Psychological Assessment
1995, Vol. 7, No. 3, 238-247

1040-3590/95/$3.00

Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods

Stephen N. Haynes
University of Hawaii at Manoa

David C. S. Richard
University of Hawaii at Manoa
and Department of Veterans Affairs, Honolulu

Edward S. Kubany
Department of Veterans Affairs, Honolulu

This article examines the definition, importance, conceptual basis, and functional nature of content validity, with an emphasis on psychological assessment in clinical situations. The conditional and dynamic nature of content validity is discussed, and multiple elements of content validity along with quantitative and qualitative methods of content validation are reviewed. Finally, several recommendations for reporting and interpreting content validation evidence are offered.

Psychological assessment[1] has an important impact on many clinical judgments. It provides data for the development of causal models for behavior disorders, for the design of intervention programs, for the prediction of future behavior, and for the evaluation of treatment effects. Clinical judgments are strongly influenced by the construct validity of the assessment instruments that provide the data on which the judgments are based (Haynes, 1994; Korchin, 1976; Weiner, 1976). This article addresses one component of construct validity—content validity.

We will examine the definition, importance, conceptual basis, and functional nature of content validity in psychological assessment, with an emphasis on the application of psychological assessment in clinical judgment situations. The relevance of content validity for all assessment methods and its conditional nature will also be emphasized. We will present an array of elements that are appropriate targets of content validation and stress both quantitative and qualitative methods. Finally, we will offer recommendations for reporting and interpreting content validation evidence.

Introduction to Content Validity

Definition and Components of Content Validity

Many definitions of content validity have been published (e.g., *Standards for educational and psychological testing*, 1985; Anastasi, 1988; Messick, 1993; Nunnally & Bernstein, 1994; Suen, 1990; Walsh, 1995).[2] Although worded differently, most of these definitions encompass concepts embodied in the following definition: *Content validity* is the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose.

Several components of this definition need to be defined and are also addressed in greater detail in subsequent sections of this article. The term *assessment instrument* is meant to reflect the applicability of content validity for all assessment methods (see footnote 1).

The term *elements*, of an assessment instrument, are all the aspects of the measurement process that can affect the obtained data. For example, the elements of questionnaires include individual items, response formats, and instructions. The elements of behavioral observation include observation codes, time-sampling parameters, and the situations in which observation occurs.

Stephen N. Haynes, Department of Psychology, University of Hawaii at Manoa; David C. S. Richard, Department of Psychology, University of Hawaii at Manoa, and Department of Veterans Affairs, Honolulu; Edward S. Kubany, Pacific Center for Posttraumatic Stress Disorder, Department of Veterans Affairs, Honolulu.

Correspondence concerning this article should be addressed to Stephen N. Haynes, Department of Psychology, University of Hawaii at Manoa, Honolulu, Hawaii 96822. Electronic mail may be sent via Internet to sneil@uhunix.uhcc.hawaii.edu.

[1] *Psychological assessment* refers to the systematic measurement of a person's behavior. It incorporates measurement strategies and targets and the inferences and clinical judgments derived from the obtained measures. Psychological assessment includes many assessment paradigms, such as behavioral assessment and personality assessment, many assessment methods, such as direct observation and self-report questionnaire, and many assessment instruments, such as self-report questionnaires on depression, psychophysiology assessment protocols for posttraumatic stress disorders, and parent–child behavior observation systems for clinic use. An *assessment instrument* refers to the particular method of acquiring data in psychological assessment (e.g., questionnaires, behavioral observation, and psychophysiological measurement). An assessment instrument includes all aspects of the measurement process that can affect the data obtained (e.g., instructions to participants, situational aspects of instrument stimuli, individual behavior codes, and questionnaire items). This article focuses on psychological assessment as applied in clinical decision-making (e.g., diagnosis, clinical case modeling, and treatment design).

[2] There are exceptions. Some have rejected content validity as a category of validity (Messick, 1993) or have suggested that it is more accurately considered the process of operationalizing a construct (e.g., Guion, 1977).

The phrase *the degree to which* refers to the fact that content validity is a quantitatively based judgment (e.g., quantitative estimates of relevance and representativeness). This phrase also suggests that content validity is a dimensional, rather than categorical, attribute of an assessment instrument (Lennon, 1956).

The term *construct* refers to the concept, attribute, or variable that is the target of measurement. Constructs can differ in their level of specificity from molar-level, latent variable constructs[3] such as conscientiousness to microlevel, less inferential variables such as hitting and alcohol ingestion. Most targets of measurement in psychological assessment, regardless of their level of specificity, are constructs in that they are theoretically defined attributes or dimensions of people.[4]

The phrase *for a particular purpose* refers to the fact that indices of relevance and representativeness of an assessment instrument can vary depending on the functions of the assessment. As noted by Messick (1993), content validity is a state, not a trait of an obtained assessment instrument score—content validity varies with the inferences that are to be drawn from the assessment data. For example, the content validity of a questionnaire measuring symptoms of depression may be satisfactory when the questionnaire is used as a brief screening instrument but not when used for treatment planning.

The *relevance* of an assessment instrument refers to the appropriateness of its elements for the targeted construct and function of assessment (Ebel & Frisbie, 1991; Guion, 1977; Messick, 1993; Suen, 1990). For example, the relevance of a self-report questionnaire measuring posttraumatic stress disorder (PTSD) symptom severity would covary with the degree that the measure contains items that reflect facets of PTSD, such as recurrent and distressing recollections and dreams, insomnia, and hypervigilance (*Diagnostic and Statistical Manual of Mental Disorders,* fourth edition [*DSM-IV*]; American Psychiatric Association [APA], 1994). Relevance would decrease to the degree that the questionnaire contained items outside the domain of PTSD (e.g., the degree that it contained items reflecting symptoms of substance dependence or conversion disorders).[5]

The *representativeness* of an assessment instrument refers to the degree to which its elements are proportional to the facets of the targeted construct (Lynn, 1986; Nunnally & Bernstein, 1994; Suen & Ary, 1989). In classical test theory, most often applied in educational and personnel evaluation, item content of an instrument is representative to the degree that the entire domain of the targeted construct can be reproduced (e.g., the entire spectrum of clerical or supervisory skills). The representativeness of a particular questionnaire purporting to assess PTSD depends on the degree to which its items are proportionally distributed or scored across the three major symptom clusters of PTSD (*DSM-IV;* APA, 1994).

Content Validity as Construct Validity

Content validation provides evidence about the construct validity of an assessment instrument (Anastasi, 1988). *Construct validity* is the degree to which an assessment instrument measures the targeted construct (i.e., the degree to which variance in obtained measures from an assessment instrument is consistent with predictions from the construct targeted by the instrument).

Construct validity subsumes all categories of validity (see Messick, 1993, and *Standards for Educational and Psychological Testing,* 1985). Measures of the predictive, concurrent, and postdictive validity, discriminant and convergent validity, criterion-related validity, and factor structure provide evidence about the construct validity of an assessment instrument. Content validity is an important component of construct validity because it provides evidence about the degree to which the elements of the assessment instrument are relevant to and representative of the targeted construct.[6]

Content validation of an assessment instrument unavoidably involves validation, and sometimes refinement, of the targeted construct (Smith & McCarthy, 1995). Constructs are synthesized variables, and evidence about the validity of instruments designed to measure them also provides evidence about the utility, domain, facets, boundaries, and predictive efficacy of the construct. The "personality" assessment literature is replete with examples of trait constructs that have been modified, partitioned, or discarded because of disconfirming validity evidence (Haynes & Uchigakiuchi, 1993).

In psychological assessment, the importance of content validation for the validation of the target construct varies depending on how precisely the construct is defined and the degree to which "experts" agree about the domain and facets of the construct. Content validation is particularly challenging for constructs with fuzzy definitional boundaries or inconsistent definitions (Murphy & Davidshofer, 1994). For example, in 1988 there were at least 23 self-report questionnaires that measured social support (Heitzmann & Kaplan, 1988). These questionnaires were developed on the basis of divergent ideas about the domain and facets of social support.

The multiple categories of construct validity can yield discor-

---

[3] *Latent variables* are unobserved variables hypothesized to explain the covariance between observed variables. In latent variable modeling, observed variables such as a Beck Depression Inventory score (Beck, 1972), self-monitored mood ratings, and the Hamilton Rating Scale for Depression (Hamilton, 1960), are presumed to be observable but imperfect indices of the latent variable, depression (see Loehlin, 1992, for an in-depth discussion).

[4] Some authors (e.g., Suen & Ary, 1989) have argued that molecular variables such as hitting, interruptions, or heart rate are not constructs in the usual sense of being indirectly measured latent variables: They are more appropriately considered as "samples" or "categories" of events. However, highly specific variables can be synthesized and measured in different ways and are, consequently, amenable to content validation.

[5] A number of behavior problems and cognitive disorders have been found to significantly covary with PTSD severity (e.g., Figley, 1979; Foa, Steketee, & Rothbaum, 1989) but are not specific to the domain of PTSD (i.e., are correlates but not "prototypic"; Smith & McCarthy, 1995). If the function of the instrument is to aid in differential diagnosis, inclusion of correlates may be appropriate.

[6] Several authors (Groth-Marnat, 1990; Guion, 1978; Messick, 1993; Mitchell, 1986; Tallent, 1992) have questioned the relevance of traditional concepts of validity, including content validity, for psychological assessment. They have suggested that validity concepts are less applicable to the higher-level, inconsistently defined constructs often targeted in psychological assessment. There is also significant disagreement among psychometricians about the language and procedures of validation. We have adopted traditional definitions of validity in our discussion of the importance of content validity for psychological assessment.

dant results. An assessment instrument with inadequate content validity (e.g., an observational system for marital communication that omits important paralinguistic behaviors) may be valid in other ways. The instrument might still accurately measure the observed behaviors, predict subsequent marital status, discriminate between couples seeking and not seeking marital therapy, exhibit excellent interobserver agreement, provide temporally and situationally stable data, and yield high coefficients of internal consistency (Haynes & Waialae, 1994). In addition, strong indices of criterion-related validity could be obtained for a content-invalid instrument if the indices of shared variance between the instrument and criterion are the result of shared variance in elements outside the construct domain.

The Importance of Content Validity

As noted earlier, content validity affects the clinical inferences that can be drawn from the obtained data. For sake of illustration, presume we are attempting to measure the efficacy of a psychosocial treatment for panic attack[7] (as defined in *DSM-IV;* APA, 1994) with a self-report questionnaire. Scores from the questionnaire on panic attacks would reflect the panic attack construct (i.e., would evidence content validity) to the extent that the items measured all facets of the construct, namely, (a) tapped the 13 criteria for panic attacks (*DSM-IV;* APA, 1994, pp. 395), (b) targeted the appropriate time frame estimate for peak response (<10 min), (c) solicited reports of panic attack frequency, (d) measured respondents' degree of concern and worry about the panic attacks, and (e) tapped the sequelae to the effects of panic attack.

The content validity of the panic attack questionnaire would be compromised to the degree that (a) items reflecting any of the five facets just listed were omitted, (b) items measuring constructs outside the domain of panic attacks were included (e.g., items on the magnitude of depression or obsessive–compulsive behaviors), or (c) the aggregate score on the questionnaire was disproportionately influenced by any facet of the panic attack construct (e.g., if the questionnaire contained three items on the "fear of dying" and only one on "cardiovascular" symptoms of a panic attack, or if items measuring different facets of the construct were equal in number but disproportionately weighted when deriving an aggregate score).

The use of a content-invalid assessment instrument degrades the clinical inferences derived from the obtained data because variance in obtained scores cannot be explained to a satisfactory degree by the construct. Data from an invalid instrument can overrepresent, omit, or underrepresent some facets of the construct and reflect variables outside the construct domain. A content-invalid assessment instrument could erroneously indicate the occurrence or nonoccurrence of clinically significant treatment effects. Similarly, erroneous inferences could be drawn about causes for panic attacks (e.g., the immediate triggers for attacks or the factors that affect the severity or duration of attacks) because estimates of shared variance would be based on erroneous measures of the construct. For example, shared variance with cardiovascular symptoms of panic attacks could not be identified if the symptoms were not proportionately measured by the assessment instrument. Changes in the questionnaire scores could also reflect changes in constructs outside the domain of panic attacks, thus leading to erroneous inferences about treatment effects and causal relationships.

Content validity also affects the latent factor structure of an assessment instrument. Instrument items (e.g., questions and behavior codes) are often selected to represent the facets, or latent factor structure, of an instrument. It is presumed that items measuring the same facet will demonstrate significant covariance. An instrument with inadequate content validity will fail to confirm the hypothesized latent structure of the assessment instrument because the items will not demonstrate significant magnitudes of covariance and because the instrument will not sufficiently tap the facets of the construct or will tap variables outside the construct domain.

Content validity is important for any aggregated measure derived from an assessment instrument (e.g., factor or scale score, summary score, or composite score). An aggregated variable is a combination of multiple measures. Components of an aggregate should be relevant to and representative of the aggregate construct and should evidence significant covariance. Aggregation can occur across time samples (e.g., averaging the observed rates of peer interactions of an elementary school child across several observation periods), across responses (e.g., generating an index of cardiovascular reactivity by combining heart rate, blood pressure, and peripheral blood flow responses to a laboratory stressor), across situations, across persons (e.g., generating an index of aggression in a classroom by summing aggressive behaviors across a sample of children), and across component items (e.g., generating an index of depression by summing responses to multiple questionnaire items).

Aggregation has been presumed to increase predictive efficacy because the measurement errors associated with individual elements of an aggregate often cancel each other out (Rushton, Philippe, Charles, & Pressley, 1983). However, the representativeness and relevance of the aggregated elements significantly affect the clinical judgments that can be drawn from the obtained data (e.g., presume that the sample of aggressive children omitted, or contained only, the most aggressive children in the classroom).

In summary the content validity of assessment instruments affects estimates of the parameters of behavior disorders (e.g., magnitude and duration), estimates of causal and functional relationships, diagnosis, the prediction of behavior, participant selection in clinical research, and estimates of treatment effects. Clinical inferences from assessment instruments with unsatisfactory content validity will be suspect, even when other indices of validity are satisfactory.

Validation of Assessment Inferences and Assessment Instruments

Functional Context

Content validation provides information about the data obtained from an assessment instrument and the inferences that can be drawn from those data (Guion, 1978; Hambleton & Rogers, 1991; Messick, 1993; Suen, 1990). Sometimes, validation procedures also provide information about the assessment

[7] Panic attacks are one component of the diagnostic construct panic disorder, as defined in *DSM-IV* (APA, 1994).

instrument. Examples of clinical inferences derived from assessment instrument data include (a) assigning a person's relative position on a trait construct (e.g., characterizing a person as high trait anxiety derived from a self-report questionnaire), (b) estimating a client's mean daily resting blood pressure from measurements on an ambulatory electrosphygmomanometer, and (c) estimating the proportion of a child's prosocial behavior that receives parental reinforcement based on measurements taken in a structured clinic observation setting.

The data, and judgments based on the data, are the primary object of validation studies. However, in the preceding examples we would want to know the degree to which reading difficulties, instrument malfunction, or observer drift, respectively, affected the obtained data. The validity of the data is a limiting factor for the validity of the clinical inferences. A "true score" may have been obtained from each instrument (e.g., no observer drift and high interobserver agreement for item c); however, inferences from the data would be compromised to the degree that the instrument elements were inappropriate for the targeted construct and assessment purpose (e.g., if some important parent behaviors were omitted from the observation coding system) or to the extent that sampling errors occurred (e.g., if blood pressure was sampled during exercise and stressful periods).

Several points regarding the conditional nature of assessment inferences and the role of content validity are particularly important: (a) the superordinate function of psychological assessment is to assist clinical judgment, (b) an assessment instrument has content validity to the degree that it taps the targeted construct and facilitates valid clinical judgments, and (c) inferences about the content validity of an assessment instrument are not necessarily generalizable across specific functions.

Assessment instruments can have different functions, and indices of validity for one function of an instrument are not necessarily generalizable to other functions of the instrument (Ebel, 1983; Guion, 1978; Hartmann, 1982; Mitchell, 1986). Consequently, validity indices are conditional—they pertain to an assessment instrument, when used for a particular purpose.[8]

Inferences about the *unconditional validity* of an assessment instrument (its validity, regardless of function) vary directly with the homogeneity of separate validity indices from studies across different assessment instrument functions. Because of the conditional nature of validation, it should rarely be assumed that an assessment instrument has unconditional validity. Statements such as ". . . has been shown to be a reliable and valid assessment instrument" do not reflect the conditional nature of validity and are usually unwarranted. In rare instances, supportive evidence for the content validity of an assessment instrument, accumulated across assessment functions, can support its generalized content validity.

Because content validity indices are specific to its function, an assessment instrument's construction should be guided by its intended function (DeVellis, 1991; Guion, 1978): The elements of an instrument that are most relevant and representative will vary with its intended use and the inferences that will be drawn from the obtained data.[9] For example, the most content-valid elements of a self-report questionnaire to measure depression are likely to differ, depending on whether the instrument is designed for brief screening, for multidimensional and multimodal assessment of causal relationships, or for the global evaluation of treatment outcome. The same could be said of a behavioral observation system for measuring social isolation or a psychophysiological assessment system for measuring cardiovascular reactivity.

Similarly, the most relevant and representative elements of an assessment instrument that measures social skills, parenting skills, or problem solving will vary depending on whether the function of the assessment is to measure abilities or current behavior (Murphy & Davidshofer, 1994). Also, the most relevant elements of an assessment instrument will vary depending on whether its purpose is to measure (a) situation-specific or situation-nonspecific behaviors, (b) maximum or average behaviors, and (c) typical or atypical behaviors. Elements would also differ depending on the parameter of interest, that is, the frequency, magnitude, or duration of a behavior problem (Franzen, 1989; Haynes, 1992).

Content validity can be conditional also for the targeted population (Nunnally & Bernstein, 1994; Suen, 1990). Content validity can vary across populations, and validity should be established for the population that will be sampled for the intended function. For example, a brief screening instrument for depression may demonstrate adequate content validity for use in the workplace but not in outpatient or inpatient psychological service centers, or the instrument may be content valid for White Americans and not for Asian Americans (Marsella & Kameoka, 1989).[10]

Finally, content validity is conditional for a particular construct domain. Many constructs have similar labels but dissimilar domains and facets. For example, Kubany et al. (1995) noted various conceptualizations of guilt; Franzen (1989) noted many different models of memory; and Somerfield and Curbow (1992) noted multiple, multifaceted definitions for coping. An assessment instrument may have satisfactory content validity for one definition of a construct but not for others.

The Dynamic Nature of Content Validity

Assessment instrument development is conducted in the context of contemporaneous theories about the targeted construct. Because the definition, domain and facets of many constructs evolve over time, the relevance and representativeness of the elements of an assessment instrument for the targeted construct are unstable. That is, content validity often degrades over time as new data are acquired and theories about the targeted construct evolve (Cronbach, 1971; Haynes & Waialae, 1994). For example, behavior observation systems for marital communication developed in the 1960s have less content validity in the 1990s to the degree that they omit the range of para-

---

[8] Although the conditional nature of content validity is frequently acknowledged, we located no studies that examined the differential content validity of an assessment instrument across different functions.

[9] Other dimensions of an assessment instrument, such as length, format, and cost, are also affected by its function.

[10] Content validity of an assessment instrument is also conditional on other dimensions, such as the situation in which measurement occurs, the state of the respondents (e.g., medication state or hospitalization state), instructions to assessment participants, and contingencies on obtained data (e.g., admittance into a treatment program).

linguistic and nonverbal elements of dyadic communication that have more recently been shown to be correlated with communication efficacy and satisfaction (Gottman, Markman, & Notarius, 1977; see reviews of marital observation systems by Weiss & Heyman, 1990). The evolution of constructs over time is exemplified by the refinements in constructs such as learned helplessness, Type-A behavior patterns, trauma-related guilt, aggression, and social support (Haynes & Uchigakiuchi, 1993).

The dynamic nature of construct definitions has four implications for content validity: (a) indices of content validity cannot be presumed to remain stable across time, (b) the content validity of psychological assessment instruments should be periodically examined, (c) psychological assessment instruments should be revised periodically to reflect revisions in the targeted construct, and (d) erroneous inferences regarding revised constructs may be drawn from unrevised assessment instruments.

Elements of Content Validity

Content validity is relevant to all elements of an assessment instrument that affect the obtained data, including item content, presentation of stimuli, instructions, behavior codes, time-sampling parameters, and scoring. All instrument elements affect the data obtained from the instrument, the degree to which the data obtained can be assumed to tap the targeted construct, and the clinical judgments that can be based on the data.

Content validity is relevant for all assessment methods, but the specific elements of content validity can differ in relevance across assessment methods. Table 1 outlines the relative importance of various content validity elements for four methods of psychological assessment.[11]

Most published articles on content validity have focused primarily on the content validity of self-report questionnaires, and almost exclusively from the perspective of educational and personnel assessment (see Hartmann, 1982, and Suen & Ary, 1989, as notable exceptions). However, content validity is also important for other assessment methods such as physiological or behavioral observation assessment because their resultant data affect clinical judgments. For example, in psychophysiological assessment, cardiovascular reactivity and poststress recovery are latent variables that can be defined and measured using different physiological systems, measurement procedures, time-sampling parameters, and data aggregation and reduction techniques (Cacioppo & Tassinary, 1990)—all of which will affect our inferences. Similarly, in behavioral observation, aggression, prosocial behavior, and self-injury are latent variables that can be defined and measured using different behavior codes, operational definitions, time-sampling parameters, observation situations, and data aggregation and reduction procedures (Hartmann & Wood, 1982).

The relevance of content validity for an assessment method is related to the level of specificity of the target construct and the degree to which the primary focus is on the obtained measure, independent of its relationship to a higher order latent-variable construct.[12] An emphasis on assessment data, independent of its implications for a higher order construct, is rare. For example, assessors are rarely interested in heart rate apart from of its *implications for higher order* physiological mechanisms, such as sympathetically mediated arousal. In contrast, blood pressure is sometimes the variable of primary interest to the clinician or researcher, independent of its function as a marker of some higher order construct.

Similarly, behavioral assessors are often not interested in the rate of interruptions during dyadic communication, in isolation from the construct of which interruptions are a marker. Interruptions are often measured because they are presumed to be one sign of negative communication behaviors that can covary with relationship satisfaction. However, the interruptions variable can be the primary target of assessment when it has been identified as an important causal variable for communication and problem-solving difficulties or marital distress (see discussions of behavioral marital assessment in Margolin, Michelli, & Jacobson, 1988, and Weiss & Heyman, 1990).

Content validity can still be relevant when measuring "samples" rather than "signs": Many elements of the measurement process can affect clinical inferences. For example, the definition of interruptions used by the observers, the situations in which this class of behaviors is observed, how the data are aggregated across codes and time, instructions to participants, and the time-sampling parameters of the observations will affect the obtained data and the inferences that can be derived from them.

Differences among assessment methods in the applicability of the various content validity elements are also influenced by the underlying assumptions of the assessment paradigm. For example, situational factors are frequently of interest in behavioral assessment. Therefore, the representativeness and relevance of situational factors are particularly important considerations in behavioral assessment.[13] Situation sampling would be less important for an assessment instrument designed to provide an aggregated "trait" score (Haynes & Uchigakiuchi, 1993).

Many behavior observation coding systems are designed to measure a construct, or response class. A *response class* is a group of dissimilar behaviors that have the same function—they operate on the environment in a similar manner or are maintained by the same contingencies. For example, both a hand gesture and "speaking over" can function as an interruption in dyadic communication (see the discussion of response classes in Donahoe & Palmer, 1994; Suen & Ary, 1989). The degree to which the behavior codes selected represent the targeted response class is an element of content validity because

---

[11] With broad definitions of the elements in Table 1, it could be argued that all elements are relevant for all assessment methods; Table 1 is meant to portray the relative importance of the various elements.

[12] This is sometimes referred to as the "sign" versus "sample" dimension of measurement. This issue is discussed by Hartmann (1982) and Suen and Ary (1989) and is also related to a latent-variable modeling discussed by Loehlin (1992).

[13] Fagot (1992) described a content validation procedure for situations depicted in a video-based parental discipline assessment instrument. Representative videotaped scenes of "risky behavior" by young children were developed from statements from 20 parents of toddlers. These situations were then rated on their degree of risk (a measure of relevance) and annoyance by 30 additional mothers and fathers. Fourteen of the most risky, annoying scenes (e.g., riding a tricycle into the street) were then filmed and used as stimuli to obtain self-reports of parents as to their probable responses.

Table 1
Relevance of Content Validity Elements to Four Methods of Psychological Assessment

Element	Assessment method			
	Questionnaire	Behavioral observation	Psychophysiology	Self-monitoring
The array of items selected (e.g., questions, codes, measures)	R	R	R	R
Precision of wording or definition of individual items	R	R	N	R
Item response form (e.g., scale)	R	N	N	R
Sequence of items or stimuli	R	N	R	N
Instructions to participants	R	R	R	R
Temporal parameters of responses (interval of interest; timed vs. untimed)	R[a]	N	N	R
Situations sampled	R	R	R	R
Behavior or event samples	R	R	R	R
Components of an aggregate, factor, response class	R	R	R	R
Method and standardization of administration	R	R[b]	R	R
Scoring, data reduction, item weighting	R	R	R	R
Time sampling parameters[c]	N	R	R	R
Stimulus presentation[d]	N	R	R	N
Definition and domain of the construct	R	R	R	R
Method–mode match	R	R	R	R
Function–instrument match	R	R	R	R

Note. R = relevant; N = not relevant.
[a] In most self-report instruments, it is important to note whether responses refer to "at the moment," "during the past day," etc. [b] Standardization is important across participants, or across assessment occasions, depending on whether a nomothetic or idiographic approach was taken. [c] The rate and temporal pattern of measurement. [d] For example, audio versus video presentations of stressors in a psychophysiological laboratory assessment; scenarios of lab observations.

it indicates the relevance and representativeness of the obtained data for that class. However, behavioral observation systems rarely undergo systematic content validation. Developers most often rely on the face validity of the selected codes.[14]

Two other important elements of content validity are the method–mode match (Suen, 1990) and the method–function match. The *method–mode match* is the degree to which a particular assessment method is appropriate for the targeted construct. The method–mode match issue has been frequently raised in discussions about the appropriateness of self-report versus other-person report measures of internal versus external events in child assessment (Kazdin, 1990). The *method–function match* is the degree to which a particular assessment method is appropriate for the purposes of the assessment. For example, an interview may be appropriate for narrowing the range of possible diagnoses for a client reporting anxiety symptoms but may not be the most appropriate assessment method for measuring treatment effects.

Content validity is also relevant to the array of instruments used in clinical assessment—the degrees to which the instruments selected are relevant to the characteristics of the client and purposes of the assessment. As noted earlier, assessment instruments vary in the constructs that they tap, in the degree to which they tap their targeted constructs, and in their relevance for specific assessment functions. For example, for treatment design for adolescent antisocial behaviors, an assessment strategy that relies on a limited number of sources (Patterson, 1993) would evidence a low level of content validity because self-report measures do not adequately sample from the domain of adolescent antisocial behaviors. Similarly, an assessment strategy for developing a causal model of PTSD that omitted measures of trauma-related guilt (Kubany et al., 1995) or for developing an intervention program that did not assess the client's goals (Evans, 1993) would not include variables that were important for the functions of the assessment.

Methods of Content Validation

Content validation is a multimethod, quantitative and qualitative process that is applicable to all elements of an assessment instrument. During initial instrument development, the purpose of content validation is to minimize potential error variance associated with an assessment instrument and to increase

[14] Face validity is a component of content validity. It refers to the degree that respondents or users judge that the items of an assessment instrument are appropriate to the targeted construct and assessment objectives (Allen & Yen, 1979; Anastasi, 1988; Nevo, 1985). It is commonly thought to measure the acceptability of the assessment instrument to users and administrators.

the probability of obtaining supportive construct validity indices in later studies. Because sources of error vary with the targeted construct, the method of assessment, and the function of assessment, the methods of content validation will also vary across these dimensions (Hartmann, 1982).

Many authors have outlined recommended methods of content validation but have focused primarily on the content validation of questionnaire items. The Appendix integrates these recommendations with other recommendations inferred from the expanded array of content validity elements outlined in previous sections of this article. DeVellis (1991) illustrated a general sequence of content validation. Fagot (1992) described the content validation of a videotape-aided assessment instrument for parenting skills. Frank-Stromborg (1989) and Kubany et al. (1995) described content validation procedures for cancer reaction and trauma-related guilt questionnaires, respectively.

A detailed examination of the 35 recommended steps and judgments outlined in the Appendix is beyond the domain of this article. Instead, we will focus on a few general principles and provide a list of recommendations to help guide the complex process of content validation.

Content Validation Guidelines

Carefully define the domain and facets of the construct and subject them to content validation before developing other elements of the assessment instrument (Nunnally & Bernstein, 1994; Suen, 1990; Walsh, 1995). This first step is essential to the development of a content-valid assessment instrument, and is the most difficult phase of content validation (Murphy & Davidshofer, 1994). A construct that is poorly defined, undifferentiated, and imprecisely partitioned will limit the content validity of the assessment instrument. For example, in developing a questionnaire on trauma-related guilt (Kubany et al., 1995), the proposed definition, domain, and facets of trauma-related guilt should be subjected to expert review before generating items to tap the construct. The proposed modes and dimensions of trauma-related guilt to be tapped (e.g., beliefs of personal responsibility, feelings of distress, and guilt frequency and severity) should also be carefully articulated and evaluated. A precise differentiation among theoretically related constructs (e.g., trauma-related guilt versus depression) is particularly important (Ebel & Frisbie, 1991). A grid of the facets of the construct can facilitate the representativeness of the item content (Messick, 1993).[15]

Subject all elements of an assessment instrument to content validation (Murphy & Davidshofer, 1994). Elements such as instructions to participants during role-play assessment, questionnaire response formats and response scales, the audiotaped and videotaped scenes presented during psychophysiological assessments, the situations depicted in questionnaires and presented in observation sessions, and the behaviors observed in social interaction studies can all affect the obtained data, the relevance and the representativeness of the elements for the targeted construct, and the clinical inferences that can be drawn from the data. All such elements, regardless of their level of specificity and face validity, are amenable to content validation. For example, in developing a psychophysiological PTSD assessment instrument for use with veterans, the battle scenes can be reviewed by combat veterans for their relevance; the selected psychophysiological measures can be reviewed by PTSD experts and psychophysiologists.

Use population and expert sampling for the initial generation of items and other elements. Although population and expert sampling is frequently recommended by psychometricians, these procedures are infrequently used by the developers of psychological assessment instruments. Carefully structured, open-ended interviews with persons from the targeted population and experts can increase the chance that the items and other elements are representative of and relevant to the facets of the construct. This process can also suggest additional facets and the need for construct refinement.

Use multiple judges of content validity and quantify judgments using formalized scaling procedures (Guion, 1978; Hambleton & Rogers, 1991; Lawshe, 1975; Lynn, 1986; Tittle, 1982). Every element of an assessment instrument (see Table 1) should be judged by multiple experts, using 5- or 7-point evaluation scales, on applicable dimensions such as relevance, representativeness, specificity, and clarity. The resulting descriptive statistics (even without formalized criteria for interpretation) can guide judgments about the content validity of the elements (Nunnally & Bernstein, 1994). The data from this evaluative pilot testing can help identify elements of the assessment instrument that require refinement and items that should be omitted.[16]

Instruments that are refined following initial content validation should undergo further evaluation. Hambleton and Rogers (1991) suggested that new assessment instruments also be reviewed for technical quality (e.g., for grammar, wording, randomization of items, and scaling) by measurement specialists.

The optimal number of judges will vary with the element under consideration, the internal consistency of the ratings, and practical considerations (e.g., instrument length and availability of experts; see discussion by Crocker, Llabre, & Miller, 1988; Lynn, 1986). However, confidence in the robustness of the ratings (the standard error of measurement) will increase as the number of judges increases. In addition, increasing the number of raters (e.g., more than five) facilitates the detection and exclusion of rater outliers (Carmines & Zeller, 1979; Lynn, 1986). Similar procedures can be used with target population samples (e.g., mothers and fathers, when developing a parental discipline assessment instrument, and combat veterans and rape and incest survivors, when developing a PTSD questionnaire). Quantitative indices of content validity can be supplemented with qualitative feedback from evaluators (e.g., suggested additions and rewordings).

Examine the proportional representation of items. The items in an assessment instrument should be distributed, or weighted, in a way that reflects the relative importance of the

[15] To formally establish the "representativeness" of the elements of an assessment instrument, the proportions of variance in the overall construct associated with various facets of the construct would have to be independently established. The partitioned variance in the assessment instrument should match that independently established for the instrument (e.g., the relative contribution of somatic vs. cognitive facets in a questionnaire measure of depression).

[16] Self-administered computerized assessment can be particularly helpful with this task because the computer can identify the items about which participants frequently request clarification.

various facets of the targeted construct (Anastasi, 1988). If items overrepresent or underrepresent facets of a construct, the obtained scores and inferences from these scores will be biased. For example, a questionnaire that disproportionately targets somatic elements of depression relative to cognitive or behavioral elements illustrates the inferential difficulties associated with disproportionate item representation.

Report the results of content validation when publishing a new assessment instrument. Indices of content validity can help potential users evaluate the targeted construct and the relevance and representativeness of the instrument elements for a particular assessment function. Content validation procedures and content validity indices, as well as the assessment functions for which the validity indices are applicable, should be treated as important categories of construct validation and should be reported systematically in the same detail as other components of construct validation.

Use subsequent psychometric analyses for assessment instrument refinement. All indices of validity have implications for content validity. Low indices of other categories of construct validity suggest that the instrument items may be insufficiently representative or relevant, or that the construct may not be precisely or appropriately defined. However, high indices of construct validity are necessary, but insufficient, to infer a satisfactory degree of construct validity. As noted earlier, high magnitudes of shared variance between scores from the newly developed instrument and criterion instruments can result from variance in items outside the domain of the targeted construct. Low indices of criterion-related validity can erroneously suggest content validity difficulties when the criterion instrument (a) is based on a different definition of the construct, (b) contains items outside the construct domain, or (c) disproportionately taps some facets of the construct. Item analysis, internal consistency indices, and the obtained factor structure also provide essential information about the degree to which an item taps the intended constructs and facets (Smith & McCarthy, 1995). Facets are constructs, and the degree to which assigned items covary and tap that construct can be examined empirically.

Content Validity of Existing Instruments and Recommendations for Reevaluation

To examine current practices in content validation, we examined all ($N = 19$) articles published in 1992–1994 in *Psychological Assessment* and *Behavior Research and Therapy* that reported on the development of a new assessment instrument (all were self-report questionnaires or rating scales). Each article was reviewed to determine if the assessment instrument elements were derived from (a) items from previously published instruments (5), (b) clinical experience or deductive reasoning by the developers (5), (c) theories and literature about the target behavior problems (12), (d) expert sampling (4), (e) population sampling (14), and (f) the results of empirical research (e.g., item discrimination indices; 4).

In addition, we examined all articles published in 1993–1994 in the *Journal of Applied Behavior Analysis* that reported on the clinical application of a new behavior observation coding system. Of the 18 behavioral observation studies rated, 7 did not provide information about how the behavior codes or observation system were developed. Only three studies reported systematic approaches to assessment instrument development. The methods included interviews with "experts" (parents and teachers of target children), informal classroom observation of target children before developing a coding system, and a review of a target child's school and medical records. In most cases, idiosyncratic behavior codes were constructed rationally by the investigators, apparently without reference to existing codes and without evidence that the codes selected were the most relevant and representative for a particular target or for a particular assessment function.

Although many previously published assessment instruments have been subjected to extensive psychometric evaluation, most of the thousands of available psychological assessment instruments were rationally derived and not subjected to systematic, quantitative content validation as outlined in the Appendix. We suggest that the most frequently used assessment instruments for a given construct and function be subjected to expert review of their comparative content validity according to the dimensions outlined in Method 10 of the Appendix. Content validation of these instruments would help establish (a) the relative degree to which they tap the targeted construct, (b) their most appropriate functions, (c) the inferences that can be drawn from the resultant data, and (d) elements that may benefit from refinement. It would be particularly helpful to users in cases where there are multiple, frequently used instruments for the assessment of a construct (e.g., the multiple questionnaires on depression, anxiety, and quality of life). A "grid" format in which many experts evaluate the content validity of multiple measures of a construct on multiple dimensions would be helpful to users and for instrument refinement.

Summary

Content validity is a category of construct validity: It is the degree to which the elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose. Content validation is applicable across assessment methods because it addresses the inferences that are based on the obtained data. Content validity has implications for the prediction of behavior and for causal models of behavior disorders, diagnosis, and estimates of treatment effects.

There are multiple elements of content validity. All aspects of an assessment instrument that can affect the obtained scores, and the interpretation of these scores, are appropriate targets for content validation. The importance of various elements varies across methods and instruments, and most can be evaluated quantitatively.

Content validity indices are specific to a particular function of the assessment instrument and to other factors such as the population to which the instrument is applied and the assessment situation in which the instrument is used. Because the definition, domain, and facets of many constructs evolve over time, the relevance and representativeness of an assessment instrument are likely to degrade.

We have outlined many methods of content validation in this article. We stressed the desirability of (a) a careful definition and quantitative evaluation of the targeted construct, (b) a multielement approach to content validation, (c) the use of population and expert sampling in initial item development, (d) quantitative

evaluations from experts and potential respondents, (e) an evaluation of the proportionate representativeness of items, (f) a detailed reporting of the results of content validation, and (g) the relevance for content validity of subsequent psychometric analyses.

Finally, we noted that many psychological assessment instruments were developed without following the content validation methods outlined in this article. We recommended that comparative studies be conducted on the content validity of multiple instruments with a similar construct focus.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.) Washington, DC: Author.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Beck, A. T. (1972). *Depression: Causes and treatment.* Philadelphia: University of Pennsylvania Press.

Cacioppo, J. T., & Tassinary, L. G. (1990). *Principles and psychophysiology: Physical, social, and inferential elements.* New York: Cambridge University Press.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity.* Beverly Hills: Sage.

Crocker, L., Llabre, M., & Miller, M. D. (1988). The generalizability of content validity ratings. *Journal of Educational Measurement 25,* 287–299.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement.* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

DeVellis, R. F. (1991). *Scale development: Theory and applications.* Newbury Park: Sage.

Donahoe, J. W., & Palmer, D. C. (1994). *Learning and complex behavior.* Boston: Allyn & Bacon.

Ebel, R. L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice, 2,* 7–10.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement.* (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Evans, I. (1993). Constructional perspectives in clinical assessment. *Psychological Assessment, 5,* 264–272.

Fagot, B. I. (1992). Assessment of coercive parent discipline. *Behavioral Assessment, 14,* 387–406.

Figley, C. R. (Ed.). (1979). *Trauma and its wake: Vol. 1. The study of post-traumatic stress disorder.* New York: Brunner/Mazel.

Foa, E. B., Steketee, G., & Rothbaum, B. O. (1989). Behavioral/cognitive conceptualizations of post-traumatic stress disorder. *Behavior Therapy, 20,* 155–176.

Frank-Stromborg, M. (1989). Reaction to the diagnosis of cancer questionnaire: Development and psychometric evaluation. *Nursing Research, 38,* 364–369.

Franzen, M. D. (1989). *Reliability and validity in neuropsychological assessment.* New York: Plenum.

Gottman, J., Markman, H., & Notarius, C. (1977). A sequential analysis of verbal and nonverbal behavior. *Journal of Marriage and the Family, 39,*461–477.

Groth-Marnat, G. (1990). *Handbook of psychological assessment* (2nd ed.). New York: Wiley.

Guion, R. M. (1977). Content validity—The source of my discontent. *Applied Psychological Measurement, 1,* 1–10.

Guion, R. M. (1978). "Content validity" in moderation. *Personal Psychology, 31,* 205–213.

Hambleton, R. K., & Rogers, H. J. (1991). Advances in criterion-references measurement. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 3–43). Boston: Kluwer Academic.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23,* 56–62.

Hartmann, D. P. (Ed.). (1982). *Using observers to study behavior.* San Francisco: Jossey-Bass.

Hartmann, D. P., & Wood, D. D. (1982). Observational methods. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (pp. 109–138). New York: Plenum.

Haynes, S. N. (1992). *Models of causality in psychopathology: Toward synthetic, dynamic and nonlinear models of causality in psychopathology.* Des Moines, IA: Ayllon & Bacon.

Haynes, S. N. (1994). Clinical judgment and the design of behavioral intervention programs: Estimating the magnitudes of intervention effects. *Psichologia Conductual, 2,* 165–184.

Haynes, S. N., & Uchigakiuchi, P. (1993). Incorporating personality trait measures in behavioral assessment: Nuts in a fruitcake or raisins in a mai tai? *Behavior Modification, 17,* 72–92.

Haynes, S. N., & Waialae, K. (1994). Psychometric foundations of behavioral assessment. In R. Fernández-Ballestros (Ed.), *Evaluacion Conductual Hoy.* Madrid: Ediciones Piramide.

Heitzmann, C. A., & Kaplan, R. M. (1988). Assessment of methods for measuring social support. *Health Psychology, 7,* 75–109.

Kazdin, A. E. (1990). Assessment of childhood depression. In A. M. La Greca (Ed.), *Through the eyes of the child: Obtaining self-reports from children and adolescents* (pp. 189–233). Boston: Ayllon and Bacon.

Korchin, S. J. (1976). *Modern clinical psychology.* New York: Basic Books.

Kubany, E., Haynes, S. N., Abueg, F. R., Marke, F. P., Brennan, J. & Stahura, C. (1995). *Development and validation of the Trauma-Related Guilt Inventory.* Manuscript submitted for publication.

Lawshe, C. H. (1975). The quantitative approach to content validity. *Personnel Psychology, 28,* 563–575.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16,* 294–304.

Loehlin, J. C. (1992). *Latent variable models: An introduction to factor, path, and structural analysis.* Hillsdale, NJ: Erlbaum.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35,* 382–385.

Margolin, G., Michelli, J., & Jacobson, N. (1988). Assessment of marital dysfunction. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (pp. 441–489). New York: Pergamon.

Marsella, A. J., & Kameoka, V. (1989). Ethnocultural issues in the assessment of psychopathology. In S. Wetzler (Ed.), *Measuring mental illness: Psychometric assessment for clinicians* (pp. 231–256). Washington, DC: American Psychiatric Association.

Messick, S. (1993). Validity. In R. L. Linn (Ed), *Educational measurement* (2nd ed., pp. 13–104). Phoenix: American Council on Education and Oryx Press.

Mitchell, J. V., Jr. (1986). Measurement in the larger context: Critical current issues. *Professional Psychology: Research and Practice, 17,* 544–550.

Murphy, K. R., & Davidshofer, C. O. (1994). *Psychological testing: Principles and applications* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22,* 287–293.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Patterson, G. R. (1993). Orderly change in a stable world: The antisocial trait as a chimera. *Journal of Consulting and Clinical Psychology, 61,* 911–919.

Rushton, J., Philippe, B., Charles, J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.

Smith G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7,* 300–308.

Somerfield, M., & Curbow, B. (1992). Methodological issues and research strategies in the study of coping with cancer. *Social Science Medicine, 34,* 1203–1216.

Standards for educational and psychological testing. (1985). Washington, DC: American Psychological Association.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale, NJ: Erlbaum.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative observation data.* Hillsdale, NJ: Erlbaum.

Tallent, N. (1992). *The practice of psychological assessment.* Englewood Cliffs, NJ: Prentice-Hall.

Tittle, C. K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Ed.), *Handbook of methods for detecting bias* (pp. 31–63), Baltimore: Johns Hopkins University Press.

Walsh, W. B. (1995). *Tests and assessment.* New York: Prentice-Hall.

Weiner, E. B. (1976). *Clinical methods in psychology.* New York: Wiley.

Weiss, R. L., & Heyman, R. E. (1990). Observation of marital interaction. In F. D. Fincham & T. N. Bradury (Eds.), *The psychology of marriage: Basic issues and applications* (pp. 87–117). New York: Guilford.

Appendix

Procedures and Sequence of Content Validation

Asterisked components are those most frequently overlooked. Not all methods of validation are relevant for all methods of assessment. Components listed here have been drawn from Anastasi (1988), DeVillis (1991), Ebel and Frisbie (1991), Franzen (1989), Hambleton and Rogers (1991), Hartmann (1982), Lynn (1986), Messick (1993), Murphy and Davidshofer (1994), Nunnally and Burnstein (1994), Suen (1990), and Walsh, (1995).

1. Specify the construct(s) targeted by the instrument
   a. Specify the domain of the construct
      i. what is to be included
      ii. what is to be excluded*
   b. Specify the facets and dimensions of the construct
      i. factors of construct to be covered
      ii. dimensions (e.g., rate, duration, and magnitude)*
      iii. mode (e.g., thoughts and behavior)*
      iv. temporal parameters (response interval and duration of time-sampling)*
      v. situations*
2. Specify the intended functions of the instrument (e.g., brief screening, functional analysis, and diagnosis)
3. Select assessment method to match targeted construct and function of assessment*
4. Initial selection and generation of items (e.g., questionnaire items, behavior codes, psychophysiological measures, and behaviors monitored)
   a. from rational deduction
   b. from clinical experience
   c. from theories relevant to the construct
   d. from empirical literature relevant to the construct (e.g., studies on construct validity of potential items)
   e. from other assessment instruments (i.e., borrowing items from other instruments that have demonstrated validity)
   f. from suggestions by experts*
   g. from suggestions by target population*
5. Match items to facets and dimensions
   a. use table of facets to insure coverage (include all relevant dimensions, modes, temporal parameters, and situations)
   b. generate multiple items for each facet
   c. insure proportional representation of items across facets (i.e., the relative number of items in each facet should match the importance of that facet in the targeted construct)
6. Examine structure, form, topography, and content of each item
   a. appropriateness of item for facet of construct
   b. consistency and accuracy, specificity and clarity of wording, and definitions
   c. remove redundant items
7. Establish quantitative parameters
   a. response formats and scales
   b. time-sampling parameters (sampling intervals and durations)
8. Construct instructions to participants
   a. match with domain and function of assessment instrument
   b. clarify; strive for specificity and appropriate grammatical structure
9. Establish stimuli used in assessment (e.g., social scenerios, and audio and video presentations) to match construct and function
10. Have experts review the results of methods 1–3 and 5–9
    a. quantitative evaluations of construct definition, domain, facets, mode, and dimensions*
    b. quantitative evaluation of relevance and representativeness of items and stimuli
    c. quantitative evaluation of response formats, scales, stimuli, situations, time-sampling parameters, data reduction, and aggregation
    d. match of an instrument attributes to its function*
    e. qualitative evaluation—suggested additions, deletions, and modifications
11. Have target population sample the results—review quantitative and qualitative evaluation of items, stimuli, and situations*
12. Have experts and target population sample rereview the modified assessment instrument*
13. Perform psychometric evaluation and contingent instrument refinement—criterion-related and construct validity, and factor analysis

Received April 10, 1995
Revision received April 12, 1995
Accepted April 14, 1995 ■