Published in final edited form as:

Res Nurs Health. 1993 December ; 16(6): 459–465.

Focus on Psychometrics Single-Item Indicators in Nursing Research

JoAnne M. Youngblut, PhD, RN [Associate Professor] and Gail R. Casper, MS, RN [Doctoral student]

Frances Payne Bolton School of Nursing, Case Western Reserve University

Abstract

Single-item indicators that ask respondents for their global rating of a specific concept are congruent with nursing's emphasis on wholism and individualism. They allow the subject to take personally salient features of the situation into account when providing a response. The psychometric performance of single-item indicators in published research and in a sample data set using measures of the mother's choice and satisfaction with her employment decision support the validity and reliability of the measures, suggesting that these indicators deserve more attention in nursing research. Recommendations for the use of single-item indicators are provided.

The purpose of this article is to explore the use of single-item indicators in nursing research. Single-item indicators can be classed into two fundamentally different groups. In the first group are single-item instruments that were designed as single-item measures. These measures generally are used to obtain the subject's perception of particular dimensions of multidimensional concepts or of an overall concept. Global single-item measures allow the subject to define the concept in a way that is personally meaningful, providing a measure that can be responsive to individual differences. Global single-item indicators require that subjects consider all aspects of a phenomenon, ignore aspects that are not relevant to their situations, and differentially weight the other aspects according to their values and ideals in order to provide a single rating. They represent a holistic way to measure subjects' perceptions of many concepts that are of interest to nursing and are consistent with nursing's perspective. Whenever nurse researchers are interested in individuals' perceptions of a particular situation, perhaps in order to predict their behavior, a global single-item indicator may be a more valid measure of the concept of interest.

The second group consists of single-item indicators that were designed as part of a multi-item scale. These single-item measures often are used in a particular study when the multi-item scale does not perform satisfactorily or as proxy measures for a concept that seems important in understanding the findings but that was not measured in a systematic way. The problem with this practice lies in the way multi-item measures are constructed. Items in a multi-item instrument generally are chosen so that each represents one aspect of the concept, with the goal of adequately sampling from the domain of possible items (Nunnally, 1978). Although this strategy is appropriate and important to the construction of a multi-item scale, it means that a composite of the items is necessary in order to validly measure the concept. Using one item originally designed as part of a multi-item scale may not provide a complete picture of the concept.

A study of single- versus multi-item measures serves as an example of this problem. Bukowski, Ferber-Goff, and Newcomb (1990) measured antisocial behavior in school children with five items, each measuring a different behavior that is part of the larger construct. In order to compare 1-month test-retest reliability for scales made of different numbers of items, they chose one item as the single-item measure and then progressively added the other items to it. Test-retest reliability for the two- to five-item measures ($r = .67$ to $.85$) was higher than for the single-item measure ($r = .55$ and $.69$), even after correction for attenuation due to measurement error. They reason that the increased reliability of the multi-item scales is due to a factor other than the number of items included in the scale. Since scales based on more items usually have an increased range of scores which increases the correlation coefficients, other things being equal (Pedhazur, 1982), the higher reliability for the multi-item measures could be due to differences in score range and not to real differences in reliability. Bukowski et al. did not explain how they chose the single item or the order of item entry into the score. Thus, the differences in correlations also could be due to the content of the specific items and not to the number of items. Single-item measures in this second group should not be used for testing study hypotheses in planned, a priori analyses, but may be beneficial in exploring the data for suggestions for future research.

PSYCHOMETRIC PERFORMANCE OF GLOBAL SINGLE-ITEM INDICATORS

Studies that included single-item indicators to measure physiologic and psychologic concepts are reviewed. The studies were selected not to convey an exhaustive review of studies using single-item indicators but, rather, to present a range of how single-item indicators have performed in various situations. Data from a study of parental and family reactions to the preterm birth of an infant are used as an additional example of the psychometric performance of global single-item indicators. The two single-item indicators (*How much choice did you have regarding your employment decision?* and *How satisfied are you with your decision?*) were used in studying maternal employment issues for mothers with preterm infants (Youngblut, Loveland-Cherry, & Horan, 1990; 1991; 1993).

Single-item measures have been used frequently in large population surveys such as *The quality of American life* (Campbell, Converse, & Rodgers, 1976), *Social indicators of well-being: Americans' perceptions of life quality* (Andrews & Withey, 1976), and *Functional status and well-being of patients with chronic conditions: Results from the medical outcomes study* (Stewart et al., 1989). The use of single-item indicators in clinical research is increasing, especially in the measurement of symptom intensity. Hoddes, Zarcone, Smythe, Phillips, and Dement (1973) used the one-item Stanford Sleepiness Scale in a study of the effects of sleep deprivation on memory and vigilance, and Lee, Hicks, and Nino-Murcia (1991) used the Stanford Sleepiness Scale to evaluate the psychometric properties of their newly developed, multi-item measure of fatigue. Many researchers measure pain intensity with single-item scales in adults (Bondestam et al., 1987; Price, McGrath, Rafii, & Buckingham, 1983; Sjoden, Bates, & Nyren, 1983) and children (Beyer, Denyes, & Villarruel, 1992). Sensations of dyspnea also have been measured with single-item scales in both adults (Gift, Plaut, & Jacox, 1986; Janson-Bjerklie, Ruma, Stulbarg, & Carrieri, 1987) and children (Carrieri, Kieck-hefer, Janson-Bjerklie, & Souza, 1991). Perceived degree of success in managing chronic illness (Lowery & Jacobsen, 1984) and satisfaction with health care (Sutherland et al., 1989) have been measured by a single item. An analysis of how single-item indicators have performed in these studies follows.

Reliability

In the studies reviewed, single-item indicators were reliable measures of the phenomenon under study and were influenced by factors that affect reliability estimates for multiple-item scales. In early work by Andrews and Withey (1976), test-retest reliability for a single-item

rating scale to measure quality of life was about .70. The interval between testings was very short, with the measure being administered twice in the same interview. In a national study of the quality of American life, Campbell et al. (1976) found a correlation of .43 between the two ratings of a single-item rating scale of global life satisfaction measured 8 months apart and a correlation of .53 between two ratings of a nine-item composite Index of Well-Being (Cronbach's $\alpha = .89$).

Cella and Perry (1986) measured anxiety, depression, and distress each with a single visual analogue scale at four time points in one day (first thing in the morning, noon, 4 p.m., and bedtime). They found test-retest correlations among these four measures that ranged from .50 to .83, with higher coefficients between measures taken when the subject's activity level was similar. Thus, as expected, correlations between measures from inactive times (morning and bedtime) and between measures from active times (noon or 4 p.m.) were higher than correlations between an active time measure and an inactive time measure.

Gift et al. (1986) compared ratings of dyspnea on vertical and horizontal visual analogue scales provided at 6 time points at half-hour intervals by adults who were seen in an emergency room for an acute asthma attack. Correlations for each subject ranged from .52 to .99. Most of these reliability estimates are acceptable, with longer time intervals between testing and changes in the phenomenon under study reflected in lower test-retest estimates.

Mothers in the premature infant project rated degree of choice and satisfaction with their employment decision each on a single 10-point rating scale ranging from 1 (*no choice or not at all satisfied*) to 10 (*totally my choice or very satisfied*). These ratings were obtained when the infant was 3, 9, and 18 months of age. Correlations among measures of choice ($r = .52$ to $.60$) and among measures of satisfaction ($r = .45$ to $.65$) indicate moderate stability of the measures across time (Youngblut, 1993). Due to the nature of the underlying concepts, some fluctuation in the ratings over time is expected. However, repeated-measures ANOVA testing for choice and satisfaction was not statistically significant. This finding raises the question of whether these fluctuations reflect error variance or variance due to the effect of omitted variables. Some of the women changed employment status between time points and others experienced a subsequent pregnancy and birth. Information about characteristics of the job was not collected, so it is not known how these characteristics might have affected the women's ratings. These factors would produce fluctuations that vary across women, resulting in lower correlations than desired for test-retest reliability. Thus, the magnitude of the correlations suggests that the measures are sensitive to change over a 15-month period (Stewart & Archbold, 1993). These results support the reliability of these two single-item measures and are consistent with the findings of the other studies reviewed.

Validity

Evidence available to support the construct validity of single-item measures is substantial, unlike that for reliability, and indicates that single-item measures are generally valid and sensitive to change in the phenomenon under study. Ratings of a single visual analogue scale for depression, anxiety, and distress by families of burn patients correlated with the Beck Depression Inventory ($r = .58$), the Spielberger State Anxiety Inventory ($r = .52$), and Perceived Stress Scale ($r = .63$), respectively (Cella & Perry, 1986). Carrieri et al. (1991) used three different single-item indicators to measure children's perception of dyspnea: a visual analogue scale, a color scale, and a 4-point rating scale with word anchors for each point. Children's ratings of dyspnea intensity were significantly different ($p < .001$) for good breathing days and for bad breathing days with each scale, supporting the construct validity of each scale. Ratings of dyspnea intensity on the visual analogue scale by adults with asthma were significantly related to number of cigarette pack years ($r = .34$, $p = .05$) and frequency of asthma attacks ($r = .35$, $p = .05$), as expected (Janson-Bjerklie et al., 1987).

Gift et al. (1986) used the visual analogue dyspnea scale to identify times of high, medium, and low dyspnea in each of the COPD patients in the study. Use of accessory muscles was significantly greater during self-rated high dyspnea times compared to low dyspnea times ($p < .01$). Subjects' scores for anxiety on Spielberger's State Anxiety Inventory and for anxiety and somatization on the Brief Symptom Inventory also were significantly higher during high dyspnea times compared to low dyspnea times ($p < .01$).

Hoddes et al. (1973) used the Stanford Sleepiness Scale, a 7-point scale with phrase description anchors, to measure degree of sleepiness when subjects were fully rested and during a period of sleep deprivation. Subjects' ratings of sleepiness were significantly higher (indicating greater sleepiness) during the sleep deprivation period compared to ratings made after a night's sleep. Lee et al. (1991) found significantly higher sleepiness ratings on the Stanford Sleepiness Scale in the evening compared to ratings the next morning for healthy subjects, but not for subjects undergoing evaluation for sleep disorders.

Cunney and Perri (1991) report a correlation of .86 between the Medical Outcomes Study short-form General Health Survey and a single-item indicator of overall health-related quality of life (*In general would you say your health is Excellent, Very Good, Good, Fair, or Poor?*) However, it is not clear from the report whether this item was removed from the General Health Survey score before computing the correlation coefficient. If it was included in the total score, the correlation coefficient may be artificially inflated.

Lowery and Jacobsen (1984) reported a 79% agreement between the single-item rating by patients and by physicians regarding how well the patients were doing with their chronic illness. Bondestam et al. (1987) report 76% agreement between Coronary Care Unit (CCU) patients' pain ratings on a numerical rating scale and CCU nurses' ratings of the patients' pain on the same numerical scale. Beyer et al. (1992) have recently reported the results of their extensive validity testing of the Oucher scale for the measurement of pain in children. The Oucher has been sensitive to expected changes in postoperative pain as time since surgery increased, to expected changes in pain after analgesia administration, and to differences in postoperative pain scores for children experiencing less-extensive surgical procedures compared with children undergoing more-extensive surgical procedures. Children's ratings of pain on the Oucher, the Poker Chip Tool (Hester, 1979), and a visual analogue scale were strongly correlated ($r = .70$ to $.98$). Findings in all of these studies support the construct validity of the single-item scales used.

Construct validity testing of the single-item measures of choice and satisfaction in the preterm infant project (Loveland-Cherry & Horan, 1986) was based on two hypotheses. The first hypothesis tested was: Women whose employment attitudes and employment status are inconsistent will report less choice and satisfaction regarding their employment decision than women whose employment attitudes and employment status are consistent. At each of the three time points, consistent women rated their degree of choice and satisfaction significantly higher than inconsistent women (Table 1). The second hypothesis tested was: Choice and satisfaction will be positively related to the positive affect scales (contentment, vigor, affection, joy) and negatively related to the negative affect scales (depression, hostility, anxiety, guilt) of the Affects Balance Scale (Derogatis, 1975). Alpha coefficients for these affect scales at each time point were adequate, ranging from .72 to .88 in this sample, except for anxiety which had an alpha of .65 (T1), .64 (T2), and .72 (T3). At each time point, choice and satisfaction were positively related to contentment. In addition, satisfaction at each time point was negatively related to depression. At T3, choice was significantly related to six of the eight scales and satisfaction to seven of the eight scales. The direction of the significant relationships were all as expected (Table 2). These findings support the construct and discriminant validity of these two single-item indicators.

DISCUSSION

In the studies reviewed as well as in the data analysis presented for illustration, most of the reliability and validity estimates were acceptable. Although it is possible that studies with single-item indicators whose reliabilities were low are more common but are not reported in the literature because of their low reliabilities, findings of the studies reviewed suggest that single-item measures can yield acceptable reliability estimates. Factors that affect reliability for single-item measures appear to be the same factors that affect reliability in multi-item measures, such as the wording of the item(s), characteristics of the sample, and specifics of the testing situation (Nunnally, 1978).

Global single-item indicators generally performed well in validity testing. When the variable of interest is the person's overall perception, global single-item measures may yield more valid data than multi-item measures. Using a multi-item measure means that the researcher selects items that address different aspects of the concept and then computes a score based on the responses to these items. If a simple summative score is used, the researcher is weighting each item equally. If a factor score is used, the weights are derived from the total sample (Nunnally, 1978) and may not represent the weights each individual would choose. Single-item indicators that ask the respondents for their global appraisal or perception of a concept allow the respondents, rather than the investigators, to consider the factors that are important to them and to differentially weight these aspects in a way that makes sense for them as individuals. Thus, a global single-item indicator often provides a valid measure of the concept that can be sensitive to individual differences.

Based on evidence about the use of single-item measures, several recommendations can be offered. First, when the researcher's focus is on the individual as a whole, the use of holistic measures is appropriate. Single-item indicators often provide valuable information about an individual's perception of the concept under study. These global perceptions may be important when studying a subject's appraisal of health status, quality of life, satisfaction with health care, or level of symptom intensity experienced. If the researcher believes that the person's perception of a specific concept is important, then a single-item indicator may be the most appropriate measurement method.

Second, when choosing a single-item indicator, it is preferable to construct one that asks the respondent for a global rating rather than to use one item from a multi-item scale. Reliability and validity estimates of single-item measures show a consistent pattern across studies regardless of the response format (e.g., numerical, with or without word anchors; visual analogue; graphic representations) used. Thus, the format that is most appropriate for the sample can be chosen based on characteristics such as age, education, and acuity. Test-retest reliability is the most appropriate estimate for single-item indicators but, when planning the time interval between testings, the researcher must take into consideration the concept's expected rate of change. Since reliability sets the upper limit on validity (Nunnally, 1978), reliability values always will exceed reported validity estimates. Construct validity testing of single-item measures provides an estimate of both the measure's validity and a lower bound on its reliability. Thus, if test-retest reliability estimates are not available, validity correlations between the single-item measure and a measure that is posited to be related may be considered as a substitute for reliability estimates for the single-item indicator.

Finally, for studies where multiple indicators of the same concept are desirable, a single-item scale may be an acceptable second measure. In this case, the investigator would choose a multi-item measure of the concept as the primary indicator and then construct a single-item measure that captures the respondent's overall perception of the concept. This method would provide the necessary second indicator of the concept of interest without a

considerable increase in the response burden for the subjects. Using a multi-item measure and a single-item measure in the same study also would provide the opportunity to investigate the construct validity of the single-item measure in relationship to the multi-item one.

Support for the use of single-item measures often has been framed in terms of limited time and resources and condition of the respondent. These factors frequently do affect the length and content of the instrument (Cunney & Perri. 1991). However, the relationship of the method to the research question is the priority. Simplicity, economy, and ease of administration may be factors in the decision to use single-item instruments, but a more important consideration is that they capture the phenomenon of interest (Wewers & Lowe, 1990). Single-item indicators often have acceptable psychometric properties and, thus, are a viable alternative for measuring global concepts of interest to nursing.

Acknowledgments

The authors acknowledge Dr. Carol Loveland-Cherry and Dr. Mary Horan for granting access to the data set. The larger project was supported by Grant No. NR01390 from the National Center for Nursing Research awarded to C. Loveland-Cherry and M. Horan.

References

- Andrews, FM.; Withey, SB. Social indicators of well-being: Americans' perceptions of life quality. New York: Plenum; 1976.
- Beyer JE, Denyes MJ, Villarruel AM. The creation, validation, and continuing development of the Oucher: A measure of pain intensity in children. *Journal of Pediatric Nursing*. 1992; 7:335–346. [PubMed: 1479552]
- Bondestam E, Hovgren K, Johansson FG, Jern S, Herlitz J, Holmberg S. Pain assessment by patients and nurses in the early phase of acute myocardial infarction. *Journal of Advanced Nursing*. 1987; 12:677–682. [PubMed: 3693724]
- Bukowski WM, Ferber-Goff J, Newcomb AF. The stability and coherence of aggregated and single-item measures of antisocial behavior. *British Journal of Social Psychology*. 1990; 29:171–179. [PubMed: 2372666]
- Campbell, A.; Converse, PE.; Rodgers, WL. The quality of American life. New York: Russell Sage; 1976.
- Carrieri VK, Kieckhefer G, Janson-Bjerklie S, Souza J. The sensation of pulmonary dyspnea in school-age children. *Nursing Research*. 1991; 40:81–85. [PubMed: 2003078]
- Cella DF, Perry SW. Reliability and concurrent validity of three visual-analogue mood scales. *Psychological Reports*. 1986; 59:827–833. [PubMed: 3809343]
- Cunney KA, Perri M. Single-item vs. multiple-item measures of health-related quality of life. *Psychological Reports*. 1991; 69:127–130. [PubMed: 1961779]
- Derogatis, L. Affects Balance Scale. Baltimore: Clinical Psychometric Research; 1975.
- Gift AG, Plaut M, Jacox A. Psychologic and physiologic factors related to dyspnea in subjects with chronic obstructive pulmonary disease. *Heart & Lung*. 1986; 15:595–601. [PubMed: 3639857]
- Hester N. The preoperational child's reaction to immunizations. *Nursing Research*. 1979; 28:250–254. [PubMed: 255924]
- Hoddes E, Zarcone V, Smythe H, Phillips R, Dement WC. Quantification of sleepiness: A new approach. *Psychophysiology*. 1973; 10:431–436. [PubMed: 4719486]
- Janson-Bjerklie S, Ruma SS, Stulbarg M, Carrieri VK. Predictors of dyspnea intensity in asthma. *Nursing Research*. 1987; 36:179–183. [PubMed: 3646619]
- Lee KA, Hicks G, Nino-Murcia G. Validity and reliability of a scale to assess fatigue. *Psychiatry Research*. 1991; 36:291–298. [PubMed: 2062970]
- Loveland-Cherry, CJ.; Horan, M. Parental/family factors in high risk infant development. National Center for Nursing Research; 1986. No. R01 NR01390

- Lowery BJ, Jacobsen BS. Attributional analysis of chronic illness outcomes. *Nursing Research*. 1984; 34:82–88. [PubMed: 3844736]
- Nunnally, JC. *Psychometric theory*. New York: McGraw-Hill; 1978.
- Pedhazur, EJ. *Multiple regression in behavioral research*. 2. New York: Holt, Rinehart, and Winston; 1982.
- Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain*. 1983; 17:45–55. [PubMed: 6226917]
- Sjoden PO, Bates S, Nyren O. Continuous self-recording of epigastric pain with two rating scales: Compliance, authenticity, reliability and sensitivity. *Journal of Behavioral Assessment*. 1983; 5:327–345.
- Stewart AL, Greenfield S, Hays RD, Wells K, Rogers WH, Berry SD, McGlynn EA, Ware JE. Functional status and well-being of patients with chronic conditions: Results from the Medical Outcomes Study. *Journal of the American Medical Association*. 1989; 262:907–913. [PubMed: 2754790]
- Stewart BJ, Archbold PG. Nursing intervention studies require outcome measures that are sensitive to change: Part two. *Research in Nursing & Health*. 1993; 16:77–81. [PubMed: 8488315]
- Sutherland HJ, Lockwood GA, Minkin S, Tritchler DL, Till JE, Llewellyn-Thomas HA. Measuring satisfaction with health care: A comparison of single with paired rating strategies. *Social Science and Medicine*. 1989; 28:53–58. [PubMed: 2928814]
- Wewers ME, Lowe NK. A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health*. 1990; 13:227–236. [PubMed: 2197679]
- Youngblut, JM. Consistency between employment attitudes and behaviors for mothers of preterm infants. 1993. Submitted for publication
- Youngblut JM, Loveland-Cherry CJ, Horan M. Factors related to maternal employment status following the premature birth of an infant. *Nursing Research*. 1990; 39:237–240. [PubMed: 2367205]
- Youngblut JM, Loveland-Cherry CJ, Horan M. Maternal employment effects on family and preterm infants at three months. *Nursing Research*. 1991; 40:272–275. [PubMed: 1896324]
- Youngblut JM, Loveland-Cherry CJ, Horan M. Maternal employment, family functioning, and preterm infant development at 9 and 12 months. *Research in Nursing & Health*. 1993; 16:33–43. [PubMed: 8488311]

Discriminant Validity: Comparison of Consistent and Inconsistent Women on Ratings of Choice and Satisfaction

Table 1

		T1 Consistent (n=72)	T1 Inconsistent (n=30)	t value
Time 1 (T1)	Choice ^a	M (SD) 8.44 (2.66)	5.30 (3.73)	4.19 **
	Satisfaction ^b	M (SD) 8.64 (1.97)	5.93 (2.91)	4.66 **
		T2 Consistent (n=66) T2 Inconsistent (n=22)		t value
Time 2 (T2)	Choice	M (SD) 8.08 (2.80)	5.14 (2.92)	4.22 **
	Satisfaction	M (SD) 8.70 (1.84)	5.50 (2.77)	5.05 **
		T3 Consistent (n=71) T3 Inconsistent (n=30)		t value
Time 3 (T3)	Choice	M (SD) 7.94 (2.97)	5.80 (3.28)	3.21 *
	Satisfaction	M (SD) 8.56 (1.87)	6.17 (3.38)	3.65 **

^aChoice scored (1) no choice to (10) totally my choice.
^bSatisfaction scored (1) not at all satisfied to (10) totally satisfied.
* p < .01.
** p < .001.

Table 2
Construct Validity: Correlations Between Choice and Satisfaction and the Affects Balance Scales

	Time 1		Time 2		Time 3	
	Choice	Satisfaction	Choice	Satisfaction	Choice	Satisfaction
Joy	.02	.10	.02	.14	.24**	.25**
Contentment	.17*	.22**	.17*	.27**	.40***	.42***
Vigor	.04	.20*	.01	.15	.09	.17*
Affection	.02	.12	-.05	.10	.19*	.21*
Anxiety	-.09	-.14	.07	-.10	-.17*	-.21*
Depression	-.10	-.16*	-.03	-.19*	-.25**	-.29***
Guilt	-.09	-.08	.09	-.04	-.13	-.14
Hostility	-.11	-.05	-.02	-.19*	-.25**	-.26**

Note. All correlations are cross sectional, T1 choice and satisfaction with T1 affect scales, T2 choice and satisfaction with T2 affect scales, T3 choice and satisfaction with T3 affect scales.

* $p < .05$.

** $p < .01$.

*** $p < .001$.