Quality of Life Research, 1, pp. 73–75

Controversy

Responsiveness to change: an aspect of validity, not a separate dimension

R. D. Hays* and D. Hadorn

RAND, Social Policy Department, 1700 Main Street, Santa Monica, CA 90407–2138, USA.

Assessment of health-related quality of life is accelerating in naturalistic observational studies, clinical trials, and clinical practice. Some researchers have argued that the ability of a quality of life instrument to detect clinically important changes over time, "responsiveness," is a distinct psychometric property from the measure's reliability and validity. We discuss the important implications of this argument and counter that responsiveness is actually one indication of a measure's validity.

Key words: Reliability, responsiveness, validity.

Interest in assessing health-related quality of life in naturalistic observational studies, clinical trials, and clinical practice has burgeoned.[1–3] Inclusion of quality of life measures in evaluations of the impact of experimental interventions and natural events is becoming standard practice. The quality of these evaluations depends in large part on the extent to which these measures satisfy accepted psychometric standards.[4]

Traditionally, reliability and validity are considered the two fundamental characteristics of a measuring instrument.[5] Reliability refers to the extent to which a consistent score is obtained on different administrations of the instrument when all relevant conditions remain essentially constant. Validity is the extent to which an instrument's scores reflect the construct it is intended to measure and not what it was unintended to measure.

Guyatt, Walter, and Norman[6] recently suggested that another important property of quality of life measures is their responsiveness to clinically important changes. These investigators operationalized responsiveness as the change in quality of life score due to a minimal clinical intervention divided by the fluctuation in quality of life score due to error of measurement. Guyatt *et al.*[6] proposed that responsiveness is distinct from reliability and validity, contending that an instrument can be: (1) reliable, but unresponsive; (2) responsive, but not valid; and (3) unreliable, yet responsive.

Although we agree with their first contention, we have substantial reservations about their second point and note that the third is inconsistent with accepted psychometric theory and practice. We argue that responsiveness is an aspect of validity rather than a separate entity.

Reliable, but unresponsive

We agree that a measure can be reliable, yet unresponsive. This is true because consistent, reproducible (reliable) measurement does not in and of itself indicate that the desired construct (e.g., quality of life) is in fact being measured. Therefore, whether or not an instrument will be responsive to a clinical intervention depends on more than the instrument's reliability.[7] For example, the 20-item Medical Outcomes Study Short-form Health Survey has been shown to be quite reliable,[8] but floor effects on some of the scales have been shown to limit the responsiveness of scores to change in health status.[9]

Responsive, but not valid

Guyatt *et al.*[10] presented results purportedly indicating that the Eastern Co-operative Oncology

Preparation of this paper was supported in part by the World Health Organization (Regional Office for Europe) and by RAND. The views expressed are those of the authors and do not necessarily represent those of the sponsor or RAND.

* To whom correspondence should be addressed.

Group Criteria (ECOG) toxicity questionnaire is responsive, but not a valid health status measure. As hypothesized, cancer patients who received a short trial of chemotherapy reported significantly less deterioration in quality of life than patients receiving a longer trial (because of the toxic side-effects of treatment). This type of testing—wherein an instrument's performance is compared with expected findings—provides an indication of construct validity.[4]

Despite these findings, Guyatt *et al.*[10] claimed that the ECOG instrument was not valid because it is composed of many laboratory measures related to drug toxicity rather than quality of life measure *per se*. Guyatt *et al.*[10] appear to be suggesting that the ECOG measures lack content validity (because they assess domains not considered reflective of quality of life) and that the apparent responsiveness of this instrument is therefore illusory (or irrelevant).

It is certainly true that a measure might change (respond) in a manner similar to that expected for a valid measure, but actually not measure what it is supposed to measure. However, this fact does not support the contention that responsiveness is distinct from validity as a psychometric construct. Rather, it demonstrates that a measure might perform well on one test of validity, but not on another. Validation is an ongoing process of obtaining multiple sources of information and empirical evidence to assess whether the instrument actually measures what it purports to. Each piece of evidence, including the instrument's responsiveness, provides important information about the validity of the measure. If one can argue that a responsive instrument may not be valid, one could just as well argue that an instrument that exhibits content validity may not actually be valid because it does not discriminate known groups, or that an instrument that displays known groups validity may not be valid because of its failure to detect changes in quality of life over time.

A measure that is valid at one time point should also be valid at another time point—"intermittently valid" measures are unlikely to be identified, or even postulated. Accordingly, valid instruments in theory should be responsive to changes over time. To maintain otherwise is to claim that an initially valid instrument may somehow lose its validity and thus no longer be able to measure the underlying construct (i.e., quality of life) at a later time point. Therefore, responsiveness simply incorporates longitudinal information (change) into the process of evaluating validity.

Thus, a quality of life instrument that measures what it is supposed to measure is expected to be responsive to a clinical intervention—it should detect real change in quality of life whether the change is induced experimentally or naturally. Hence, conceptualizing responsiveness as an indicator of validity is consistent with common usage in the psychometric literature (i.e., Does the instrument measure what it is supposed to?). Indeed, other investigators have suggested that sensitivity to change is one test of a measure's construct validity.[11–12]

Unreliable, yet responsive

Guyatt *et al.*[6] provide a hypothetical example of an instrument that is unreliable (test-retest reliability is zero), but responsive to a clinical intervention. This exercise demonstrates that it is possible to conjure up an example to prove almost any point, but hypothetical examples do not always reflect reality.

Extending the argument given above that responsiveness is an indication of an instrument's validity, it would be possible to imagine an example that suggests a measure is valid, yet unreliable. One could argue, for instance, that it is possible for a multi-item quality of life scale to have no internal consistency reliability (negative inter-item correlations), but still perform well in terms of known-groups validity. Indeed, Table 1 provides hypothetical data for a three-item quality of life scale that demonstrates no reliability, but the validity of the scale is supported by its perfect correlation ($r = 1.00$) with housing status (homeless vs. sheltered). Despite this interesting example, it is well known that adequate reliability is required for valid measurement.[5,7,13] A measure that yields inconsistent (unreliable) information about persons will not be a valid (or responsive) measure. It is unreasonable to expect a measure to hit the bullseye (measure what it is supposed to) if it can't even hit the target itself consistently.

Concluding remarks

Reliability, or the extent to which the same information is obtained on repeated administrations when no change should occur, is a necessary, but not sufficient property of a valid quality of life instrument. In addition, a quality of life measure

Table 1. Hypothetical example of no internal consistency reliability, but validity

Person	Group	Quality of life item scores			
		1	2	3	Total
1	Homeless	2	4	3	9
2	Homeless	2	4	3	9
3	Homeless	1	4	4	9
4	Homeless	2	4	3	9
5	Homeless	2	5	2	9
6	Sheltered	5	3	3	11
7	Sheltered	5	3	3	11
8	Sheltered	5	3	3	11
9	Sheltered	5	3	3	11
10	Sheltered	5	3	3	11

Note: Potential item range was 1 to 5. Items were all worded so that a higher score indicates better quality of life. All intercorrelations among items are negative.

should reflect (i.e., be responsive to) the effects of a clinical intervention that changes underlying quality of life. If an instrument is responsive to a clinical intervention, this fact provides some support for the validity of the instrument.

The artificial distinction between responsiveness and validity may in part stem from an emphasis on classifying quality of life instruments dichotomously as either valid or not. In fact, instruments are valid to varying degrees. The most valid quality of life measure should perform favourably on multiple tests of validity including a test of its ability to detect change in quality of life over time.

Acknowledgements

The authors gratefully acknowledge Kirsten Staehr Johansen for her assistance.

References

1. Nelson EC, Berwick DM. The measurement of health status in clinical practice. *Med Care* 1989; **27**: S77–S90.
2. Rubenstein LV, Calkins DR, Greenfield S, *et al*. Health status assessment for elderly patients: Report of the Society of General Internal Medicine task force on health assessment. *J Am Geriat Soc* 1988; **37**: 562–569.
3. Wells KB, Stewart AL, Hays RD, *et al*. The functioning and well-being of depressed patients: Results from the Medical Outcomes Study. *J Am Med Assoc* 1989; **262**: 914–919.
4. Stewart AL, Hays RD, Ware JE. Methods of validating health measures. In: Stewart AL, Ware JE, eds. *Measuring Functioning and Well-being: The Medical Outcomes Study Approach*. Durham, NC: Duke University Press, in press.
5. Stewart AL. Psychometric considerations in functional status instruments. In: Lipkin M, ed. *Functional Status Measurement in Primary Care*. New York: Springer-Verlag, 1990: 3–26.
6. Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. *J Chron Dis* 1987; **40**: 171–178.
7. Spitzer RL, Fleiss JL. A re-analysis of the reliability of psychiatric diagnosis. *Br J Psychiat* 1974; **125**: 341–347.
8. Stewart AL, Hays RD, Ware JE. The MOS short-form general health survey: Reliability and validity in a patient population. *Med Care* 1988; **26**: 724–735.
9. Bindman AB, Keane D, Lurie N. Measuring health changes among severely ill patients: The floor phenomenon. *Med Care* 1990; **28**: 1142–1152.
10. Guyatt G, Deyo RA, Charlson M, *et al*. Responsiveness and validity in health status measurement: A clarification. *J Clin Epidemiol* 1989; **42**: 403–408.
11. Boyle MH, Torrance GW. Developing multiattribute health indexes. *Med Care* 1984; **22**: 1045–1057.
12. Chambers LW. Physical and emotional function of primary care patients: Scientific requirements for the measurement of functional health status. *J Am Med Assoc* 1983; **249**: 3353–3355.
13. Thorndike RL. Reliability. In: Jackson DN, Messick S, eds. *Problems in Human Assessment*. New York: McGraw-Hill, 1967: 217–240.

(*Received 8 October 1991; accepted 13 November 1991*)