PERSONNEL PSYCHOLOGY
1975, 28, 563–575.

A QUANTITATIVE APPROACH TO CONTENT VALIDITY[1]

C. H. LAWSHE

Purdue University

CIVIL rights legislation, the attendant actions of compliance agencies, and a few landmark court cases have provided the impetus for the extension of the application of content validity from academic achievement testing to personnel testing in business and industry. Pressed by the legal requirement to demonstrate validity, and constrained by the limited applicability of traditional criterion-related methodologies, practitioners are more and more turning to content validity in search of solutions. Over time, criterion-related validity principles and strategies have evolved so that the term, "commonly accepted professional practice" has meaning. Such is not the case with content validity. The relative newness of the field, the proprietary nature of work done by professionals practicing in industry, to say nothing of the ever present legal overtones, have predictably militated against publication in the journals and formal discussion at professional meetings. There is a paucity of literature on content validity in employment testing, and much of what exists has eminated from civil service commissions. The selection of civil servants, with its eligibility lists and "pass-fail" concepts, has always been something of a special case with limited transferability to industry. Given the current lack of consensus in professional practice, practitioners will more and more face each other in adversary roles as expert witnesses for plaintiff and defendant. Until professionals reach some degree of concurrence regarding what constitutes acceptable evidence of content validity, there is a serious risk that the courts and the enforcement agencies will play the major determining role. Hopefully, this paper will modestly contribute to the improvement of this state of affairs (1) by helping sharpen the content

[1] A paper presented at Content Validity II, a conference held at Bowling Green State University, July 18, 1975

validity concept and (2) by presenting one approach to the quantification of content validity.

A Conceptual Framework

Jobs vs. curricula. Some of our difficulties eminate from the fact that the parallel between curriculum content validity and job content validity is not a perfect one. Generally speaking, an academic achievement test is considered content valid if and when (a) the curriculum universe has been defined (called the "content domain") and (b) the test adequately samples that universe. In contrast, the job performance universe and its parameters are often ill defined, even with careful job analysis. What we can do is isolate specific segments of the job performance universe. In this paper, a *job performance domain*[2] is defined as: an identifiable segment or aspect of the job performance universe (a) which has been operationally defined and (b) about which inferences are to be made. Hence, a particular job may have a single job performance domain; more often jobs have several. For example, the job of Typist might have a single job performance domain, i.e., "typing straight copy from rough draft." On the other hand, the job of Secretary might have a job performance universe from which can be extracted several job performance domains, only one of which is "typing straight copy from rough draft." The distinction made here is that, in the academic achievement field we seek to define and sample the entire universe; in the job performance field we sample a job performance domain which *may* or *may not* approximate the job performance universe. More often it is not the total universe but rather is a segment of it which has been identified and operationally defined.

The nature of job requirements. If a job truly requires a specific skill or certain job knowledge, and a candidate cannot demonstrate the possession of that skill or knowledge, defensible grounds for rejection certainly exist. For example, a retail clerk may spend less than five percent of the working day adding the prices on sales slips; however, a candidate who cannot demonstrate the ability to add whole numbers may defensibly be rejected. Whether or not we attempt to sample other required skills or knowledges is irrelevant to this issue. Similarly, it is irrelevant that other aspects of the job (other job performance domains) may *not* involve the ability to add whole numbers.

The judgment of experts. Adkins[3] has this to say about judgments in discussing the content validity approach:

[2] The author is aware that certain of his colleagues prefer the term "job content domain." The term, "job performance domain" is used here (a) to distinguish it from the content domain concept in achievement testing and (b) to be consistent with the *Standards*, pp. 28–29.

[3] Dorothy C. Adkins, as quoted in Mussio and Smith, p. 8.

> In academic achievement testing, the judgment has to do with how closely test content and mental processes called into play are related to instructional objectives.

> In employment testing, the content validation approach requires judgment as to the correspondence of abilities tapped by the test with abilities requisite for job success.

The crucial question, of course, is, "Whose judgment?" In achievement testing we normally use subject matter experts to define the curriculum universe which we then designate as the "content domain." We may take still another step and have those experts assign weights to the various portions of a test item budget. From that point on, content validity is established by demonstrating that the items in the test appropriately sample the content domain. If the subject matter experts are generally perceived as true experts, then it is unlikely that there is a higher authority to challenge the purported content validity of the test.

When a personnel test is being validated, who are the experts? Are they job encumbents or supervisors who "know the job?" Or, are they psychologists or other professionals who are expected to have a greater understanding of the organization of human personality and/or greater insight into "what the test measures?" To answer these questions requires a critical examination of job performance domains and their characteristics.

The nature of job performance domains. The behaviors constituting job performance domains range all the way from behavior which is directly observable, through that which is reportable, to behavior that is highly abstract. The continuum extends from the exercise of simple proficiencies (i.e., arithmetic and typing) to the use of higher mental processes such as inductive and deductive reasoning. Comparison of the behavior elicited by a test to behavior required on a job involves little or no inference at the "observation" end; however, the higher the level of abstraction, the greater is the "inferential leap" required to demonstrate validity by other than a criterion-related approach. For example, it is one thing to say, "This job performance domain involves the addition of whole numbers; Test A measures the ability to add whole numbers; therefore, Test A is content valid for identifying candidates who have this proficiency." It is quite another thing to say, "This job performance domain involves the use of deductive reasoning; Test B purports to measure deductive reasoning; therefore, Test B is valid for identifying those who are capable of functioning in this job performance domain." At the "observation" end of the continuum, where the "inferential leap" is small or virtually nonexistent, sound judgments can normally be made by incumbents, supervisors, or others who can be shown to "know the job." The more closely the behavior elicited by the test approximates a true "work sample" of the job performance domain, the more competent are people who know the job to assess the content validity of the test. When a job knowledge test is under consideration, they are similarly competent to judge whether or not knowledge of a given bit of job information is relevant to the job performance domain.

Construct validity. On the other hand, when a high level of abstraction is involved and when the magnitude of the "inferential leap" becomes significant, job incumbents and supervisors normally do not have the insights to make the required judgments. When these conditions obtain, we transition from content validity to a construct validity approach. Deductive reasoning, for example, is a psychological "construct." Professionals who make judgments as to whether or not deductive reasoning (a) is measured by *this test* and (b) is relevant to *this job performance domain* must rely upon a broad familiarity with the psychological literature. To quote the "Standards",[4]

> Evidence of construct validity is not found in a single study; judgments of construct validity are based upon an accumulation of research results.

An operational definition. Content validity is the extent to which communality or overlap exists between (a) performance on the test under investigation and (b) ability to function in the defined job performance domain. In summary, content validity analysis procedures are appropriate *only* when the behavior under scrutiny in the job performance domain falls at or near the "observation" end of the continuum; here, those who "know the job" are normally competent to make the required judgments. However, when the job behavior approaches the abstract end of the continuum, a construct validity approach is indicated; job incumbents and supervisors are normally not qualified to judge. Operationally defined, content validity is: the extent to which members of a *Content Evaluation Panel* perceive overlap between the test and the job performance domain. Such analyses are essentially restricted to (1) simple proficiency tests, (2) job knowledge tests, and (3) work sample tests.

Measuring the Extent of Overlap

Content evaluation panel. How, then, do we determine the extent of overlap (or communality) between a job performance domain and a specific test? The approach outlined here uses a *Content Evaluation Panel* composed of persons knowledgeable about the job. Best results have been obtained when the panel is composed of an equal number of incumbents and supervisors. Each member of the *Panel* is supplied a

[4] *Standards for Educational and Psychological Tests,* p. 30.

number of items, either prepared for the purpose or constituting a "shelf" test. Independent of the other panelists, he is asked to respond to the following question for each of the items:

> Is the skill (or knowledge) measured by this item
> —Essential
> —Useful but not essential, or
> —Not necessary
> to the performance of the job?

Responses from all panelists are pooled and the number indicating "essential" for each item is determined.

Validity of judgments. Whenever panelists or other experts make judgments, the question properly arises as to the validity of their judgments. If the panelists do not agree regarding the essentiality of the knowledge or skill measured to the performance of the job, then serious questions can be raised. If, on the other hand, they *do* agree, we must conclude that they are either "all wrong" or "all right." Because they are performing the job, or are engaged in the direct supervision of those performing the job, there is no basis upon which to refute a strong consensus.

Quantifying consensus. When all panelists say that the tested knowledge or skill is "essential," or when none say that it is "essential," we can have confidence that the knowledge or skill *is* or *is not* truly essential, as the case might be. It is when the strength of the consensus moves away from unity and approaches fifty-fifty that problems arise. Two assumptions are made, each of which is consistent with established psychophysical principles:

—Any item, performance on which is perceived to be "essential" by more than half of the panelists, has some degree of content validity.

—The more panelists (beyond 50%) who perceive the item as "essential," the greater the extent or degree of its content validity.

With these assumptions in mind, the following formula for the *content validity ratio* (CVR) was devised:

$$\mathrm{CVR} = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

in which the n_e is the number of panelists indicating "essential" and N is the total number of panelists. While the CVR is a direct linear transformation from the percentage saying "essential," its utility derives from its characteristics:

—When fewer than half say "essential," the CVR is negative
—When half say "essential" and half do not, the CVR is zero

TABLE 1
Minimum Values of CVR and CVR_t
One Tailed Test, $p = .05$

No. of Panelists	Min. Value*
5	.99
6	.99
7	.99
8	.75
9	.78
10	.62
11	.59
12	.56
13	.54
14	.51
15	.49
20	.42
25	.37
30	.33
35	.31
40	.29

—When all say "essential," the CVR is computed to be 1.00, (It is adjusted to .99 for ease of manipulation).

—When the number saying "essential" is more than half, but less than all, the CVR is somewhere between zero and .99.

Item selection. In validating a test, then, a CVR value is computed for each item. From these are eliminated those items in which concurrence by members of the *Content Evaluation Panel* might reasonably have occurred through chance. Schipper[5] has provided the data from which Table 1 was prepared. Note, for example, that when a *Content Evaluation Panel* is composed of fifteen members, a minimum CVR of .49 is required to satisfy the five percent level. Only those items with CVR values meeting this minimum are retained in the final form of the test. It should be pointed out that the use of the CVR to reject items does *not* preclude the use of a discrimination index or other traditional item analysis procedure for further selecting those items to be retained in the final form of the test.

The content validity index. The CVR is an item statistic that is useful in the rejection or retention of specific items. After items have been identified for inclusion in the final form, the *content validity index* (CVI) is computed for the whole test. The CVI is simply the mean of the CVR values of the retained items. It is important to emphasize that

[5] The author wishes to acknowledge the contribution of Dr. Lowell Schipper, Professor of Psychology, Bowling Green State University at Bowling Green, Ohio, who did the original computations that were the basis of Table 1.

the *content validity index* is not to be confused with a coefficient of correlation. Returning to the earlier definition, the CVI represents the extent to which perceived overlap exists between capability to function in a defined job performance domain and performance on the test under investigation. Operationally it is the average percentage of overlap between the test items and the job performance domain. The following sections of this paper present examples in which these procedures have been applied.

Example No. 1; Basic Application

Background. This first example was developed in a multi-plant, heavy industry in which the skilled crafts are extremely important. The objective was to generate validity evidence on one or more tests for use in the preliminary screening of candidates who will be given further consideration for selection as apprentices in the mechanical and electrical crafts.

Job analysis. One facet of the job analysis resulted in the identification of 31 mathematical operations which are utilized by apprentices in the performance of their job duties. Of these operations, 19 are sampled in a commercially available test, the *SRA Arithmetic Index,* which is the subject of this discussion. Not discussed in this paper is another test, tailor-made to sample the remaining 12 operations, which was validated in the same manner.

The Content Evaluation Panel. The *Content Evaluation Panel* consisted of 175 members of subpanels, one from each of 17 plants, which were composed of (1) craft foremen, (2) craftsmen, (3) apprentices, (4) classroom instructors of apprentices, and (5) apprentice program coordinators. Each panel member was supplied with (a) a copy of the *SRA Arithmetic Index* and (b) an answer sheet. The "essentiality" question presented earlier was modified by changing the second response, "useful but not essential", to "useful but it can be learned on the job." Each panelist indicated his response on the answer sheet for each of the 54 problems.

Quantifying the results. The number indicating that being able to work the problem was "essential" when the individual enters apprenticeship was determined for each of the 54 items, and the *content validity ratio* was computed for each. These CVR values are tabulated in Table 2; all but one satisfies the 5% level discussed earlier. When the mean of the values is computed, the *content validity index* for the total test is .67; when the one item is omitted it is .69.

Example No. 2; Modified Application

Applicability. In the first example, individual test items were evaluated against the job performance domain as an entity; CVR values

TABLE 2
Distribution of Values of CVR for 54 Items In the SRA Arithmetic Index

CVR	*f*
90–94	13
85–89	12
80–84	3
75–79	0
70–74	1
65–69	2
60–64	7
55–59	0
50–54	1
45–49	2
40–44	2
35–30	2
30–34	3
25–29	3
20–24	2
15–19	0
10–14	0
5–9	0
0–4	1*
	N = 54

* Not significant.

were determined, and the CVI for the total test was computed. The modified approach used in Example No. 2 is an extension of the author's earlier work in synthetic validity (Lawshe and Steinberg, 1955) and may be used:

—When the job performance domain is defined by a number of statements of specific job tasks, and

—When the test under investigation approaches "factorial purity" and/or has high internal consistency reliability.

Clerical Task Inventory. This study utilized the *Clerical Task Inventory*[6] (CTI) which consists of 139 standardized statements of frequently performed office and/or clerical tasks of which the following are examples:

16. Composes routine correspondence or memoranda, following standard operating procedures.

38. Checks or verifies numerical data by recomputing original calculations with or without the aid of a machine.

[6] The *Clerical Task Inventory* is distributed by the Village Book Cellar, 308 State Street, West Lafayette, Indiana 47906. This is an updated version of the earlier publication "Job Description Check-list of Clerical Operations."

Virtually any office or clerical job or position can be adequately described by selecting the appropriate task statements from the inventory.

Tests utilized. Tests utilized were the six sections of the Purdue Clerical Adaptability Test[7] (PCAT): 1. Spelling; 2. Arithmetic Computation; 3. Checking; 4. Word Meaning; 5. Copying; and 6. Arithmetical Reasoning.

Establishing CVR_t Values. The *Content Evaluation Panel* was composed of fourteen employees, half of whom were incumbents who had been on the job at least three years and half of whom were supervisors of clerical personnel. All were considered to be broadly familiar with clerical work in the company. Each member, independently, was supplied with a copy of the CTI and a copy of the Spelling Section of the PCAT. He was also provided with an answer sheet and asked to evaluate the essentiality of the tested skill against each of the 139 tasks, using the same question presented earlier. The number of panelists who perceived performance on the Spelling Section as *essential* to the performance of task No. 1 was determined, and the CVR_t[8] was computed. Similarly, spelling CVR_t values were computed for each of the other 138 tasks in the inventory. This process was repeated for each of the other five sections of the PCAT. The end product of this activity, then, was a table of CVR_t values (139 tasks times six tests). All values which did not satisfy the 5% level as shown in Table 1 (i.e., .51) were deleted. The remaining entries provide the basic data for determining the content validity of any of the six sections of the PCAT for any clerical job in the company.

Job description. In this company, each clerical job classification has several hundred incumbents, and employees are expected to move from position to position within the classification. Seniority provisions apply to all persons within the classification. The job description for each classification was prepared by creating a Job Analysis Team for that classification. In the case of Clerk/Miscellaneous (here referred to as Job A), the team consisted of eight analysts,[9] four incumbents and four supervisors, all selected because of their broad exposure to the job. Each analyst was supplied with a copy of the *Clerical Task Inventory* and was asked to identify each task which ". . . in your judgment is included in the job." He was also asked to "add any tasks which are not listed in the inventory." One analyst added one task which was

[7] The *Purdue Clerical Adaptability Test* is distributed by the University Book Store, 360 State Street, West Lafayette, Indiana 47906.

[8] The CVR_t differs from the CVR only in the notation. The "t" designates "task."

[9] These eight analysts, plus six for another clerical job, constitute the 14 member *Content Validity Panel* discussed earlier.

later withdrawn by consensus. The data were then collated in the personnel office; each task that was checked by one or more analysts was identified, and the following procedure was followed:

—A list of those tasks identified by *all* analysts was prepared.

—A list was prepared of those tasks identified by all but one, another for those checked by all but two, etc.

—All analysts for the job were convened under the chairmanship of a personnel department representative.

In the meeting, members of the team considered those tasks previously identified by all members of the group, and they further defined the CTI standardized statements by adding examples which (a) designated company forms processed, (b) enumerated kinds of data treated, or (c) otherwise supplied specific information which added "in house" meaning to the statements. This process yielded a job description consisting of 47 tasks for Job A to which was appended the following certificate:

> Each of the undersigned independently analyzed the classification of Job A and selected those tasks in the *Clerical Task Inventory* which he considered to be present in the job.
>
> We then met as a group and discussed each task. By consensus, we identified those tasks which make up the attached list as the true content of the job. We also agreed on the specific example that is listed with each task.
>
> We individually and collectively certify that, in our opinion, the content of the job, is adequately and fairly represented by the attached document.

This document, signed by the eight members of the Job Description Team, becomes a part of the "compliance review trail," if and when such review is conducted.

The Content validity index. The two procedures produced (a) a table of CVR_t values for each test perceived as being relevant to each task in the CTI and (b) the list of the 47 tasks constituting Job A. In order to determine the *content validity index* for each test it is necessary to (a) identify those tasks (called determinants) which have significant CVR_t values for that test and (b) compute the mean. For example, of the 47 tasks constituting the defined job performance domain of Job A, seven tasks had significant CVR_t values (i.e., greater than .51) for the Computation Test. They are shown in Table 3 along with their respective CVR_t values and the resulting CVI of .92. The CVI values for all of the tests for Job A are shown in Table 4, column 2. Examination of column 1 in Table 4 shows that the number of task determinants

TABLE 3

Task Determinants in Job A for the CVI of the Computation Test

Task No.	Clerical Task Determinants	CVR_t
1	Makes simple calculations such as addition or subtraction with or without using a machine.	.99
2	Performs ordinary calculations requiring more than one step,, such as multiplication or division, without using a machine or requiring the use of more than one set or group of keys on a calculating machine.	.99
4	Balances specific items, entries, or amounts periodically with or without using a machine.	.99
45	Determines rates, costs, amounts, or other specifications for various types of items, selecting and using tables or classification data.	.99
3	Performs numerous types of computations including relatively complicated calculations involving roots, powers, formulae, or specific sequences of action with or without using a machine.	.83
38	Checks or verifies numerical data by recomputing original calculations with or without the aid of a machine.	.83
39	Corrects or marks errors found in figures, calculations, operation forms, or record book data by hand or using some type of office machine or typewriter.	.83
CONTENT VALIDITY INDEX (Mean CVR_t)		.92

ranges from a low of three, for Word meaning and Copying, to seven for Computation.

Minimum requirements. It is important to emphasize that the *number* of tasks for which each test is perceived as essential is irrelevant. When the *content validity ratio* (CVR_t) for a unitary test has been computed for each of a number of job tasks which constitute the job performance domain, the development of an index of content validity (for the total job performance domain) must recognize a fundamental fact in the nature of job performance:

TABLE 4

Content Validity Index Values for Six Tests for Job A

Test Section	No. of Determinants (1)	CVI (2)	Weighted Mean (3)
1. Spelling	4	.87	.87
2. Computation	7	.92	.95
3. Checking	5	.73	.79
4. Word Meaning	3	.72	.71
5. Copying	3	.94	.96
6. Arithmetical Reasoning	4	.87	.89

> If a part of a job requires a certain skill, then this skill must be reflected in the personnel specifications for that job. *The fact that certain other portions of the job may not similarly require this skill is irrelevant.*

In other words, that portion of the job making the greatest skill demand establishes the skill requirement for the total job. Theoretically, that single task with the highest CVR_t establishes the requirement for the job. The problem is that we do not have a "true" *content validity ratio* for each task. Instead, we have CVR_t values which are estimates, or approximations, arrived at through human judgments known to be fallible. The procedure utilizes the CVR_t values of a *pool* of tasks in order to minimize the impact of any inherent unreliability in the judgments of panelists.

The Weighting Issue

Weighting procedures. In any discussion of job analysis, the subject of weighting inevitably arises. It naturally extends into content validity analysis considerations. Mussio and Smith (1973) use a five point "relevance" rating. Drauden and Peterson (1974) use "importance" and "usefulness" ratings and, in addition, use ratings of "time spent" on the various elements of the job. These procedures, and similar weighting systems, no doubt have a certain utility when used solely for job *description* purposes. For example, they help job applicants and others to better understand the nature of the job activity. However, the rating concept is not compatible with the content validity analysis procedures presented in this paper; the rationale rests upon both logical considerations and empirical evidence.

Logical considerations. Perhaps when the job performance domain is viewed as a reasonable approximation of the job performance universe, some justification exists for incorporating evaluations of importance or time spent. However, if the definition of job performance domain presented in this paper is accepted, such justification is markedly diminished. To repeat, if arithmetic, or some other skill, is essential to functioning in a defined job performance domain, estimates of relevance or time spent are not germaine, i.e., the sales clerk who spends less than 5% of the work day adding numbers on a sales slip. Of course, judgments of panelists as to whether or not a specific item of knowledge or a specific skill is "essential" may be regarded as ratings. However, the question asked and the responses elicited from panelists are not oriented to relative amounts or degrees, but rather assume discrete and independent categories.

Empirical evidence. Despite these arguments, an experimental weighting procedure was incorporated in the original design of the study reported in Example No. 2. When the Job Description Team was convened, in addition to identifying the 47 tasks which constitute the job, team members also reached a consensus on the "five most important" tasks and "five next most important" tasks making up the job description. Experimental content validity index values were computed for each test using the following weights: "most important," 3; "next most important," 2; and all others, 1. The resulting weighted means appear in Table 4, column 3. There seem to be no practical differences between the weighted and the unweighted results. This outcome which was replicated in other jobs in the study confirms and further reinforces the author's earlier position that most weighting schemes are not worth the candle and that, ". . . very often, the statistical nicety of present methods suggests or implies an order of precision which is not inherent in the data. Psychological measurements, at this point in time, are quite unreliable; to suggest otherwise by using unwarranted degrees of statistical precision is for the psychologist to delude others, and perhaps to delude himself (Lawshe 1969)."

REFERENCES

Drauden, G. M. and Peterson, N. G. *A domain sampling approach to job analysis.* Test Validation Center. St. Paul, Minn. 1974.

Lawshe, C. H. Statistical theory and practice in appllied psychology. PERSONNEL PSYCHOLOGY, 1969, 22, 117–123.

Lawshe, C. H. and Steinberg, M. C. Studies in synthetic validity I: An exploratory investigation of clerical jobs. PERSONNEL PSYCHOLOGY 1955, 8, 291–301.

Mussio, S. J. and Smith; M. K. *Content validity: A procedural manual.* International Personnel Association, Chicago, 1973.

Standards for Educational and Psychological Tests, American Psychological Association, Washington, 1974.