NOTE ON RESEARCH METHODOLOGY

Phrase completions: An alternative to Likert scales

David R. Hodge and David Gillespie

Likert scaling, introduced by Rensis Likert (1932, 1970), is the most widely used method of measuring personality, social, and psychological attitudes (Babbie, 1998; Nunnally, 1978). For example, prominent measures of self-esteem, depression, alienation, locus of control, ethnocentrism, racism, religiosity, spirituality, and homophobia have all used Likert scales to make operational the underlying latent construct (Hill & Hood, 1999; Raja & Stokes, 1998; Robinson, Shaver, & Wrightsman, 1991).

In addition to numerous established measures, a review of the social work literature reveals that researchers commonly use Likert scales in the development of new instruments that tap a broad array of constructs. Likert scales have been used to measure adolescent concerns that foster runaway behavior (Springer, 1998); appropriate practitioner responses to suicidal clients (Neimeyer & Bonnelle, 1997); homesickness and contentment among Asian immigrants (Shin & Abell, 1999); social worker empowerment (Frans, 1993); Spanish-speaking clients' perceptions of social work interventions in prenatal care programs (Julia, 1993); willingness to seek help (Cohen, 2000); attitudes toward illegal aliens (Ommundsen & Larsen, 1998); punishment (Chung & Bagozzi, 1997); and working with clients with HIV/AIDS (Riley & Greene, 1993).

The popularity of Likert scales can be traced to a number of factors, including ease of construction, intuitive appeal, adaptability, and usually good reliability (Babbie, 1998; Nunnally, 1978). Yet, despite these assets there are significant problems associated with Likert scales. This article delineates the problems, with a particular emphasis on multidimensionality and coarse response categories, and proposes a new measurement method called "phrase completions," which has been designed to circumvent the problems inherent in Likert scales. We also conducted an exploratory test, in which Likert items were adapted to phrase completions.

OVERVIEW OF LIKERT SCALES

In contemporary usage, Likert scales present individuals with positively or negatively stated propositions and solicit respondents' opinions about the statements through a set of response keys. Typically, participants are asked to indicate their level of agreement or disagreement with a proposition on a graded four- or five-point scale (for example, strongly disagree, disagree, agree, strongly agree). The fifth point, when used, allows for a neutral or undecided selection to be incorporated into the response key as a midpoint response option.

Agreement with a positively stated proposition is hypothesized to reveal the underlying construct. The responses are usually equated with integers (for example, strongly agree = 0, strongly disagree = 4). Negatively worded items are reverse scored. The items are summed, creating an index that indicates the degree to which the respondent exhibits the traits in question (Duncan & Stenbeck, 1987; Roberts, Laughlin, & Wedell, 1999).

Although the agree–disagree format is perhaps the most common form of Likert scale, other types of response keys also are widely used (for example, very unmotivated, moderately unmotivated, indifferent, moderately motivated, very motivated; below average, slightly below average, average, slightly above average, above average; and so forth). Instruments using these formats are sometimes referred to as Likert-type scales or more generically as rating scales. However, this basic approach, a positively or negatively stated proposition followed by a graduated response key using adverbs and verbs, is commonly understood as the distinguishing characteristic of Likert scales (Brody & Dietz, 1997; Duncan & Stenbeck, 1987).

PROBLEM OF MULTIPLE DIMENSIONS

One of the principle tenets in constructing instruments is that items be as clear and concise as possible. The more items are characterized as cognitively complex, the more likely respondents are to misunderstand the question and answer incorrectly. Even small differences in wording can increase the level of cognitive noise and dramatically alter response patterns (Babbie, 1998). In short, indicators incorporating more than one dimension into an item may increase measurement error by increasing the level of cognitive complexity.

Furthermore, to sum items in the creation of an index, the response keys must be unidimensional (Brody & Dietz, 1997). In other words, the summation of ordinal-level data requires that the units

CCC Code 1070-5309/03 $3.00

that make up the response key consist of ordered categories along a single dimension. If more than one dimension exists in the response key, then the results are confounded, because it is impossible to distinguish the dimensions.

Because of their design, Likert questions ask individuals to think along at least two different dimensions—content and intensity (Brody & Dietz, 1997; Duncan & Stenbeck, 1987). Respondents must evaluate the content of each stated proposition; they must examine the item and decide whether they agree or disagree with the content of the stated proposition. In addition, respondents must assess their level of intensity regarding the stated proposition—they must evaluate how strongly they feel about the proposition (for example, strongly or not strongly).

In effect, Likert scales confound cognitive (content) and affective (intensity) dimensions by incorporating both dimensions into the response key. Researchers have attempted to address the problem of multiple dimensions by separating the cognitive and affective dimensions into two discrete response keys, one that asks participants to indicate whether they agree or disagree with the proposition and the other which asks participants whether they feel strongly or not strongly about the statement. Analysis of the two-question approach, however, suggests little improvement (Brody & Dietz, 1997).

In short, Likert items do not produce unidimensional ordinal responses, thus violating a central measurement tenet. Furthermore, the multiple dimensions inherent in Likert response items may increase measurement error by increasing the level of cognitive noise, a problem that is accentuated by the use of negatively worded items.

Negatively Worded Items and Increased Complexity

As mentioned earlier, Likert scales commonly incorporate the use of negatively worded statements (for example, propositions that use the term "not" and similar variations). Negatively worded items are used to circumvent the problem of response set bias, the tendency of respondents to agree with a series of positively worded items (Cronbach, 1946). To counter this tendency, negatively stated statements are interspersed with positively worded statements and then reverse scored before summing the items to achieve a total score (Nunnally, 1978).

The use of negatively worded statements, however, increases the level of cognitive complexity. The Raja and Stokes (1998) update of Hudson and Ricketts's (1980) widely used instrument designed to measure attitudes toward homosexuals illustrates the added cognitive load that occurs with negatively worded items. Although agreement with roughly half the positively worded statements indicates favorable attitudes toward lesbians, item number 9 states, "I would not vote for a political candidate who was openly lesbian." Respondents must disagree with this negatively worded proposition to be scored as having a positive attitude toward lesbians. Thus, in addition to having to think across two dimensions, respondents have the added cognitive load of having to conceptually synthesize a double negative, **not** vote for, and **dis**agree to be coded as exhibiting positive attitudes toward lesbians.

In short, some individuals have difficulty expressing agreement with the underlying construct by disagreeing with a negatively phrased item (Garg, 1996). Consequently, the added complexity associated with negatively stated items results in lower levels of validity and reliability (Barnette, 2000; Chan, 1991).

Problematic Nature of Disagreement

A similar set of issues also may come into play when respondents disagree with positively worded statements. As mentioned earlier, the Likert response key is divided into positive (agree, strongly agree) and negative (disagree, strongly disagree) response categories. When a positively worded proposition is theorized to indicate the underlying attribute, it seems reasonably clear that agreement indicates the presence of the hypothesized construct, because the item's designer and the respondent are operating in concert with one another (Roberts et al., 1999). Similarly, it is also probable that the level of agreement is relatively likely to indicate the degree of the underlying attribute. In other words, within the constraints of the Likert method, stronger agreement may well denote a greater level of the underlying construct.

It is, however, much more nebulous what disagreement signifies. All the designer really knows is that the respondent does not agree with the positively worded statement that was hypothesized to indicate the underlying trait. From a theoretical standpoint, what sort of information disagreement yields is questionable because the response is not in concert with the designer's theorizing. In short, disagreement is more problematic because respondents may disagree with a statement for any number of reasons.

Consequently, it is of particular concern when disagreement is hypothesized to indicate the presence

of a particular trait. For example, agreement with the positively worded statement "school curricula should include positive discussion of lesbian topics," from Raja and Stokes' (1998) instrument, can reasonably be assumed to indicate a favorable attitude toward lesbians. However, it is far from clear that disagreement connotes homophobia, as Raja and Stokes posit. Respondents may, for example, strongly disagree, not because they are homophobic but simply because they believe that school curricula should be devoted to "the three Rs" in light of the low levels of achievement recorded by Americans in international scholastic ratings. In effect, this approach takes the added noise inherent in the disagree categories and codes it as homophobia, inflating the incident of homophobia.

Midpoint Response Option and Its Attributed Value

Additional problems occur with the use of an odd number of response categories, such as five-point scales. These response keys use a midpoint response category (for example, undecided, indifferent, neutral, no opinion, neither agree nor disagree). The central issue is an understanding of exactly what is being signified by the selection of the midpoint response option in light of the multiple dimensions innate in Likert scales.

The midpoint response is commonly placed between the agree and disagree responses, as the name implies midway in the response key. A midpoint response is theorized to indicate a level of the underlying trait or attribute that is somewhere between the levels signified by agreement and disagreement in a categorical continuum; midpoint responses are coded to indicate a value higher than the disagree options and lower than the agree options. In essence, the middle response option signifies the midpoint on the intensity dimension and is coded as such in the creation of an index (Raaijmakers, van Hoof, Hart, Verbogt, & Vollebergh, 2000).

Many respondents, however, understand the midpoint category as a discrete response option unrelated to the intensity of their response (Raaijmakers et al., 2000). Although some respondents grade their intensity on a continuum, from strongly agree through neutral to strongly disagree at the other end of the continuum, others understand the midpoint option in a manner analogous to a "don't know" or "not applicable" (NA) response category. These individuals see the midpoint option as an extension of the content dimension, viewing it as an option when they do not have enough information to answer the question. For instance to follow up on the earlier example, individuals may not feel sufficiently familiar with school curricula to be able to comment on what material should or should not be included in school settings.

When the midpoint is understood as an NA response, it is appropriate to remove such responses when calculating the scored total (Raaijmakers et al., 2000); that is, an NA response is attributed no specific value. Yet, with five-point Likert scales, what in many cases is an NA response is attributed a higher value than disagree or strongly disagree. Thus, whenever a five-point scale is used, a considerable degree of error may be incorporated because the midpoint category is always attributed a value, regardless of whether the midpoint reflects an NA response related to content or an actual midpoint in terms of intensity.

Summed Ordinal Data

The values attributed to a set of Likert questions are usually summed to create an index. The total is typically treated as either interval- or ratio-level data for statistical analysis. Thus, a rough five-point ordinal scale has been converted into apparent interval- or ratio-level data that are presumed to be suitable for analysis with, for example, multiple regression (Byrne, 1998; Duncan & Stenbeck, 1987).

It is, however, unclear how ordinal data are transformed into interval- let alone ratio-level data merely through the summation process. By definition, interval data comprise equal or uniform units of measurement such as meters, years, and degrees of temperature. These characteristics distinguish interval data from ordinal data, which provide a rank order to data made up of units having an unknown distance between them.

The difference also can be understood in terms of information (Russell & Bobko, 1992). Ordinal items contain less information than interval items that contain less information than ratio items. Consequently, it is possible to collapse higher levels of information into lower levels by discarding some of the obtained information. Interval data, for example, may be reduced to ordinal-level data.

Although it is possible to discard information collected, it is not possible to go in the other direction. Information that was not originally compiled is lost forever. An ordinal measure cannot be converted into an interval one (Babbie, 1998). From a strict theoretical standpoint, summing ordinal-level data cannot transform that data into interval data.

It is merely summed ordinal-level data. Whereas new information can be added, in the form of more items, there is no guideline as to how the distance between any two values is affected.

Thus, although many statistical procedures are robust and can withstand violation of the underlying assumptions, it is important to note that the use of Likert scales violates many of the assumptions of parametric approaches because of the ordinal level of the data (Nanna & Sawilowsky, 1998). Furthermore, and perhaps more important, the coarse nature of the data limits the amount of information available. The limited information provided by Likert scales can result in a substantial reduction in the ability to detect interaction effects between variables (Russell & Bobko, 1992).

Given the constraints of ordinal data, at least two steps can be taken to alleviate the loss of information. First, the measurement can be conducted in units that approximate uniform or equal units to the greatest extent possible. Second, the number of units can be increased.

Approximating Interval-Level Units. Although ordinal-level data, by definition, comprise unequal data units, there is variation in how unequal the units are. Some ordinal data are very coarse, made up of grossly unequal units, whereas other data may approximate interval data as the measurement units are relatively equal. Consequently, to collect the highest amount of information possible, measurement units should be relatively similar (Russell & Bobko, 1992).

Parks, Parks, and Ogden's (1999) research revealed how dissimilar Likert categories can be. As mentioned earlier, Likert scales have been adapted into other formats using response keys other than the agree–disagree keys. These researchers compiled a number of response terms (for example, always, average, and so forth) commonly used in Likert scales to express probability. Graduate students were then asked to assign a probability score (for example, always = 1.00 or 100 percent, average = .50 or 50 percent) for each of the terms. Table 1, typical of the obtained results, reveals the values assigned to response terms that are often understood by researchers to represent a relatively smooth continuum.

As can be seen, no measurement units are quantitatively similar. Furthermore, a wide range of responses occurred despite the sample being highly educated and having some degree of familiarity with probabilities. The results highlight the fact that responses are specific to the context the individual has in mind at the time the item is answered (Chang, 1994). A value of 90 may be deemed to be below average if respondents envision situations (for example, graduate examinations) in which the range is 80 to 100 with a mean in the low 90s. Conversely, if the item elicits a different picture, a different probability would be assigned.

Terms commonly associated with probabilities in everyday use may be more likely to yield quantitatively similar units than terms that are even more amorphous, such as the widely used agree–disagree format. A sentiment, such as agreement, is more unlikely to be parsed into equal units because of its abstract nature. Feelings of intensity are much more difficult to quantify than probabilities. In addition, the difficulties associated with different dimensions, negatively worded items, and the midpoint option may all function to increase the amount of dissimilarity between measurement units. In sum, Likert response keys seem to be a coarse ordinal-level measure that falls short of approximating interval-level data and consequently fails to tap a significant amount of the available information (Russell & Bobko, 1992).

TABLE 1—Probabilities Associated with Likert-Type Response Key, in Percentages

Statistic	Below Average	Slightly Below Average	Average	Slightly Above Average	Above Average
M	47	47	57	60	68
Median	45	45	50	60	70
SD	18	17	13	17	12
Minimum	10	10	40	10	50
Maximum	90	90	90	90	90

Increasing the Number of Units. Increasing the number of units in the response key often increases the amount of information collected, with a resulting increase in reliability (Chang, 1994). Nunnally (1978) observed that reliability increases with the number of intervals up to 20 steps, but that the increase in reliability is relatively minor after 11 steps. Conversely, reliability drops sharply as the number of response options falls below seven. In short, the closer a response key approximates a continuous measure, the more information is captured (Russell & Bobko, 1992).

Consequently, the information supplied by the commonly used four- and five-point Likert response keys is relatively coarse (Russell & Bobko, 1992). It is, of course, possible to simply increase the number of response categories. Agreement and disagreement could each be subdivided into five categories, which, with the addition of a neutral response option, would allow for a response key with 11 steps.

Response options, however, should correspond to something in the respondents' actual experience for the collected data to have meaning. One of the reasons the standard five-point key is so widely used is that many researchers believe its response options are consistent with individuals' actual experiences. It is highly questionable whether individuals can parse their intensity of agreement into five discrete categories that have substantive meaning (Chang, 1994). Chang's comparison of four- and six-point Likert scales led Chang to suggest that more categories may actually increase the amount of measurement error as respondents may skip response categories that have little meaning for them. In addition, increasing the number of steps with little substantive meaning also may increase the "primacy effect," the tendency for respondents' responses to be the first option acceptable to them (Albanese, Prucha, Barnet, & Gjerde, 1997; Chan, 1991).

Another difficulty with increasing the number of response categories is related to the positively or negatively stated proposition. The response key is linked to the opinion expressed in the proposition. Further dividing the range of intensity does little to increase reliability and validity because the statement itself expresses an opinion and functions to limit the range of responses (Roberts et al., 1999).

More specifically, by necessity statements are commonly designed to express a moderately positive (or negative) opinion posited to indicate the existence of the underlying construct. Individuals who exhibit an extreme degree of the construct have to circumscribe their responses to fit the options presented on the scale. If the opinion expressed in the statement is changed to capture this more extreme element, then more moderate responses are not captured as accurately. Thus, the opinion expressed in the statement limits the ability of Likert scales to tap fully the underlying construct because individuals who hold relatively extreme attitudes must circumscribe their responses to fit the presented options (Roberts et al., 1999; Russell & Bobko, 1992).

In short, the Likert method seems to have built-in limitations that limit the information they can capture. In place of a number of relatively equal units that approximate interval-level data, Likert scales offer a restricted number of relatively unequal units.

Numerical Values and Constants

A final problem with Likert scales is the attribution of numerical values to sentiments in conjunction with an added constant to obtain positive integers. As implied earlier, the attribution of integers to personality traits is based on fiat (Cicourel, 1964). There is no mathematically demonstrable association between agree–disagree and any particular set of integers. Indeed, some qualitative researchers have suggested that attempting to quantify attitudes is essentially an invalid exercise (Franklin & Jordan, 1995; Rodwell, 1987; Scott, 1989). However, given the reality of having to obtain data for statistical analysis, an attempt should be made to be as theoretically grounded as possible in the attribution of values to sentiments.

A normal distribution has both positive and negative numbers. Likert response scales can perhaps best be understood numerically as positive integers (agreement) and negative integers (disagreement). In other words, no opinion or a neutral sentiment is equivalent to zero, and larger positive integers are associated with higher levels of agreement (for example, 2, 1). Conversely, larger negative integers are associated with greater levels of disagreement (for example, –1, –2). Research has demonstrated that respondents may use numbers as a guide to interpretation of the sentiment; consequently, numbers are frequently incorporated into item design (Schwarz, Knauper, Hippler, Noelle-Neumann, & Clark, 1991). Assuming normality for the concept being measured results in a bell-shaped distribution. From a theoretical perspective, there is a curvilinear distribution on each side of zero sloping down to 2 and –2, respectively.

With negative integers, a constant must be added to achieve a positive score for each item. With the

use of only positive integers, one is conceptually restricted to half of the normal curve. Adding a constant does not change the shape of the distribution; however, it does change the meaning of the values underlying the distribution. Researchers can, by fiat, equate −2 with 0 or 1, for example, but the shift from the negative to the positive is lost. Consequently, information is lost, and the analysis is not as accurate.

AN ALTERNATIVE—PHRASE COMPLETIONS

It has been suggested that researchers develop alternative methods for measuring attitudes (Russell & Bobko, 1992). Phrase completions represent an attempt to address the problems inherent in Likert scales.

In place of multiple dimensions, it is clearly optimal that items be designed to concisely assess a single dimension with response scales that approximate a continuous response (Brody & Dietz, 1997; Russell & Bobko, 1992). From a theoretical perspective, most abstract concepts exist along a continuum. Accordingly, there is both theoretical warrant and intuitive appeal for attempting to operationalize this continuum and to specify how it might be measured. For example, by delineating the underlying theoretical continuum on a response key, individuals do not have to circumscribe their responses to the limited number of options contained in Likert scales, thus increasing the accuracy of the instrument (Roberts et al., 1999).

To illustrate the phrase completion method, we adapted the six items Genia (1993) retained after her factor analysis of Allport and Ross's (1967) intrinsic measure of religion, as religion and spirituality are areas in which the profession is experiencing growing interest (Canda & Furman, 1999). The intrinsic measure is among the most influential instruments in the psychology of religion (Hill & Hood, 1999) and also has appeared in the social work literature with some degree of frequency (Gibbs & Achterberg, 1978; Hodge, 2000a, 2000b; Spalding & Metz, 1997).

Allport and Ross (1967) posited the existence of a psychological trait that they referred to as intrinsic religiosity—individuals who find their internal motivation, or in Allport and Ross's terminology their "master motive" for life, in their spiritual tradition. The six items offer the twin advantages of being amenable to transformation but still allowing us to make a comparison to an instrument that has an extensive psychometric history. (For convenience, we have listed the six items in Table 2).

TABLE 2—Original Six Items from the Allport and Ross Intrinsic Scale

(1) I try hard to carry my religion over into all other dealings in life.
(2) Quite often I have been aware of the presence of God or the Divine Being.
(3) Religion is especially important to me because it answers many questions about the meaning of life.
(4) My religious beliefs are what really lie behind my whole approach to life.
(5) I read literature about my faith.
(6) It is important to me to spend periods of time in private religious thought and meditation.

SOURCE: Adapted from Genia, V. (1993). A psychometric evaluation of the Allport-Ross I/E Scales in a religiously heterogeneous sample. *Journal for the Scientific Study of Religion, 32,* 284–290.

From these six items we devised our phrase-completion items (Table 3). Each item consists of an introductory phrase. Building on the work of other researchers (Inglehart & Rabier, 1986), this phrase is followed by an 11-point continuum that serves as the response key. The continuum is anchored on each end with phrase completions that represent the absence of the construct and the maximum amount of the construct. In keeping with research that reveals that respondents use numbers to guide their thinking regarding response keys, individuals are oriented toward the continuum in an introductory statement and asked to circle the number that best reflects their response (Schwarz et al., 1991).

As recommended by Schwartz and associates (1991), the construct is measured by items that emphasize the strength of the attribute. Zero is associated with the absence of the attribute, whereas 10 is associated with a theorized maximum amount; that is, the integers work in concert with the anchor phrases to directly imply the degree of the attribute.

Although other researchers have used similar 11-point response keys in Likert applications (for example, 0 = very dissatisfied, 10 = very satisfied) (Inglehart & Rabier, 1986), this method improves on earlier 11-point response keys by providing a more direct link between the integers and the underlying construct. Furthermore, in place of an opinion conveying a statement that limits response options (Roberts et al., 1999), the full theoretical continuum is delineated in the phrase completion response key.

TABLE 3—Measuring Intrinsic Religion with Phrase Completions

The following questions use a sentence completion format to measure various attributes associated with religion. An incomplete sentence fragment is provided, followed directly below by two phrases that are linked to a scale ranging from 0 to 10. The phrases, which complete the sentence fragment, anchor each end of the scale. The 0 to 10 range provides you with a continuum on which to reply, with zero corresponding to absence or 0 amount of the attribute, whereas 10 corresponds to the maximum amount of the attribute. In other words, the end points represent extreme values, whereas five corresponds to a medium, or moderate, amount of the attribute. Please circle the number along the continuum that seems to best reflect your initial feeling.

(1) My religious beliefs affect

No aspect of my life										Absolutely every aspect of my life
0	1	2	3	4	5	6	7	8	9	10

(2) I am aware of the presence of God or the Divine

Continually										Never
10	9	8	7	6	5	4	3	2	1	0

(3) In terms of the questions I have about life, my religion answers

Absolutely none of my questions										Absolutely all my questions
0	1	2	3	4	5	6	7	8	9	10

(4) My religion is

The master motive of my life, directing every other aspect of my life										Not a factor of my life
10	9	8	7	6	5	4	3	2	1	0

(5) I read literature about my faith

Never										Every day, without fail
0	1	2	3	4	5	6	7	8	9	10

(6) I spend periods of time in private religious thought and meditation

Every day, without fail										Never
10	9	8	7	6	5	4	3	2	1	0

The continuum was reversed with alternating questions to avoid response set bias. This approach allows all items to consist of a single dimension, avoiding the cognitive complexity that occurs with negatively worded items. The end points are highlighted to alert respondents to alternating response keys. This approach is consistent with the work of Barnette (2000), who found that the highest reliability was obtained by using positively worded items and in tandem with alternating response keys (for example, half going from strongly agree to strongly disagree, half going from strongly disagree to strongly agree).

By using a numerical continuum as the response key instead of sentiments that reflect intensity of agreement, respondents may be able to quantify their responses in more equal units. By presenting individuals with an interval-level continuum in the form of integers, researchers offer respondents the opportunity to assess the degree of the attribute and directly indicate their views. Instead of going through the intermediary step of recording sentiments and then transforming the sentiments into integers, individuals are presented with the underlying hypothesized continuum up front. Because the theoretical cards are on the table for respondents to

see and respond to, reliability and validity should be increased.

Whereas the choice of the zero to 10 continuum is arbitrary, rating scales from 0 to 10 are commonly used among the general population (for example, on a scale of 0 to 10, how would you rate the movie?). This familiarity may allow respondents to more accurately quantify their responses. Furthermore, because the response key approximates a continuous measure along a single dimension, more substantively meaningful options are available (Russell & Bobko, 1992). Individuals do not have to circumscribe their responses to fit the limited options in Likert scales (Roberts et al., 1999), and more information is retained (Russell & Bobko), potentially increasing reliability and validity (Chang, 1994; Nunnally, 1978).

Finally, the use of the phrase completion response key better meets the assumptions associated with parametric statistics because it approximates a continuous measure along a single dimension (Brody & Dietz, 1997; Duncan & Stenbeck, 1987; Nanna & Sawilowsky, 1998). Compared with Likert scales, there are more substantively meaningful units that may be more equally spaced. Because the scale taps into a set of ordered categories along a single dimension, items can be summed into an index without violating assumptions (Brody & Dietz).

Psychometric Analysis

On the basis of the preceding material, we expected the phrase completion method to yield higher reliability coefficients and individual items to have stronger factor loadings relative to Likert scales. To test these two hypotheses, the six items listed in Table 3, in conjunction with the orienting material, were administered to 78 students enrolled in a graduate social work program at the start of classes.

The nine-item intrinsic measure, which is normally paired with the Allport and Ross (1967) extrinsic measure, has been used in more than 150 studies with Cronbach's alphas commonly ranging from the low to the mid .80s (Burris, 1999; Trimble, 1997). For example, with a religiously heterogeneous sample (N = 309), Genia (1993) recorded a coefficient of .79 using the traditional scale. Genia was able to increase the reliability to .86 by pairing the six intrinsic items we have transformed in our phrase completions with three reverse-scored extrinsic items.

Although the original nine intrinsic items typically load on a single factor, the loadings have not been particularly strong (Genia, 1993; Kirkpatrick, 1989). Genia, using a principal axis factor analysis with Equamax rotation, reported values ranging from .24 to .80 for the nine items and .45 to .80 for the six retained intrinsic items in her study.

TABLE 4—Interitem Correlations for Six Items on the Allport-Ross Intrinsic Scale

	1	2	3	4	5
1					
2	.733				
3	.887	.716			
4	.930	.759	.922		
5	.691	.624	.679	.748	
6	.795	.715	.742	.819	.774

With our sample, our first hypothesis was confirmed. For the six items, a Cronbach's alpha of .95 was obtained. This reliability coefficient is substantially higher than the coefficient obtained by Genia (1993) with the same items in Likert format. The interitem correlations were relatively high, ranging from .62 to .93 (Table 4).

The six items were subjected to an exploratory principal axis factor analysis. The scree plot indicated a single strong factor. An eigenvalue of 4.86 was obtained, which accounted for 81.06 percent of the variance. All items loaded strongly, confirming our second hypothesis (Table 5). As a point of compari-

TABLE 5—Means, Standard Deviations, and Factor Loadings for Six Items on the Ross-Allport Intrinsic Scale

			Factor Loadings	
Item	*M*	*SD*	Phrase Completions	Likert Format
1	6.56	2.84	.93	.62
2	7.38	2.87	.79	.70
3	5.73	3.01	.91	.57
4	6.13	2.99	.97	.80
5	4.59	3.13	.88	.45
6	5.28	3.31	.87	.70

son, we provide the factor loadings obtained by Genia (1993) with the same six items in Likert format.

Limitations

Although higher reliability and factor loadings were obtained with the phrase completion format, it is important to acknowledge that this initial empirical work is exploratory and limited in nature. For example, although the initial results are encouraging, the results being compared are not drawn from the same sampling frame. Future research, which we are currently conducting, could compare the psychometric properties of both formats using the same sampling frame.

Another limitation associated with the phrase-completion method may be the challenge of operationalizing the underlying continuum in a manner that presents individuals with a relatively simple response key. For example, respondents may need an elevated level of education to understand terms such as "master motive." If respondents have to struggle to understand the terminology used in the response key, the added cognitive noise may increase the amount of measurement error.

CONCLUSION

This article has delineated some of the problems associated with Likert scales, namely multidimensionality and coarse response categories that result in lost information. To circumvent these problems we proposed an alternative approach—phrase completions.

Likert scales require individuals to think across at least two dimensions: content and intensity. Further cognitive complexity is added with reverse-worded items that result in double negatives. When five-point response keys are used, individuals may equate the midpoint option with a not applicable response related to content, whereas the score is recorded as a midlevel intensity response.

Phrase completions reduce cognitive complexity by presenting individuals with a single dimension. Because of the flexibility innate in the phrase completion approach, the double negatives associated with reverse-worded items are averted. Similarly, phrase completions avoid the problem of equating the neutral option with a not applicable response and coding the resulting score as midlevel intensity data by providing individuals with the option of selecting a zero amount of the construct to be assessed.

In addition to the lack of parsimony and the violation of assumptions inherent in multiple dimensions, traditional Likert five-point response keys result in lost information because of the coarse categories. The limited number of substantive response categories means that a limited amount of information is collected. Furthermore, by attempting to tap the level of intensity, significantly dissimilar measurement units are obtained. Finally, a constant must be added to negative integers that also result in lost information.

Phrase completions tap higher levels of information through the use of more refined response keys. By specifying the underlying theoretical continuum in an 11-step response key, more substantively useful information can be collected. Integrating the continuum with a scale ranging from 0 to 10 suggests that the measurement units are more likely to be similar while, concurrently, dispensing with the need to use a constant.

Thus, phrase completions go one step further than Occam's Razor—which stipulates that approaches that accomplish the same ends while reducing cognitive complexity are to be preferred. Phrase completions provide a more parsimonious approach while simultaneously assessing higher levels of information. Accordingly, in keeping with the promising psychometric data reported in this article, phrase completions may offer a better approach to measurement than traditional Likert scales. ■

REFERENCES

Albanese, M., Prucha, C., Barnet, J. H., & Gjerde, C. L. (1997). The effects of right or left placement on the positive response on Likert-type scales used by medical students for rating instruction. *Academic Medicine, 72,* 627–630.

Allport, G. W., & Ross, M. J. (1967). Personal religious orientation and prejudice. *Journal of Personality and Social Psychology, 5,* 432–443.

Babbie, E. (1998). *The practice of social research* (Vol. 8). New York: Wadsworth.

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60,* 361–370.

Brody, C. J., & Dietz, J. (1997). On the dimensionality of two-question format Likert attitude scales. *Social Science Research, 26,* 197–204.

Burris, C. T. (1999). Religious orientation scale. In P. C. Hill & R. W. Hood, Jr. (Eds.), *Measures of religiosity* (pp. 144–154). Birmingham, AL: Religious Education Press.

Byrne, B. M. (1998). *Structural equation modeling with Lisle, Prelis, and Simplis: Basic concepts, applications and programming.* Mahwah, NJ: Lawrence Erlbaum.

Canda, E. R., & Furman, L. D. (1999). *Spiritual diversity in social work practice.* New York: Free Press.

Chan, J. C. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement, 51,* 531–540.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18,* 205–215.

Chung, W. S., & Bagozzi, R. P. (1997). The construct validity of measures of the tripartite conceptualization of punishment attitudes. *Journal of Social Science Research, 22,* 1–25.

Cicourel, A. V. (1964). *Method and measurement in sociology.* London: Free Press of Glencoe.

Cohen, B. Z. (2000). Measuring the willingness to seek help. *Journal of Social Service Research, 26,* 67–82.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6,* 474–494.

Duncan, D., & Stenbeck, M. (1987). Are Likert scales unidimensional? *Social Science Research, 16,* 245–259.

Franklin, C., & Jordan, C. (1995). Qualitative assessment: A methodological review. *Families in Society, 76,* 281–295.

Frans, D. J. (1993). A scale for measuring social worker empowerment. *Research on Social Work Practice, 3,* 312–328.

Garg, R. K. (1996). The influence of positive and negative wording and issue involvement on responses to Likert scales in marketing research. *Journal of the Market Research Society, 38,* 235–256.

Genia, V. (1993). A psychometric evaluation of the Allport-Ross I/E scales in a religiously heterogeneous sample. *Journal for the Scientific Study of Religion, 32,* 284–290.

Gibbs, H. W., & Achterberg, L. J. (1978). Spiritual values and death anxiety: Implications for counseling with terminal cancer patients. *Journal of Counseling Psychology, 25,* 563–569.

Hill, P. C., & Hood, R. W. (Eds.). (1999). *Measures of religiosity.* Birmingham, AL: Religious Education Press.

Hodge, D. R. (2000a). Do faith-based providers respect client autonomy? A comparison of client and staff perceptions in faith-based and secular residential treatment programs. *Social Thought, 19,* 39–57.

Hodge, D. R. (2000b). The spiritually committed: An examination of the staff at faith-based substance abuse providers. *Social Work and Christianity, 27,* 150–167.

Hudson, W. W., & Ricketts, W. A. (1980). A strategy for the measurement of homophobia. *Journal of Homosexuality, 5,* 357–372.

Inglehart, R., & Rabier, J.-R. (1986). Aspirations adapt to situations—But why are the Belgians so much happier than the French? A cross-cultural analysis of the subjective quality of life. In F. M. Andrews (Ed.), *Research on the quality of life* (pp. 1–56). Ann Arbor: University of Michigan.

Julia, M. (1993). Developing an instrument to measure the Spanish-speaking client's perception of social work intervention in prenatal care. *Research on Social Work Practice, 3,* 329–342.

Kirkpatrick, L. A. (1989). A psychometric analysis of the Allport-Ross and Feagin measure of intrinsic–extrinsic religious orientation. *Research in the Social Scientific Study of Religion, 1,* 1–31.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 5–53.

Likert, R. (1970). A technique for the measurement of attitudes. In G. F. Summers (Ed.), *Attitude measurement* (pp. 149–158). Chicago: Rand McNally.

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods, 3,* 55–67.

Neimeyer, R. A., & Bonnelle, K. (1997). The suicide intervention response inventory: A revision and validation. *Death Studies, 21,* 59–81.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Ommundsen, R., & Larsen, K. S. (1998). Attitudes toward illegal aliens: The reliability and validity of a Likert-type scale. *Journal of Social Psychology, 137,* 665–667.

Parks, G., Parks, M., & Ogden, T. (1999, November). *Likert type scales and probabilities.* Unpublished data, George Warren Brown School of Social Work, Washington University, St. Louis.

Raaijmakers, Q.A.W., van Hoof, A., Hart, H. T., Verbogt, T.F.M.A., & Vollebergh, W.A.M. (2000). Adolescents' midpoint response on Likert-type scale items: Neutral or missing values? *International Journal of Public Opinion Research, 12,* 208–216.

Raja, S., & Stokes, J. P. (1998). Assessing attitudes toward lesbians and gay men: The modern homophobia scale. *Journal of Gay, Lesbian, and Bisexual Identity, 3,* 113–134.

Riley, J. L., & Greene, R. R. (1993). Influence of education on self-perceived attitudes about HIV/AIDS among human service providers. *Social Work, 38,* 396–401.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59,* 211–233.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.). (1991). *Measures of personality and social psychological attitudes.* San Diego: Academic Press.

Rodwell, M. K. (1987). Naturalistic inquiry: An alternative model for social work assessment. *Social Service Review, 61,* 231–246.

Russell, C. J., & Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse for comfort. *Journal of Applied Psychology, 77,* 336–342.

Schwarz, N., Knauper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55,* 570–582.

Scott, D. (1989). Meaning construction in social work practice. *Social Service Review, 63,* 39–51.

Shin, H., & Abell, N. (1999). The homesicknesss and contentment scale: Developing a culturally sensitive measure of adjustment for Asians. *Research on Social Work Practice, 9,* 45–60.

Spalding, A. D., & Metz, G. J. (1997). Spirituality and quality of life in alcoholics anonymous. *Alcoholism Treatment Quarterly, 15,* 1–14.

Springer, D. W. (1998). Validation of the adolescent concerns evaluation: Detecting indicators of runaway behavior in adolescents. *Social Work Research, 22,* 241–253.

Trimble, D. E. (1997). The religious orientation scale: Review and meta-analysis of social desirability effects. *Educational and Psychological Measurement, 57,* 970–986.

David R. Hodge, MSW, MCS, is a Rene Sand doctoral fellow, and David

Gillespie, PhD, is professor, George Warren Brown School of Social Work, Washington University, 1 Brookings Drive, Campus Box 1196, St. Louis, MO 63130-4899. Address all correspondence to David R. Hodge.

Original manuscript received September 6, 2000
Final revision received November 20, 2001
Accepted March 20, 2002