II-5

J Clin Epidemiol Vol. 46, No. 9, pp. 1019–1023, 1993
Printed in Great Britain.

0895-4356/93 $6.00 + 0.00

THE VALIDITY OF HEALTH ASSESSMENTS: RESOLVING SOME RECENT DIFFERENCES

MICHAEL E. HYLAND
Department of Psychology, University of Plymouth, Plymouth PL4 8AA, England

(Received in revised form 2 February 1993)

Abstract—The purpose of this paper is to examine what is meant by a *valid* measure of health. Guyatt, Kirshner and Jaeschke propose that health tests should be designed so as to have one of several kinds of validity: "longitudinal construct validity" for those which are used for longitudinal research designs, and "cross-sectional construct validity" for those which are used for cross-sectional designs. Williams and Naylor argue that this approach to test classification and validation confuses what a test purports to measure with the purpose for which it is used, and that some tests have multiple uses. A review of the meanings of validity in the psychological test literature shows that both sets of authors use the term *validity* in an idiosyncratic way. Although the use of a test (evaluated by content validity) should not be conflated with whether the test actually measures a specified construct (evaluated by construct validity); if health is actually made up of several constructs (as suggested in Hyland's interactional model) then there may be an association between types of construct and types of purpose. Evidence is reviewed that people make several, independent judgements about their health: cognitive perceptions of health problems are likely to be more sensitive to change in a longitudinal research design, whereas emotional evaluations of health provide less bias in cross-sectional designs. Thus, a classification of health measures in terms of the purpose of the test may parallel a classification in terms of what tests purport to measure.

INTRODUCTION

Kirshner and Jaeschke [1, 2] have suggested a methodologic framework for health assessment which is based on a classification and validation of instruments in terms of the purpose for which the instrument is used. In particular, they distinguish between (a) health assessments which are designed to detect change in health status over time and where correlations between change scores from repeated measures provide "longitudinal construct validity" and those which are designed to compare between different groups of patients and where correlations between measures at a particular time provide "cross-sectional validity". Williams and Naylor [3] have criticized this classification system. They argue that "this approach conflates what an instrument measures with one of its potential uses", and go on to conclude that "Guyatt, Kirshner, and Jaeschke have eschewed several concepts and principles of measurement found in the behavioral and social sciences, and developed their own clinically-oriented framework."

Guyatt *et al.* certainly use the concept of validity in an idiosyncratic fashion—standard psychological texts on assessment do not refer to longitudinal or cross-sectional validity. However, it could easily be demonstrated that idiosyncratic use of the concept of validity is extremely common in accounts of health assessment. Williams and Naylor's use is also idiosyncratic, for example, in their equating of face and content validity. The purpose of this paper is to (a) describe the standard account(s) of validity in measurement theory, (b) show that classification in terms of function is acceptable within that standard account in specific circumstances, and (c) show what is needed for health measures to have construct validity.

HISTORICAL CONTEXT OF THE VALIDITY CONCEPT

The much quoted paper by Cronbach and Meehl [4], entitled "Construct validity in psychological tests" was part of a more general discussion then current in psychology. Seven years earlier, MacCorquodale and Meehl [5] had published a paper on the distinction between two types of theoretical construct in psychology, intervening variables and hypothetical constructs. In brief, hypothetical constructs have ontological status in the sense that the construct exists independently of its measurement; intervening variables, on the other hand, are exhaustively defined by their measurement and do not exist independently of it. MacCorquodale and Meehl suggested that the "truth" of hypothetical constructs and intervening variables should be established in different ways. The article by Cronbach and Meehl [4] developed this idea by suggesting that there were different kinds of validity (i.e. ways of establishing truth) depending on what the test was supposed to measure. Cronbach and Meehl suggested three different kinds of validity which form the basis for the standard view [6, 7] in measurement theory. These three types of validity are criterion validity, content validity and construct validity.

CRITERION, CONTENT AND CONSTRUCT VALIDITY

Criterion validity (or predictive validity) applies when the purpose of a test is to match some other measurement that already exists. For example, suppose there is an existing but rather long scale for measuring health, and it is desirable to create a shorter scale. The shorter measure has criterion validity to the extent that it predicts the results of the longer measure (i.e. the criterion), and the correlation coefficient between the two measures would then be the validity coefficient. Of course, criterion validity does not tell us what the original test was measuring—merely that the new test measures the same thing as the old. Contrary to Williams and Naylor [3], criterion validity is not the best type of validity—it is just a type of validity which has a particular function. And although there is certainly no independent criterion of health, criterion validity may well form part of research into health instruments.

Content validity is used where the test is designed to provide information about a particular application established by the researchers. Cronbach and Meehl [4] write: "Content validity is established by showing that the test items are a sample of a universe in which the investigator is interested. Content validity is ordinarily to be established deductively, by defining a universe of items and sampling within this universe to establish the test." Content validity is not established by correlating two measures, but by a logical process of showing that the content of the test's items correspond to what the test constructors say that their test is measuring. In fact, most writers on validity would agree with Nunnally [7, p. 92] that "rather than test the validity of measures after they are constructed, one should ensure validity by the plan and procedures of construction". Elsewhere, I have suggested [8] that this is precisely what happens with quality of life questionnaires. That is, authors define what they mean by quality of life in terms of domains, and then plan a sensible strategy of questionnaire construction which ensures that their questionnaire samples information from those domains. If those domains are defined in terms of a particular type of application for the questionnaire and there is more than one type of application, then a functional classification system such as that proposed by Guyatt *et al.* [1] will be useful. Indeed, Guyatt *et al.*'s suggested methodology for validation (cross-sectional correlations for scales intended for cross-sectional research and change correlations for scales intended for longitudinal research) is entirely consistent with an approach where the questionnaire is validated in terms of the adequacy with which it fulfils a particular research purpose.

Content validity should not be confused with face validity which is not a technical form of validity but merely the informal evaluation of apparent validity by the test users. Well established tests tend to have high face validity—but may not be more valid in any technical sense.

Construct validity exists when the purpose of the test is to measure a construct—i.e. a theoretical construct (originally called a hypothetical construct [5] and later a person variable [9]) which exists independently of the test. Construct validity is analogous to theory testing, and as with theory testing there are a variety of scientific procedures which can be used to provide confirmatory evidence that a test measures a particular construct. Consequently, there are many types of construct validity—types listed by Anastasi [6] include developmental changes,

correlations with other tests, factor analysis, internal consistency, convergent and discriminant validity, and experimental interventions. In contrast to Guyatt's proposal that longitudinal and cross-sectional validity require different kinds of methodology, construct validity is best obtained from a multi-methodologic framework where a variety of different methodologies are used.

CAN HEALTH MEASURES BE VALID IN MORE THAN ONE WAY?

Williams and Naylor [3] argue that Guyatt *et al.* [1] conflate the function of a test with what it measures. Guyatt *et al.*'s classification system is based on research function and their proposed method of validation tests the adequacy of those functions. By "longitudinal construct validity" they actually mean "content validity in terms of the ability to detect longitudinal change" and by "cross-sectional construct validity" they mean "content validity in terms of the ability to make cross-sectional comparisons".

However, all standard accounts of measurement theory allow for the possibility that a test may be valid in more than one way. Moreover, an underlying theme of researchers working at McMaster, is that they assume their tests measure some real property of the patient, i.e. they measure a construct with ontological status. Hence, construct validity is an aim of these and probably many other researchers.

It is at this point that the standard view of measurement theory becomes less clear, because there are two kinds of approach to the relationship between different types of validity. Some authors [7] suggest that the different types of validity actually overlap and that a questionnaire is best seen as having a combination of different types of validity. However, this view is inconsistent with the original [4] formulation that different types of validity exist for different types of construct. A more common viewpoint [6], therefore, is that a test has different types of validity only if it has different types of purpose.

In introducing the concept of validity, Anastasi [6] writes "No test can be said to have 'high' or 'low' validity in the abstract. Its validity must be established with reference to the particular use for which the test is being considered [6, p. 131]." She goes on to say "The same test, when employed for different purposes, should be validated in different ways [6, p. 152]". Thus it is quite conceivable that a health measure has both content and construct validity—content validity when validation is being considered in terms of the context in which the test is used and construct validity when validation is being considered in terms of whether the test measures some specified construct. Health measures may have multiple uses, and those different uses may require different kinds of validation.

THE CONSTRUCT OF HEALTH

Although there are numerous definitions of health and quality of life, definitions are not the same as constructs. These definitions provide a useful basis for establishing content validity, but they provide scant information about the construct of health. Many definitions imply, however (and this coincides with the belief of the McMaster group), that health exists as some mental or perceptual property of patients. That is, health assessment is crucially dependent on the patient's perception of health: health is a percept.

If health is indeed a psychological construct, then it makes sense to draw on existing psychological theory to infer what kind of construct that might be. There is a considerable body of research showing that perceptions, including perceptions of the self [10] are not unitary. Two distinctions are particularly relevant to the patient's perceptions of health. The first is that knowledge of events and evaluative appraisals of events are independent [11] and involve different neurological processes [12]. Thus a patients' knowledge of their health problems are likely to be independent of evaluations of how much distress those problems cause. Second, affective evaluations of positivity tend to be independent of affective evaluations of negativity [13, 14, 15, 16, 17]. That is, a patient may be both happy and unhappy with life or happy about some aspects and unhappy about others. I have used these two distinctions to develop a framework [18] for health which is shown in Fig. 1. The patient's knowledge of problems (which is caused by an interaction between morbidity and psychological factors such as coping style and information processing bias) is causally antecedent to the patient's evaluations of those problems (which are caused by knowledge of problems and other psychological variables such as trait affect and emotional reactivity). Some people with immense problems are highly satisfied with at least some aspect of their lives [16, 17]—and by contrast some people

with no problems are highly dissatisfied. It is quite unwarranted to assume that by merely adding together the patient's experience of problems (with or without weightings) it is possible to obtain an estimate of the patient's actual experience of distress. Moreover, evaluations of life satisfaction are not unidimensional, and judgements about life richness (positive life quality) may be independent of health complaints (negative life quality).

From this psychological perspective, quality of life is not a single construct but rather several causally connected constructs, and any questionnaire may therefore reflect one or more of these different constructs. I have proposed [18] that measures of different constructs in that causal sequence should be used for answering different types of research question. Specifically, the construct *knowledge of problems* is likely to be more sensitive to longitudinal change caused by drug treatment compared with *evaluations of health status*. The reason is that psychological constructs that are highly correlated with morbidity level are more likely to demonstrate change in a clinical trial involving drug therapy compared with constructs that are highly correlated with psychological factors (assuming the drug therapy is not psychoactive), because the drug does not alter the patient's disposition. As problems are more closely related to morbidity in the causal sequence, they are more likely to be affected by changes in morbidity compared with evaluations; and because problems are likely to be less closely related to personality, the anchoring effect of personality disposition on change scores will be less on problems compared with evaluations. Hence the patient's knowledge of problems should be more sensitive to change in clinical trials compared with emotional evaluations of those problems.

By contrast, evaluations of health have a useful role to play in comparing between different diseases. One difficulty which arises when comparing between different diseases is that different diseases are associated with different kinds of problem, for example, asthmatics with sleep disturbance and arthritics with pain. Therefore, if the items in a health questionnaire tend to reflect the kinds of problem experienced by one disease but not the other, it will appear that patients with the former disease have a worse quality of life than the latter. This problem of *balance* in the selection of items [21] means that problem-focused general quality of life questionnaires inevitably have a bias towards one disease or another, in the sense that the result of comparisons between diseases are instrument specific. However, if the patient

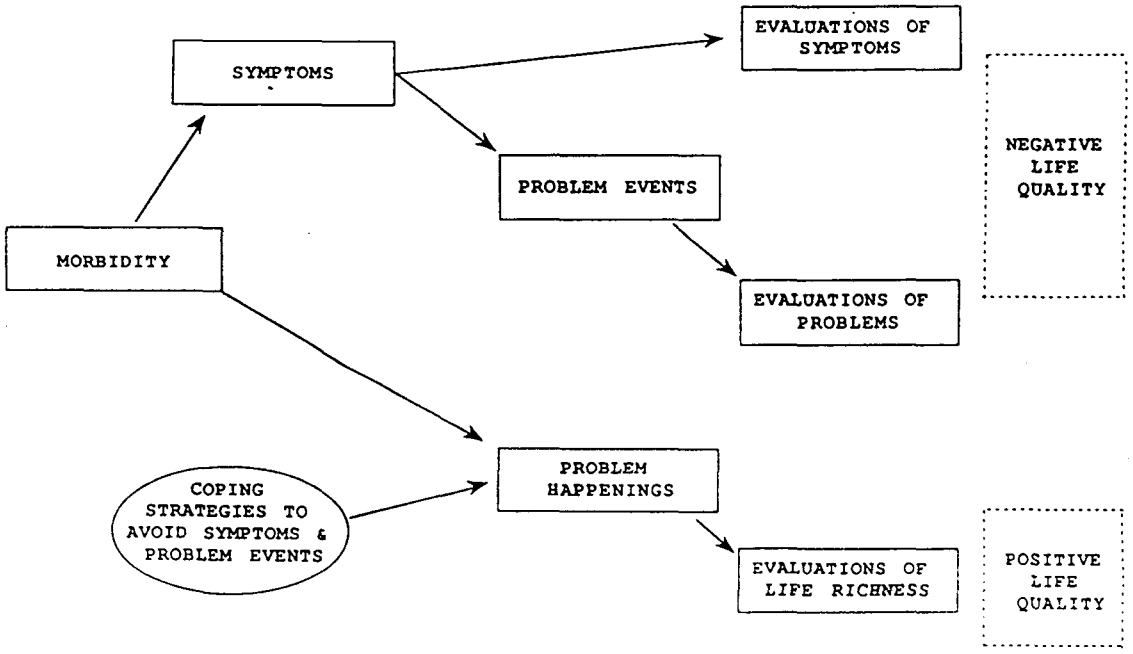

Fig. 1. Theoretical model showing causal sequence of psychological constructs which are relevant to health-related quality of life. Problem events are problems which occur directly as a result of symptoms. Problem happenings are problems which result from actions taken by the patient to avoid symptoms and so, typically, occur in the absence of symptoms. From: Hyland ME. A reformulation of quality of life for medical science. **Quality of Life Research** 1992; 1: 267–272.

makes general evaluations about health, then these evaluations form a "common currency" which is independent of the particular problems experienced. Thus, there is merit in using patient's evaluations of their health in cross-sectional comparisons as a way achieving a "valid" i.e. unbiased method of comparison.

In conclusion, the classification system proposed by Guyatt *et al.* [1] which is based on the use for which the test is put appears to parallel an alternative classification based on what the test purports to measure. If a test provides a (construct) valid measure of the construct of health evaluation, then that test will provide a (content) valid means for comparing between different diseases. However, if the test provides a (construct) valid measure of knowledge of health problems, then the test will provide a (content) valid method for making longitudinal comparisons. My own view is that it makes more sense to classify tests in terms of the constructs they measure and then see how that classification relates to intended use. However, the alternative approach of defining function first and evaluating tests in terms of the adequacy of that function as suggested by Guyatt *et al.* [1] is perfectly feasible.

If there is any recommendation to come from this debate it is this. When claiming validity for a health measure, researchers should specify the purpose of their test and what type of validity they are trying to establish. And if the intention is to establish construct validity, then its *essential* to know what construct the scale is supposed to measure validly!

REFERENCES

1. Guyatt GH, Kirshner B, Jaeschke R. Presentation: Measuring health status: what are the necesary measurement properties. **J Clin Epidemiol** 1992; 45: 1341–1345.
2. Guyatt GH, Kirshner B, Jaeschke R. Response: A methodologic framework for health status measures: clarity of oversimplification? **J Clin Epidemiol** 1992; 45: 1353–1355.
3. Williams JI, Naylor CD. Dissent: How should health status measures be assessed? Cautionary notes on procrustean frameworks. **J Clin Epidemiol** 1992; 45: 1347–1351.
4. Cronbach LJ, Meehl PE. Construct validity in psychological tests. **Psychol Bull** 1955; 52: 281–302.
5. MacCorquodale K, Meehl PE. On a distinction between hypothetical constructs and intervening variables. **Psychol Rev** 1948; 55: 95–107.
6. Anastasi A. **Psychological Testing**, 6th edn. New York: Collier Macmillan; 1988.
7. Nunnally JC. **Psychometric Theory**, 2nd edn. New York: McGraw-Hill; 1978.
8. Hyland ME. Quality of life assessment in respiratory disease: an examination of the content and validity of four questionnaires. **PharmacoEconomics** 1992; 2: 43–53.
9. Hyland ME. Do person variables exist in different ways? **Am Psychol** 1985; 40: 1003–1010.
10. Showers C. Compartmentalization of positive and negative self-knowledge: Keeping bad apples out of the bunch. **J Pers Soc Psychol** 1992; 62: 1036–1049.
11. Lazarus RS, Smith CA. Knowledge and appraisal in the cognition–emotion relationship. **Cognition Emotion** 1988; 2: 281–300.
12. LeDoux JL. Cognitive-emotional interactions in the brain. **Cognition Emotion** 1989; 3: 276–289.
13. Costa PT, McCrae RR. Influence of extraversion and neuroticism on subjective well-being: happy and unhappy people. **J Pers Soc Psychol** 1980; 38: 668–678.
14. Heady B, Wearing A. Personality, life events, and subjective well-being: toward a dynamic equilibrium model. **J Pers Soc Psychol** 1989, 57: 731–739.
15. DeLongis A, Coyne JC, Dakof G, Folkman S, Lazarus RS. Relationship of daily hassles, uplifts, and major life events to health status. **Health Psychol** 1982; 1: 119–136.
16. DeLongis A, Folkman S, Lazarus RS. The impact of daily stress on health and mood: Psychological and social resources as mediators. **J Pers Soc Psychol** 1988; 54: 486–495.
17. Watson D, Pennebaker JW. Health complaints, stress, and distress: exploring the central role of negative affectivity. **Psychol Rev** 1989; 96: 234–254.
18. Hyland ME. A reformulation of quality of life for medical science. **Qual Life Res** 1992; 1: 267–272.
19. Evans RW. Quality of life. **Lancet** 1991; 338:363.
20. Hyland ME, Kenyon CAP. A measure of positive health-related quality of life: The Satisfaction with Illness Scale. **Psychol Rep** 1992; 71: 1137–1138.
21. Hyland ME. Selection of items and avoidance of bias in quality of life scales. **PharmacoEconomics** 1992; 1: 182–190.