of Mexican-American Educators v. State of California (1996),
- Gillespie v. State of Wisconsin (1985)
- Cuesta v. State of New York Office of Court Administration (1987)
- Progressive Officers Club, Inc. v. Metropolitan Dade County (1990)

A review of what courts are saying concerning employment testing (Harris, 2008) found that the majority of court cases where employers presented validity studies relied on the content validation strategy.

The IPMA-HR Job Analysis one-day training program (2002, p. 24-26) also provides a summary of court cases relevant to job analysis and content validation. The court decisions were used as part of the process of designing the IPMA-HR job analysis model, and designing the IPAC examination planning process (IPAC, Examination Planning, 2002) for building content validity evidence which was summarized earlier in this article.

Biddle (2008) compared the *Guidelines, Standard and Principles*. He found that the *Guidelines* have been cited as the sole standard for validity review in numerous court cases. Content validation is an acceptable validation method under the *Uniform Guidelines*. However, the *Guidelines* only apply whenever a selection procedure has adverse impact. As stated by Biddle: "The professional standards embody best practice guidelines that apply to situations where adverse impact may or may not exist." It is interesting to note that federal, state and local merit system laws typically require evidence of validity and job relatedness for all selection procedures, regardless of adverse impact.

Meta-Analysis Research Demonstrating the Value of Content Validation

The next sections of this article review meta-analysis research information which provides support for content validation. First, research results for direct assessment methods will be compared to results for indirect assessment methods. Meta-analysis validity findings will be compared across a wide range of assessment methods. Then, meta-analysis validity results will be reviewed and related to content validity within three commonly used

assessment methods: written job knowledge tests, ratings of training and experience, and interviews.

A Hierarchy of Direct and Indirect Assessment Methods

Assessment procedures developed following the content validation strategy can be viewed as a hierarchy of assessment methods. By their very nature, some assessment methods produce stronger evidence that the content of the assessment procedure is representative of important aspects of job performance.

For example, the strongest evidence would be actual job performance as evidenced by productivity data or by performance evaluation information, such as evaluation of performance during a probationary period, an annual performance evaluation, or a performance evaluation conducted for research purposes. (Note that similar evidence is often used as the criterion measure in a criterion-related validation study). The next level of evidence would be measures of performance proficiency on the most important job tasks. This evidence could be obtained via performance or work sample testing. An example of a next lower level of evidence would be success on knowledge, skill or ability tests measuring KSA's which have been shown to be necessary for job performance. A still lower level of evidence would be completion of courses which prepare a person to perform relevant job tasks. A table listing examples of types of evidence, from the strongest evidence which most closely replicates actual job performance, to evidence which is less direct follows.

A Hierarchy of Assessment Evidence

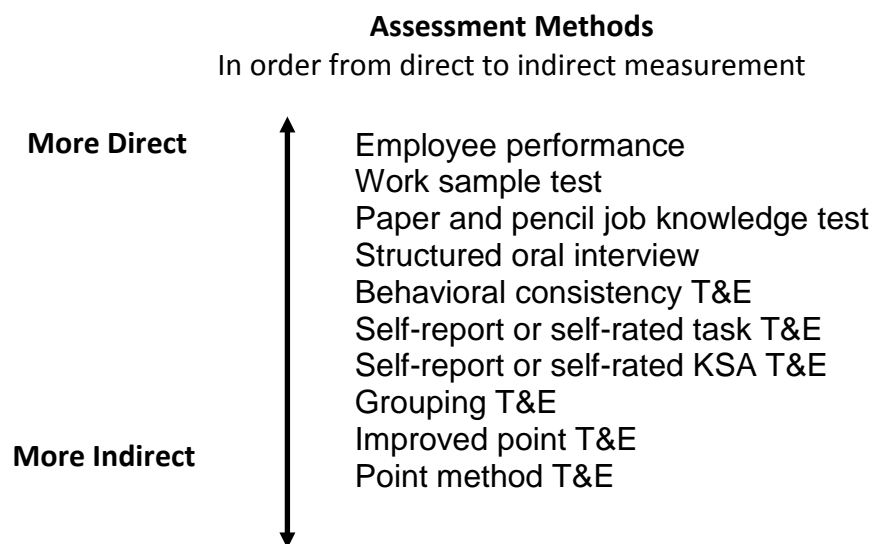
Level of Evidence	Examples of Evidence
High	<ul style="list-style-type: none"> • Actual job performance
	<ul style="list-style-type: none"> • Performance on work sample tests replicating the most important job tasks • Reports of relevant and verifiable past achievements demonstrating task or KSA proficiency
Moderate	<ul style="list-style-type: none"> • Possession of knowledges, skills and the specific abilities needed to perform job tasks • History of past performance of job tasks (i.e., the tasks were performed but we do not have information on the proficiency of the performance)
	<ul style="list-style-type: none"> • Relevant work experience
Low	<ul style="list-style-type: none"> • Possession of relevant diplomas, degrees, certifications • Completion of relevant training courses


Another way of looking at this hierarchy is from the perspective of direct vs. indirect measurement methods. Assessment methods differ in many ways. One useful way of distinguishing them is to consider the extent to which they are direct measures of the attributes we are assessing, rather than indirect indicators of attributes. Direct assessment methods require candidates to demonstrate competencies (e.g. performance tests) whereas indirect assessment methods rely on indicants of relevant competencies. Direct assessment methods possess a very high degree of content validity. It is more difficult to establish content validity evidence for indirect assessment methods, since the inferential leap between the evidence and actual job proficiency can be great.

One common example used to illustrate a content valid assessment procedure is use of a test of keyboarding for the job of Word Processing and Data Entry Specialist. A content valid keyboarding test would replicate the most important job tasks (text entry and data entry). The test would be a *direct* measure of keyboarding ability. Two *indirect* measures or indicators of keyboarding ability are completion of a high school keyboarding course, and having keyboarding work experience. The indirect measures do not inform us about the current keyboarding proficiency of the subject.

The following anecdote illustrates that one can obtain quite different results for direct and indirect measures of keyboarding proficiency. The 1985 Guinness Book of world records identifies Barbara Blackburn of Salem Oregon as holding the world record for typing speed. She could type 150 words per minute for fifty (50) minutes at a stretch. Her top speed was 212 words per minute. However, according to the Guinness book, she failed her high school typing class.

Following is a listing of examples of some commonly used assessment methods in order from those which are most the direct measures of ability to those which are indirect measures of ability.





In the illustration listing direct and indirect assessment methods, the most direct measure is employment of all candidates and assessing their actual job performance. The next most direct assessment method is job simulation or performance testing. The least direct assessment method listed is the traditional point methods of rating training and experience (T&E). This T&E method is far removed from actual job performance because it relies on indicants of relevant competencies, which are based on self-report information provided on application forms or resumes by candidates, rather than demonstration of competencies or possession of relevant KSA's.

The more direct assessments, if developed for use in assessing candidates for a particular job based upon a job analysis and the examination planning and development process described in IPAC seminars, would be supported by substantial evidence of content validity. The more direct assessment methods would be supported by stronger evidence of content validity since the test tasks would more closely replicate job tasks or assess job-specific KSA's than would be the case with the indirect assessment methods. The indirect methods only provide ability indicants.

Cumulative Validity Data on Assessment Methods

Following is a summary of the results of meta-analysis research on the validity of a variety of assessment methods, including those listed earlier. Meta-analysis is a quantitative technique which combines validity data from past research and corrects for statistical and measurement artifacts. The correlations listed are corrected meta-analysis estimates of the true validity of the measures.

Cumulative Validity Data on Assessment Methods

Assessment Method	Job Performance	Training Performance
Work Sample Tests	.54*	
General Ability Tests	.51**	.56
Structured Interviews	.51**	.35
Peer Ratings	.49	.36
Job Knowledge Tests	.48**	
Behavioral Consistency T&E	.45	
Job Tryout	.44	
Integrity Tests	.41	.38
Unstructured Interviews	.38	.35
Assessment Centers	.37**	
Biographical Data	.35**	.30
Conscientiousness Tests	.31	.30
Reference Checks	.26	.23
Self-rating KSA T&E	.20	
Years of Experience	.18	.01
Self rating Task T&E	.15	
Point Method T&E	.11	
Years of Education	.10	.20
Interests	.10	.18
Graphology	.02	

Sources: Schmidt, F., and Hunter, J. (1998). The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124(2), 262 - 274, and (for Self-rating T&E validity coefficients), McDaniel, et al., 1988.

* Note: A more recent meta-analysis of work sample tests found a corrected validity of .33 (Roth, Bobko and McFarland, 2005).

** These meta-analysis corrected correlations are the same as those reported by Ployhart and Holtz (2008).

The following is provided to help readers who are not familiar with interpreting validity coefficients.

Above .35	Very beneficial
.21 to .35	Likely to be useful
.11 to .20	Depends on circumstances
Less than .11	Unlikely to be useful

Source: U.S. Department of Labor, Employment and Training Administration (1999). *Testing and Assessment: An Employers Guide*.

Following are comments concerning some of the assessment methods listed in the table.

Work sample tests require candidates to perform tasks that closely resemble tasks performed on the job. Part of the job is thus simulated, and these methods are sometimes called job simulations or performance tests. When it is feasible to develop and to administer such tests to candidates, they are typically valid predictors of job performance as shown by the 1998 meta-analysis result of .54 for job performance prediction. Work sample tests are direct measures of ability. Note: A 2005 meta-analysis found a corrected validity of .33 for work sample tests. The earlier (Schmidt and Hunter, 1998) finding of .54 may be an overestimate of validity.

General ability tests are usually multiple-choice tests of language, mathematics, reasoning, and spatial ability. Schmidt and Hunter (1998, p. 262 & 264) state that: "...for hiring employees without previous experience in the job the most valid predictor of future performance is general mental ability...It is the best available predictor of job-related learning." The reason may be that the test "tasks" presented to candidates require cognitive abilities also required for learning and performance of job tasks. These tests are easy and inexpensive to administer. They are less job specific than many other assessment methods, and therefore are not usually supported by content validity evidence. Candidates may not be able to see a clear relationship between the test and the job. These tests often have higher adverse impact than some alternative measures.

The validity of **structured interviews** is quite high (.51) for predicting job performance. A structured interview is carefully developed, based on a job study, and all candidates are asked to respond to the same questions. The questions may present a problem to be solved, or ask the candidates how they have handled a situation in the past (the behavioral consistency approach). Note that it is much more successful in predicting job performance than the **unstructured interview** (.38). More details on interview research findings will be presented later in this article.

The validity of **job knowledge tests** (.48) is in the same range. These tests are usually multiple-choice questions requiring candidates to demonstrate their knowledge of job-relevant subject matter. Though more difficult to develop than general ability tests, they are often used by public organizations and for professional licensure and certification. Job knowledge tests can be custom developed or commercially purchased. As will be reported later in this article, job knowledge tests which have high job specificity, have higher levels of criterion-related validity. Job knowledge tests can not be used for entry-level jobs. They are not appropriate for use with jobs where no prior experience is required or where no prior job-specific training is required. They are easy and inexpensive to administer.

The validity of **behavioral consistency T&E methods** (.45) is nearly as high as that of job knowledge tests, although fewer studies of this method have been carried out. This method is based on the principle that the best predictor of future performance is past

performance. Applicants describe their past achievements and the achievements are rated by subject matter experts. The IPAC Training and Experience Rating Seminar (2001) provides guidance and instruction on how to develop and conduct behavioral consistency T&E's.

For the **self-rating and self-report T&E methods** (.15 - .20), the validities depicted are difficult to interpret not only because of the few studies available on these methods, but because they vary considerably in the T&E procedures used. More recent research summarized later in this article indicates how these T&E methods can be improved.

Years of experience (.18) appear to be a modest predictor of job performance. Some additional data related to experience measures will be presented later in this article. Years of experience are a very indirect measure of ability.

Although there have been many studies of the traditional **point method T&E's**, the cumulative validity evidence (.11) does not support their use.

Conclusion – Validity of Direct and Indirect Assessment Methods

Review of the meta-analysis results, and comparison to the list of direct and indirect assessment methods, leads to the conclusion that, except for general ability tests, the predictive value of assessment methods reflects the extent to which they more directly assess applicant competencies.

The three most direct assessment methods listed earlier were: work sample tests, job knowledge tests and structured interviews. Reviewing the earlier "Cumulative Validity Data on Assessment Methods", these methods had correlations for predicting job performance of .54 (.33 in a 2005 meta-analysis), .48, and .51, respectively. The three most indirect assessment methods for which data are available from meta-analysis are: task based T&E's, KSA based T&E's, and point method T&E's. Their correlations with job performance were .15, .20, and .11 respectively.

These data indicate that direct assessment methods have higher levels of criterion-related validity than indirect assessment methods. This is evidence that the stronger the content validity evidence supporting an assessment method, the more likely it is that the assessment method will have a high level of criterion-related validity. In the author's view, the meta-analysis criterion-related validity data provides support for the content validation model. It is recognized that the strength of content validation evidence involves a lot more than the nature of the assessment device. For example, the strength of the content validity evidence also includes the clarity of the definition of the content domain, and the specificity of the point-to-point linkages between each job task and KSA and the relevant section(s) and items of the assessment device. Additional research is needed to more fully explore this matter.

The one glaring exception (the fly in the ointment) to direct measures having higher criterion-related validity and indirect measures having lower criterion-related validity for prediction job performance is general ability tests (r of .51 with job performance). Ployhart and Holtz's (2008) review of assessment procedure validity found that cognitive ability measures have "... the highest validity but the largest White-Black subgroup differences." As summarized in the following section, general ability tests are likely to have adverse impact. Also, general ability tests are most appropriate for use at entry-level, they are generally not appropriate to use for jobs where prior job-specific preparation is required. When prior job-specific preparation is required, content valid tests are more appropriate and acceptable to users and applicants than general ability tests.

To maximize validity but minimize subgroup differences, an effective strategy recommended by Ployhart and Holtz is to use alternative predictor measurement methods instead of cognitive ability tests. Examples of alternatives they recommend include interviews, situational judgment tests and biodata. "Using alternative predictor measurement methods will reduce subgroup differences because they measure multiple cognitive and non-cognitive KSA's, frequently minimize reading requirements, may engender more favorable reactions, and/or are based on job performance tasks for which subgroup differences are smaller" (Ployhart and Holtz, 2008, Table 2, p. 158).

Adverse Impact

One consideration in selecting an assessment method is the degree of adverse impact expected. Following is a summary of two reviews of effect sizes. One review (Schmidt, Clause and Pulakos, 1996) is by "ability" and the other (Ployhart and Holtz, 2008) is by "predictor." The two reviews differ to some extent in their findings but the overall pattern of findings is consistent.

Effect size is a measure of the difference in the average scores of sub-groups measured in standard deviation units. For example, on cognitive ability, the first table reports a difference of -.83. This means that the average score of African-Americans is typically .83 standard deviation units below the average score of whites on cognitive ability measures. In the second table, the effect size of .99 for the White-Black comparison means that whites typically score .99 standard deviations above blacks on cognitive ability.

Subgroup Effect Sizes by Ability
 (African-American/White Comparisons)
 Source: Schmidt, N., Clause, C.A., and Pulakos (1996)

Ability	Weighted Effect Size
Cognitive Ability	-.83
Spatial Ability	-.66
Math Ability	-.64
Verbal Ability	-.55
Mechanical Comprehension	-.40
Job Sample/Job Knowledge	-.38
Accomplishment Record	-.33
Interview	-.15
Clerical Speed/Accuracy	-.15
Manual Dexterity	-.14
Personality	-.09

Subgroup Effect Sizes by Predictor
 Source: Ployhart and Holtz (2008)

Predictor	White-Black	White-Hispanic	White-Asian	Male-Female
Cognitive Ability	.99	.58 to .83	-.20	.00
Job Knowledge	.48	.47		
Spatial Ability	.66			
Biodata	.33			
Structured Interview	.23			
Accomplishment Record	.24*			.09
Work Sample	.52	.45		
Assessment Center	.60**			

* White-Minority comparisons ** .60 or less, depending upon the exercise.

Note that the effect sizes for the White-Black and White-Hispanic comparisons on work samples, job knowledge tests, interviews, biodata, and the accomplishment record are much lower than the effect sizes for abilities typically measured by general ability tests (cognitive ability, spatial ability, math ability, and verbal ability). There is very limited data and very small differences by gender. The White-Asian differences on cognitive ability are smaller than for the other race-ethnic groups.

Despite the effect size differences for ability tests, an analysis of the fairness of ability testing by the National Academy of Sciences concluded that there is no substantial evidence that ability tests are unfair to minorities (Wigdor, A., and Garner, W. Eds., 1982). Ability tests accurately predict job performance regardless of race. However, employers may still want to consider the above data, as well as applicant acceptance, when deciding which type(s) of tests to use, especially if diversity is an organizational objective.

Use of Assessment Methods by Public Organizations

Organizations vary in size, functions, available resources, and a host of other factors. It is helpful to know what “others” are doing, and we’ll now review some information on that question.

The International Public Management Association for Human Resources (IPMA-HR) and the National Association of State Personnel Executives (NASPE) conducted a “benchmarking” survey in 2000-2001. The 177 responding organizations were federal, state, and local agencies. Most were cities, counties, and states.

Respondents were asked to identify their three most frequently used “testing/selection” methods. The methods most often used were “written tests of job knowledges”, “T&E evaluations”, “oral exams,” and “resume screens,” followed by “assessment centers,” “written general aptitude tests,” and “skills inventories.” Following is the percentage of agencies surveyed who identified each assessment method as their most frequently used testing/selection method.

<u>Assessment Method</u>	<u>% using most frequently</u>
Written Tests (Job Knowledge)	81%
T&E Ratings	70%
Oral Exams	70%
Resume Screens	66%
Assessment Centers	44%
Written General Aptitude Tests	43%
Skill Inventories (paper-based)	39%
Personality Tests	25%
Skills Inventories (computer-based)	24%
Computerized written exams	24%
Other	23%

Note: "Other" included physical ability tests, typing and data entry tests, and other methods.

One hundred seventy-seven (177) federal, state, and local organizations responded to the survey, with the following composition:

Cities	45%
States	20%
Counties	16%
Other	13% (towns, special districts, federal agencies, schools, etc.)

Based upon the 2000-2001 survey, the three most commonly used public sector assessment methods were:

- Written tests (primarily job knowledge tests)
- Ratings of Training and Experience (T&E's)
- Oral Examinations

A "2006 Recruitment and Selection Benchmarking" survey by IPMA-HR, which was sponsored by NEOGOV, found that the above three assessment methods continued to be commonly used. Job Knowledge tests were used by 78% of the 236 survey respondents, T&E's by 72%, and Oral Examinations by 65%.

We will now review data for each of the three commonly used assessment procedures. We will review data related to the following question: Does higher levels of content validity evidence lead to higher levels of criterion-related validity for these three assessment methods?

Written Job Knowledge Tests

Eighty-one percent (81%) of public sector jurisdictions surveyed in 2000 – 2001 stated that written tests (primarily tests of job knowledge) were one of their three most frequently used assessment procedures.

The 1998 Schmidt and Hunter validity review summarized earlier found that the validity of job knowledge tests for predicting job performance was .48. This is a very respectable level of criterion-related validity and was the fifth highest level of validity reported.

The 1988 *Annual Review of Psychology* contains a chapter on "Personnel Selection and Placement" which gives this *Summary of the Validity of Written Tests*: "many kinds of predictors can be useful ... but abilities have the best track record. ... cognitive tests are likely to be good predictors of job performance ... evidence continues to be reported ... what seems clear is that **knowledge is causally related to performance** whichever measure is used, and that better performance can be expected if people are selected who either have the knowledge from experience or the aptitude for acquiring it" (Guion and Gibson, 1988, p. 363 & p. 365).

It is rare for Psychologists to speak about test - job performance relationships in other than probabilistic terms. The above statement, that higher levels of knowledge cause higher levels of job performance, is strong evidence to support use of the content validation model for the development of job knowledge tests.

Details on the validity of job knowledge tests are reported in a meta-analysis study (Dye, Reck, and McDaniel, 1987). The study summarized and analyzed the results of previous research. The previous studies were based upon 363,528 persons and included 502 validity coefficients. This study found a corrected mean validity of .45 for job performance and .47 for training success. Validities for predicting job performance were higher for high complexity jobs (.57) and when job-test similarity was high (.62). Validity for prediction training success was higher for high complexity jobs (.57) and when job-test similarity was high (.76). Note that the .62 and .76 validity finding exceed the levels of validity reported for any of the assessment methods reviewed in the 1998 Schmidt and Hunter article.

Validity of Job Knowledge Tests

Source: Dye, Reck, and McDaniel (1987)

Job Knowledge Tests	Corrected r's job performance	Corrected r's training success
All job knowledge tests reviewed	.45	.47
Tests for high complexity jobs	.57	.57
Tests for low complexity jobs	.39	.46
Tests with high test-job similarity	.62	.76
Tests with moderate test-job similarity	.35	.49
Tests with low test-job similarity	.35	.46

The research indicates that: "when tests of job knowledge are to be used there is much to be gained by developing them to be job specific" (Dye, Reck and McDaniel, 1987, p. 9.). Greater validity was found for high complexity jobs. "For the combined moderator effect (job complexity and test-job similarity) a job-specific test is always superior to an off-the-shelf test." Results of the research indicate that the validity of job knowledge tests does generalize. Two other reviews of research also found that job knowledge as measured by written tests play a significant role in job performance (Schmidt, Hunter and Outerbridge, 1986; Hunter and Hunter, 1983).

Written job knowledge tests with high job similarity have a high level of both face validity and content validity. The above results demonstrate that use of content validity as the basis for the development of job knowledge tests is likely to result in high levels of criterion-related validity.

Three advantages of job knowledge tests are high candidate acceptance, efficiency, and breadth. Since job knowledge tests are typically job-specific, applicants tend to like them because they appear to be job related (i.e. they have face validity). Multiple choice and true/false tests are extremely efficient and inexpensive when testing large numbers of candidates. They may not be cost-effective to develop for small candidate groups. Also, written tests can measure many different knowledges for a wide range of jobs. In addition, well-developed job knowledge tests have high reliability.

Concerning adverse impact, the effect sizes for minority groups reported for job knowledge tests (Schmidt, Clause and Pulakos, 1996; Ployhart and Holtz, 2008) are about half the size of those found for cognitive ability tests. This indicates that job knowledge tests are likely to have much lower adverse impact than cognitive ability tests.

A review by Ployhart and Holtz (2008) found that one effective way to balance the need to maximize validity and minimize adverse impact is to assess the full range of knowledges, skills and abilities. Job knowledge tests can assess a wide array of job requirements and thus contribute to this recommended strategy. The review also found that another way to help maximize validity and minimize adverse impact is to minimize the verbal ability requirements of the predictor. This can be done with job knowledge tests by making sure that the verbal level of the test materials and instructions do not exceed the verbal ability requirements of the job. Job analysis can be used to determine the verbal ability level requirements of the job.

Job knowledge tests are generally not appropriate for entry-level jobs where no prior job-specific training or experience is required. Some organizations use job-specific trainability tests or situational judgment tests for entry-level jobs as an alternative to use of general ability tests. Situational judgment tests are usually paper and pencil multiple-choice tests. They can be computer administered, as can job knowledge and general ability tests. Nguyen, McDaniel and Whetzel (2001) report that situational judgment tests correlate well with general ability tests, and have less racial impact. Job knowledge tests, situational judgment tests and job-specific trainability tests have higher face validity than general ability tests.

Ratings of Training and Experience (T&E's)

Seventy percent (70%) of 177 public jurisdictions surveyed in 2000 – 2001 stated that ratings of training and experience were one of their three most frequently used assessment procedures. An earlier public sector agency survey (Cook, 1980) reported that T&E's were used more than any other selection device except written tests.

T&E's are know as "unassembled examinations" since candidates do not need to report for testing. Traditional T&E methods, such as the point method, rely on information from an application form or resume, while other methods are based on responses to structured questionnaires designed to elicit the information necessary to evaluate applicants for a specific job. Some public jurisdictions collect the needed information using structured questionnaires which are completed over the internet to speed up the assessment process.

T&E's are difficult to use for entry-level jobs, but can readily be used for a wide range of jobs where prior training or experience is required. The IPAC T&E seminar describes a variety of methods for rating training and experience. The most common methods are listed below. To the right of the methods, when available, is the estimated true validity from the earlier table summarizing the 1998 Schmidt and Hunter meta-analysis research, and the 1988 McDaniel et al. meta-analysis research.

T&E Rating Methods	Corrected Meta-analysis Validity Data
1. Holistic Methods	<i>No Data</i>
2. Traditional Point Methods	.11
3. Improved Point Methods	<i>No Data</i>
4. Grouping Methods	<i>No Data</i>
5. Self-Report and Self-Rating Methods	.15 (task), .20 (KSA)
6. Behavioral Consistency Methods	.45

Schneider’s (1994) literature review reported the validity of point methods as ranging from .11 to .15, task-based methods from .15 to .28, and behavioral consistency methods from .45 to .49.

The validity levels reported for T&E methods are related to the T&E method being used. Criterion-related validity increases as the content validity evidence for the T&E procedure increases. The lowest level of validity reported (.11) is for the “traditional point method.” In this method, points are awarded for education and experience, which are indirect indicators of competencies. The next higher level of criterion-related validity reported (.15 to .20) is for task and KSA based T&E methods where self-report or self-rating information is credited for each relevant task or KSA. The self-report and self-rating task and KSA T&E methods can be supported by content validity evidence. The highest level of T&E method validity was for the behavioral consistency T&E method (.45). In this method specific and verifiable achievements of candidates are evaluated. This method is similar to job performance evaluation based on descriptions of job-related accomplishments. The behavioral consistency method is most appropriate for higher level jobs.

A review of the validity of T&E rating methods by James Johnson (IPAC T&E Seminar Participant Manual, 2001), A Summary of Research on T&E Methods of Assessment concluded that:

“The validity studies of T&E ratings clearly support use of some methods. Most strongly supported are the self-report and self-ratings methods, and the behavioral consistency methods. When appropriately developed and used, these methods may predict performance as well as other, more thoroughly studied methods including objective tests and structured oral interviews. It is clear, however, that all methods must be developed through use of a sound job analysis.

Use of traditional point methods is not supported by research evidence. It is possible, however, that future research on use of the “improved point methods” and “grouping methods” will be more promising. Use of job analyses, appropriate job experts, and the research results and models outlined here (in the IPAC T&E seminar) should be helpful.

As has been found for development and use of structured interviews, careful preparation of instructions, rating scales, and content (e.g., task statements, definitions) is likely to be critical to assure reliability and credibility of the process, as well as validity. It is well known that a poorly designed test or interview is a poor predictor of job performance; the same principle applies to development and use of T&E rating methods.”

Applicant Acceptance and Adverse Impact

Ratings of training and experience are well accepted by job applicants. This may be because the methods have high face validity. A study by Stone (1989) found that employer use of application information was perceived by applicants to be low on a scale of invasiveness. The IPMA-HR/NASPE study (2001) found that T&E methods are well accepted by using public agencies.

The author is not aware of any court case where a rating of training and experience examination procedure has been challenged.

A review of 253 adverse impact studies (Pennsylvania State Civil Service Commission, 1994) which were conducted between 1982 and 1993, and which were based upon close to 400,00 applicants and 30,000 hires, found that: “The Experience & Training (T&E) Rating type of exam was consistently adverse impact free.”

An IPAC review of information on T&E’s for development of the Examination Planning seminar found some T&E methods (i.e. those based on credentials) may have adverse impact while other T&E methods have low impact. Schmidt, Clause and Pulakos (1996) report relatively low adverse impact against African Americans for the behavioral consistency method. The effect sizes reported for the accomplishment record by Ployhart and Holtz (2008) are also relatively low. The accomplishment record effect size they reported is about one quarter of the effect size of cognitive ability measures.

Length of Experience and Validity

McDaniel, Schmidt and Hunter (1988) concluded that job experience is more highly correlated with job performance when the average amount of experience among the applicants is relatively low. They found that beyond five years of experience there is little increase in performance. This has implications for T&E crediting plans.

Specificity of Experience and Validity

Quinones, Ford and Teachout (1995) clarified the nature of relationships between experience and job performance in the largest meta-analysis study to-date on this topic. The results are based on a meta-analysis of 44 studies, which included 25,911 participants. Validity data on experience was analyzed by breaking down past studies into those which reviewed experience at the organizational level, the job level, and at the task level. The level of validity found varied depending upon the way in which experience was defined as shown below.

<u>Specificity of Experience</u>	<u>Validity</u>	<u># of studies</u>
Task Level	.41	6
Job Level	.27	30
Organization Level	.16	8

Several conclusions are implied by these results: The more specific the experience to the target job, the greater its predictive value. Experience measured at the task level is substantially more valid in predicting job performance than is experience measured at the job level or experience in an organization. These data provide support for the use of content validation. Experience measures which are more job-specific had a higher level of criterion-related validity. Evaluating the relevance of work based on tasks is more precise than evaluating experience by job or organizational level.

Quinones *et. al.* (1995) also found higher validity coefficients in studies using “hard” rather than “soft” criterion measures (.39 and .24, respectively). They found much higher correlations when experience is defined as *frequency* with which a task has been performed (.43) in contrast to *amount of time* spent performing tasks (.27). Amount of time spent performing tasks is far less predictive of job performance than asking how often the tasks have been performed. These results have significant implications for designing T&E experience crediting methods.

Interviews

Seventy percent (70%) of 177 public sector jurisdictions surveyed in 2000 – 2001 stated that oral examinations were one of their three most frequently used testing methods. An earlier survey (IPMA, 1988) of International Personnel Management Association agency members found that 76% of the responding 389 public sector jurisdictions and agencies were using structured oral examinations. The structured oral examination procedure was the selection procedure used by the largest percentage of respondents. Twelve selection procedures were included in the 1988 survey.

Interviews can be unstructured or structured. Oral examinations are structured interviews and are developed following a content validation process.

The employment interview may be the most common selection procedure in use, along with the application form. A private sector survey on the use of the interview by 852 firms in 1957 found 99% using it (Huett, 1976).

Research in the 1970's and earlier generally found low criterion-related validity for the interview (Huett, 1976). Reviews of the interview in the industrial-organizational psychology literature in the early 1980's expressed hope for the structured interview (Arvey and Campion, 1982).

Later research (Whetzel, McDaniel and Schmidt, 1985) analyzed interview reliability and validity data by type of interview using meta-analysis procedures. Job related interviews were better predictors than psychological interviews. Structured interviews were better predictors than unstructured interviews. The highest levels of validity were found for structured interviews. The highest validity generalization result was .51 for job-related structured interviews with job performance criteria collected for research purposes. This was based on 10 correlations with 978 subjects. More recent research investigating the criterion-related validity of structured and unstructured interviews is summarized below.

Validity of Unstructured Interviews

Uncorrected correlation range of .11 to .18

Corrected correlation range of .20 to .33

Study	Number of Coefficients	Uncorrected Correlation	Corrected Correlation
Weisner/Cronshaw, 1988	87	.17	.31
Wright et al., 1989	13	.14	
Huffcutt/Arthur, 1994	114	.11	.20
McDaniel, et al., 1994	145	.18	.33

Unstructured interviews generally have lower criterion-related validity than structured interviews. For example, Hunter and Hunter (1983) found an average validity for unstructured interviews of .14. Williamson et al. (1997) summarized meta-analyses that reported validities ranging from .11 to .18 (.20 to .33 corrected) for unstructured interviews.

Validity of Structured Interviews

Uncorrected correlation range of .24 to .34

Corrected correlation range of .35 to .62

Study	Number of Coefficients	Uncorrected Correlation	Corrected Correlations
Weisner/Cronshaw, 1988	87	.34	.62
Wright, et al., 1989	13	.27	.35*
Huffcutt/Arthur, 1994	114	.34	.57
McDaniel, et al., 1994	45	.24	.44
Schmidt and Hunter, 1998	(85 years of research)	n/a	.51

* Corrected for reliability only

Comparing the two data tables on interviews, the range of corrected correlations for unstructured interviews is .20 to .35, and the range for structured interviews is from .35 to .62. This is clear evidence that structured interviews have a much higher level of criterion-related validity than unstructured interviews. Structured interviews are developed and conducted using a content validation process. No such process is used for unstructured interviews. These findings provide further support for use of the content validation process.

Research reviews since the mid to late 1980's have consistently found that carefully developed and structured oral examinations have high criterion-related validity.

Wiesner and Cronshaw (1988) reported on a meta-analysis of 150 studies. 48 of these studies, with 10,080 subjects, were based upon structured interviews. The mean validity for structured interviews was .34 uncorrected and .62 corrected. Wiesner and Cronshaw state: "... The validity coefficients of structured interviews both individual and board, are comparable with the best predictors available ..." (Wiesner and Cronshaw, 1988, p. 286).

Following is a conclusion from a 1990 review of research on structured interviews: "Recent research indicates that well-developed, carefully administered structured oral examinations, based on job analysis information, using job-related questions, specific and anchored rating scales, and well-trained raters have high reliability and a level of validity comparable to that of cognitive tests, and show less adverse impact." (Sproule, 1990, p. 68).

McDaniel, Whetzel, Schmidt and Maurer (1991) found a true mean validity for structured interviews of .46, and .50 for structured situational interviews.

McDaniel et al. (1994) report a corrected validity of .44 for structured orals.

Williamson et al. (1997) summarize meta-analyses that reported validities ranging from .24 to .34 (.44 to .62 corrected).

Schmidt and Hunter (1998) report a corrected true validity of .51 for structured interviews. This is equal to their validity estimate for general mental ability tests.

A review of research on the interview (U.S. Merit System Protection Board, 2003, p. 4) concluded that: "Research indicates that unstructured interviews are, on average, little more than half as effective as structured interviews and unstructured interviews may be subject to bias and challenges." Unfortunately, the review also found that use of unstructured interviews was more common in federal agencies, and "use of structured interviews appears to be the exception rather than the norm."

The public sector has been using oral examinations (which are structured interviews) for decades. For example, the author began work in test development with the Pennsylvania State Civil Service Commission in the early 1960's and oral examinations had been used routinely by many state, city, and county merit system agencies for decades prior to that. In 1975 the Pennsylvania State Civil Service Commission gave structured oral examinations to 3,457 candidates for 95 job classes (Moreano and Sproule, 1976).

In 1976, well before the finding of high validity for structured interviews by the industrial-organizational psychology profession, Dennis Huett identified the following typical merit system selection interview components:

- Conduct of a systematic job analysis to identify the important job elements
- Development of a standard set of questions linked to each job element
- Use of questions which elicit factual, verifiable information on actual behaviors
- Development of evaluation forms for use by interviewers in recording and summarizing their observations
- Development of precise standards of evaluation
- Communicate the important elements of the job to the interviewers
- Use multiple interviewers
- Train the interviewers in questioning techniques, use of rating forms and how to avoid errors
- Have interviewer's rate candidates independently. Avoid comparisons of candidates to one another
- Evaluate the results of the interviews and take action to correct inconsistency

Many of the above steps are typical components of the process for developing, conducting and evaluating content valid assessment procedures. Research has demonstrated that this content validation process results in high criterion-related validity for structured interviews.

Two recommended references on structured interview procedures are the IPAC Oral Examination Seminar Participant Manual (2001), and the California State Personnel Board's Manual of Theory and Practice on *Development and Use of Structured Interviews* (Willihnganz and Langan, 1998).

Fairness and Adverse Impact – Structured Interviews

A major advantage of structured interviews is their fairness and lack of adverse impact. A review by Reilly and Warech (1988) concluded: "The very limited data suggest no general unfairness of the interview towards minorities, females or other individuals. The data also indicate that interviews will probably have less adverse impact than cognitive tests."

As reported earlier, Schmidt, Clause and Pulakos (1996) found an effect size of interview scores in African-American and White comparisons of $-.15$. This was a much smaller difference than found for seven other types of measures, including written ability and knowledge tests, and job sample tests. Only "personality" tests had a smaller difference ($-.09$). Ployhart and Holtz (2008) reported an effect size of $.23$ for White-Black comparisons of interview scores. This was one of the smallest effect sizes reported for the predictors they reviewed.

A comparison of alternative predictor methods which minimize racioethnic and sex subgroup differences (Ployhart and Holtz, 2008) found that using interviews was one of the most effective methods for reducing subgroup differences. They referenced research by Huffcutt and Roth, 1998 which found that Blacks and Hispanics score about one-quarter of a standard deviation lower than Whites on interviews.

Other Comments – Structured Interviews / Oral Examinations

Another advantage of structured interviews developed and conducted using the content validation model is that they have high face validity. Also, interviews are well accepted by candidates. There have been few lawsuits related to the validity of well-developed and well-conducted structured interviews.

Structured interviews can be relatively easy to develop but expensive and time consuming to administer. The most frequent problem areas related to oral examinations are their administration and security. Users of oral examinations need to take special care to maintain the security of oral examination materials. Raters need to be consistent in their

questioning of candidates, and consistent in applying the rating standards. Raters must not discuss candidate performance outside the examination room

Any oral examination becomes expensive if the number of candidates tested is large because the process must be individually administered and scored. For small candidate groups, the process can be inexpensive because test development and administration can be done relatively quickly compared to other measures.

Oral examinations are very versatile. A wide range of assessment formats can be used within the oral examination setting, including case problems, situational problems, behavioral consistency questions, oral presentations, and other formats. Behaviors involving interaction can be readily assessed. The format versatility allows better matching of candidate behaviors to required job behaviors, and can increase content validity.

Making Use of Other Validation Evidence

The last few sections of this article presented a summary of meta-analysis validation research for a variety of assessment methods. There is a wide range of other methods to collect or develop evidence of validity which readers are encouraged to pursue. A few examples follow.

McDaniel (2007) provides guidance on use of validity generalization information as a validation strategy. His chapter on “Validity Generalization as a Test Validation Approach” contains references and information on research for assessment methods not covered in this article (e.g., integrity tests, personality tests, customer service orientation measures, psychomotor tests, etc.). The McDaniel chapter also contains references on the results of cumulative reviews of research for specific occupations (e.g., law enforcement, firefighters, computer programmers, clerical occupations, petroleum industry jobs). McDaniel provides suggestions for interpreting validity generalization information and using existing research to support your test development and validation efforts. Readers are encouraged to make use of a variety of existing research in addition to conducting their own validation research. This approach will help users to meet *Standards* professional practice guidance to provide a variety of sources of validation evidence to help support their assessment methods.

Tippens and Macey (2007) provide guidance on use of consortium studies as a validation strategy. Joining with other organizations to plan and conduct cooperative studies is yet another way to improve assessments and do more with less.

In a previous publication (Sproule, 1980) guidance was provided on a resource allocation strategy for public personnel selection. One part of that strategy related validation methods to needs and situations. Although this article has focused on content validation, other validation methods are often necessary and appropriate. For example, construct and criterion-related validation was determined to be necessary in Pennsylvania for the

development of physical ability tests and medical standards for entry-level Corrections Officers (Berkley and Sproule, 2001), whereas content validation was determined to be appropriate for use in developing a written test of job-specific abilities. The written test included video-based sections. Because of very high hiring activity for entry level Corrections Officers, criterion-related test validation studies were conducted. Readers should keep in mind that content validation evidence may need to be supplemented by other evidence in some situations.

Summary – Making a Case for Content Validation

Content validation is acceptable as a method for demonstrating validity under the federal *Uniform Guidelines on Employee Selection Procedures* (1978). It is also acceptable as a source of validity evidence under the *Standards for Educational and Psychological Testing* (1999) and the *Principles for the Validation and Use of Personnel Selection Procedures* (2003). The *Guidelines* and the *Principles* indicate that content validation evidence alone may be sufficient evidence of validity.

The content validation process provides a rigorous scientific structure to help assure that the judgments made to plan and develop tests are appropriate. It is a practical method of test validation which should be used along with other validation methods. Content valid tests have high applicant and user acceptance. Job-specific tests make logical sense to candidates. Content validation has been accepted in court. Other validation methods are often not feasible or practical, and the validation process with other methods can be confusing to test users. Content valid tests have less adverse impact than general ability tests.

Research evidence demonstrates that tests with high content validity have high criterion-related validity. Testing methods which more directly assess job performance are supported by stronger evidence of content validity. Evidence from meta-analysis research demonstrates that tests which more directly assess job requirements (e.g., work sample tests, structured interviews, job knowledge tests, behavioral consistency ratings of training and experience) have high criterion-related validity for predicting job performance. In addition, these direct measures have lower adverse impact than general ability tests. Except for general ability tests, assessment methods which are more indirect measures of job requirements have lower levels of criterion-related validity.

Public organizations use three types of assessment methods most frequently. For the three commonly used methods (job knowledge tests, ratings of training and experience, and interviews), there is a variety of evidence to demonstrate that procedures which are more job-related, and which more directly assess what is being measured, have higher levels of criterion-related validity. This is further support for the use of content validation.

For job knowledge tests, cumulative research evidence shows that job knowledge tests with high test-job similarity have very high criterion-related validity for predicting job performance (.62) and training success (.76). These are levels of validity beyond what has been found for general ability tests. As with general ability tests, validity for job knowledge tests increases as job complexity increases.

For ratings of training and experience (T&E's), the T&E methods which most directly assesses achievements (behavioral consistency) has the highest level of T&E validity (.45), and the method which only assesses indicators of ability (T&E point method) has the lowest level of validity (.11). When specificity of experience is analyzed, crediting experience at the most specific level (i.e. task level) results in higher levels of validity (.41) than crediting experience at the job level (.27) or organizational level (.16). All of these findings provide support for content validation, since the strongest evidence of content validity exists for more direct and more job-specific T&E assessments.

Research on interviews shows dramatic differences in validity for structured interviews (range of corrected r 's .35 to .62) as compared to unstructured interviews (range of corrected r 's .20 to .33). The validity of structured interviews is about twice as great as the validity of unstructured interviews. Structured interviews are developed and conducted following the content validation model. This is further support for the value of content validation.

Assessment methods developed following the content validation methodology generally have high applicant and test user acceptance.

Test users are encouraged to routinely conduct content validation studies, supplement those studies with information from the research literature, and periodically conduct local criterion-related validation research when feasible. IPAC offers training courses for practitioners (i.e., Examination Planning, Ratings of Training and Experience, and Structured Interviews) to help agencies develop adequate content validity evidence for their assessment procedures, and to rely on supporting research. IPMA-HR offers a one-day training course on Job Analysis. The Mid-Atlantic Personnel Assessment Consortium (MAPAC) also offers training courses to help practitioners (i.e., Job Analysis for Content Validation, Item Writing for Selection Specialists, and Essential Statistics for Employee Selection Specialists). Some agencies have combined the six IPAC and MAPAC courses into a training curriculum for their assessment staff. This training is highly recommended.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999) *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Anastasi, A. (1988). *Psychological Testing*. Sixth Edition. New York: Macmillan Publishing Company.

Arvey, R.D. , Champion, J.E. (1982). The Employment Interview A Summary and Review of Recent Research. *Personnel Psychology*. 35, 281-322.

Berkley, Steven, Sproule, Charles F. (2001) The Selection of Entry-Level Corrections Officers: Pennsylvania Research. *Public Personnel Management*. Vol. 30, No. 3. International Personnel Management Association.

Biddle, Daniel A. (2008). Are the *Uniform Guidelines* Outdated? Federal Guidelines, Professional Standards, and Validity Generalization (VG). *The Industrial Organizational Psychologist*. Vol. 45, No. 4, pp. 17-23.

Cook, Cindy Lorentson (1980). Rating Education, Training and Experience in the Public Sector. *Proceedings of the 1980 International Personnel Management Association Assessment Council Conference on Public Personnel Assessment*. Chicago, Illinois.

Dye, D. A, Reck, M., McDaniel, M.A. (1987). *Moderators of the Validity of Job Knowledge Measures*. U.S. Office of Personnel Management. (OED Report 87-10) Washington, DC. Republished in the *International Journal of Selection and Assessment*, July 1993, Vol. 1, No. 3.

Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, and Department of Labor. (1978) *Uniform Guidelines on Employee Selection Procedures*. 41 CFR Part 603.

Guion, R. M., Gibson, W. M. (1988). Chapter on "Personnel Selection and Placement" *Annual Review of Psychology*. 39: 349-374.

Guion, R. M. (1977). Content Validity – The Source of My Discontent. *Applied Psychological Measurement*. 1 (1), 1-10.

Harris, Michael M. (2008). Employment Testing: What the Courts are Saying. *PTC Quarterly*. Washington, DC: Personnel Testing Council of Metropolitan Washington. Vol. IV, #2, p.5-7.

Huett, Dennis L. (1976). *Improving the Selection Interview in a Civil Service Setting*. Great Lakes Assessment Council and International Personnel Management Association. Chicago, Illinois.

Hunter, John E. (1982). *What is the Validity of a Content Valid Test?* June 7, 1982 presentation to the International Personnel Management Association. Minneapolis, Minnesota.

Hunter, John E., Hunter, Ronda F. (1983). *The Validity and Utility of Alternative Predictors of Job Performance*. Washington, DC: U.S. Office of Personnel Management, Office of Personnel Research and Development, OPRD-83-4.

Huffcutt, A. I., Arthur, W. (1994). Hunter and Hunter revisited: Interview Validity for Entry Jobs. *Journal of Applied Psychology*. 79, 184-190.

International Personnel Management Association. (2002). *Job Analysis*. Participant Manual and Instructor Manual. Alexandria, Virginia

International Personnel Management Association Assessment Council. (2001). *Training and Experience Rating (T&E) Seminar*. Participant Manual and Instructor Manual. Alexandria, Virginia

International Personnel Management Association Assessment Council. (2001). *Oral Examination Seminar*. Participant Manual and Instructor Manual. Alexandria, Virginia

International Personnel Management Association Assessment Council. (2002). *Examination Planning Seminar* (Planning Hiring and Promotional Assessments). Participant Manual and Instructor Manual. Alexandria, Virginia

International Personnel Management Association (2001). *2000/2001 IPMA/NASPE Benchmarking Report: Recruitment and Selection*. Alexandria, VA.

International Public Management Association for Human Resources (IPMA-HR) (2006). *Recruitment and Selection Benchmarking*. Sponsored by NEOGOV. Alexandria, VA.

Mid-Atlantic Personnel Assessment Consortium. (2003). *Job Analysis for Content Validation*. Participant and Instructor Manuals. Baltimore, Maryland.

McDaniel, Michael A. (2007). Validity Generalization as a Test Validation Approach. . In S.M. McPhail (Ed.), *Alternative Validation Strategies: Developing New and Leveraging Existing Validity Evidence* (p. 159-180). San Francisco: Jossey-Bass.

McDaniel, M. A., Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, 9, 103-113.

McDaniel, Michael A., Schmidt, Frank L., Hunter, John E. (1988). A Meta-Analysis of the Validity of Methods for Rating Training and Experience in Personnel Selection. *Personnel Psychology*. 41 (2), 283-314.

McDaniel, M.A., Schmidt, F.L., and Hunter, J.E. (1988b). Job Experience Correlates of Job Performance. *Journal of Applied Psychology*. 23, 327 -330.

McDaniel, Michael A., Whetzel, Deborah L., Schmidt, Frank L., Maurer, Steven D. (1994). The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis. *Journal of Applied Psychology*. Vol. 79, No. 4, 599-616.

Moreano, Augusto G., Sproule, Charles F. (1976). *Reliability and Other Data on Structured Oral Examinations*. Pennsylvania State Civil Service Commission. Bureau of Examinations. Harrisburg, PA

Murphy, Kevin R. (2008). *Content Validity and the Easter Bunny*. Workshop conducted for the Personnel Testing Council of Metropolitan Washington (PTC/MW). Washington, DC

Murphy, Kevin R. (2008). Explaining the Weak Relationship between Job Performance and Ratings of Job Performance. *Industrial and Organizational Psychology*. Vol. 1, # 2, pp. 148 – 160.

Mussio, S. J., Smith, M.K. *Content Validity: A Procedural Manual* (undated). Great Lakes Assessment Council (GLAC), and International Personnel Management Association (IPMA). Chicago, Illinois.

Pennsylvania State Civil Service Commission, Bureau of Personnel Assessment, Research Division (1994). *Adverse Impact Analysis: Summary of Previous Studies*. Harrisburg, PA

Ployhart, Robert E., Holtz, Brian R. (2008). The Diversity-Validity Dilemma: Strategies for Reducing Racioethnic and Sex Subgroup Differences and Adverse Impact in Selection. *Personnel Psychology*. 61 (1), 153-172.

Quinones, M.A., Ford, J.K., Teachout, M.S. (1995). The Relationship between Work Experience and Job Performance: A Conceptual and Meta-Analytic Review. *Personnel Psychology*. 48, 887-910.

Reilly, Richard R., Warech, Michael A. (1988). *The Validity and Fairness of Alternatives to Cognitive Tests*. Stevens Institute of Technology. Paper prepared for the National Commission on Testing and Public Policy.

Roth, P. L., Bobko, P., McFarland, L. A. (2005). A Meta-analysis of Work Sample Test Validity. *Personnel Psychology*. 58, 1009-1037.

Schmidt, N., Clause, C.A., Pulakos (1996) "Subgroup differences Associated with Different Measures of Some Common Job Relevant Constructs." In C. L. Cooper and I.T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology*. New York: Wiley.

Schmidt and Hunter. (1998). The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings. *Psychological Bulletin*. Vol. 124, No. 2, 262-274.

Schmidt, Hunter and Outerbridge (1986). Impact of Job Experience and Ability on Job Knowledge, Work Sample Performance, and Supervisory Ratings of Job Performance. *Journal of Applied Psychology*. 71 (3) 432-439.

Schneider, Robert E. (1994). *The Rating of Experience and Training: A Review of the Literature and Recommendations on the Use of Alternative E T Procedures*. Personnel Assessment Monograph. Vol. 3, No. 1, Alexandria, VA: International Personnel Management Association Assessment Council.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the Validation and Use of Personnel Selection Procedures*. (4th ed.) Bowling Green, OH

Sproule, Charles F. (1980). A Strategy for Resource Allocation in Public Personnel Selection. *Public Personnel Management*. Vol. 9, No.3. International Personnel Management Association.


Sproule, Charles F. (1990). *Recent Innovations in Public Sector Assessment*. Personnel Assessment Monograph. Vol. 2, No. 2, Arlington, VA: International Personnel Management Association Assessment Council.

Stelly, Damian J., Goldstein, Harold W. (2007) Application of Content Validation Methods to Broader Constructs. In S.M. McPhail (Ed.), *Alternative Validation Strategies: Developing New and Leveraging Existing Validity Evidence* (p. 252-316). San Francisco: Jossey-Bass.

Tippens, Nancy T., Macey, William J. (2007). Consortium Studies. In S.M. McPhail (Ed.), *Alternative Validation Strategies: Developing New and Leveraging Existing Validity Evidence* (p. 233-251). San Francisco: Jossey-Bass.

U.S. Merit System Protection Board ((2003). *The Federal Selection Interview: Unrealized Potential*. Washington, D.C.

U.S. Department of Labor, Employment and Training Administration (1999). *Testing and Assessment: An Employer's Guide to Good Practices*. Washington, D.C.



Whetzel, D. L., McDaniel, M. A., Schmidt, F. L. (1985). *The Validity of Employment Interviews: A Review and Meta-Analysis*. U.S. Office of Personnel Management, Office of Examination Development. OED 87-9, Washington, DC.

Wiesner, William H., Cronshaw, Steven F. (1988). A Meta-Analytic Investigation of the Impact of Interview Format and Degree of Structure on the Validity of the Employment Interview. *Journal of Occupational Psychology*. 61, 275-290.

Williamson, Laura Gollub, Campion, James E., Malos, Stanley B., Roehling, Mark V., Campion, Michael A. (1997). Employment Interview on Trial: Linking Interview Structure With Litigation Outcomes. *Journal of Applied Psychology*. Vol. 82, No. 6, 900-912.

Wigdor, A., Garner, W. (Eds.) (1982). *Ability Testing: Uses, Consequences, and Controversies*. Committee on Ability Testing, National Research Council, National Academy of Sciences. Washington, D.C.: National Academy Press.

Willihnganz, M.A., Langan, S.A. (1998) *Development and Use of Structured Employment Interviews: A Manual of Theory and Practice*. Sacramento, California State Personnel Board.

Wright, Patrick M., Lichtenfels, Philip A., Pursell, Elliot D. (1988). *The Structured Interview: Additional Studies and a Meta-Analysis*. University of Notre Dame, Center for Research in Business, Research Paper Series 88-07.

