The Sequential Approach to Measurement of Health Behavior Constructs: Issues in Selecting and Developing Measures

Colleen A. Redding[1], Jay E. Maddock[2], Joseph S. Rossi[1]

[1]University of Rhode Island
[2]University of Hawaii

Abstract

Measurement theory and practice defines how well we can measure the most important constructs in the health behavior field. This article reviews the sequential process of measurement development that builds upon both theory and evidence, as well as building toward the future of measurement development. Some basic measurement theory and concepts are reviewed, including types of reliability and validity. The process of scale development and selection is described in some detail with clear advise for choosing measures and criteria for selecting items and scales. Finally two different examples of theory-based measurement development are described in detail: one of an alcohol expectancy scale grounded in Social Learning Theory, and the other of scales assessing confidence in remaining quit and temptation to smoke, grounded in the Transtheoretical model conceptualization of self efficacy. These examples illustrate two different ways that measurement development efforts can produce good scales, with different strengths. Finally, some future directions for the field are discussed within the context of health behavior measurement.

© 2006 Californian Journal of Health Promotion. All rights reserved.
Keywords: Measurement; Measurement Development; Psychometrics; Health Behavior

Measurement is the basic foundation of science and research, including the science of health behavior. At a very basic level, measurement defines what we assess and how well. As such, measurement is essential for the intersection between theory, research, and practice for health behaviors. Fundamentally, asking and answering questions about health behaviors and theories of health behavior requires us to be able to measure constructs well. To critically assess how well something is measured requires some understanding of measurement theory, reliability and validity. These are the topics that will be covered in this article.

How does measurement fit into health behavior, health education, and public health? Theory provides the description or "map" of health behavior that tells researchers what to look for, when, and how (Redding, Rossi et al., 1999). Similarly, theory is vital to intervention development, since it describes what variables to intervene on, how to intervene, and what changes to expect. The importance of a good theory in the measurement development process cannot be overstated, since it defines the constructs clearly and specifies their relationships to other theoretical and behavioral variables. A good theory will also clarify the level of specificity or generality necessary to measure a construct well. Unfortunately, there are relatively few guides to theory-based measurement developed specifically for health behavior (e.g., Rossi et al., 1995). While health behavior scientists may debate strengths and weaknesses of different health behavior theories, measurement issues are important regardless. No matter how well a theory appears to explain behavior or behavior change or how well it informs interventions, how it performs empirically is directly related to (and potentially limited by) how well its basic constructs are measured.

Quality measures are essential for all theory-based research, whether the theory you are using is targeted at the individual-level or multiple levels. For any theory, it is important to start with a strong scale that measures a construct well. But how does one assess how well a measure performs? There are several issues to consider when selecting an existing measure or developing a new one. For an existing measure, there are two essential aspects of the scale that reflect its quality: reliability and validity. The important issues to consider when developing a new measure are almost the same as those that are important when evaluating existing ones (Meyer, Edwards, & Rossi, 1995). A measure that is reliable, but not valid, will measure something the same way consistently, but you will never be sure exactly what is being measured. To be useful for research, scales must empirically demonstrate both reliability and validity.

Figure 1 illustrates the sequential multi-step, iterative measurement development process (Comrey, 1988; Spector, 1992). Going through this process once or even only partially can yield an acceptable measure, but it is preferable to go through some or all of the steps more than once. This could be for some new purpose (e.g., generalize to a different population) or simply to refine or improve an existing measure. The process of gathering validation and cross-validation data for any construct or measure is literally ongoing.

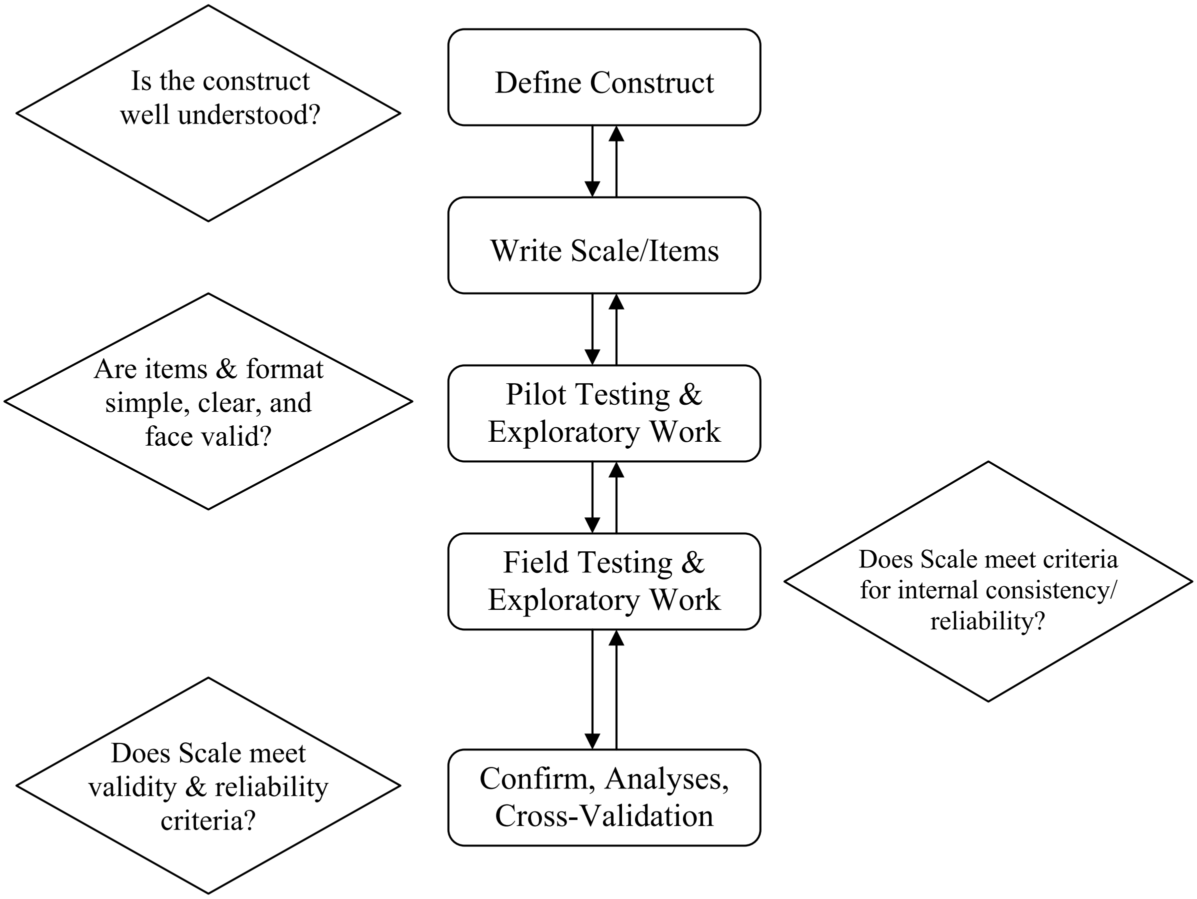
Is the construct well understood?
Define Construct
Write Scale/Items
Are items & format simple, clear, and face valid?
Pilot Testing & Exploratory Work
Field Testing & Exploratory Work
Does Scale meet criteria for internal consistency/ reliability?
Does Scale meet validity & reliability criteria?
Confirm, Analyses, Cross-Validation

Figure 1
Major steps of the sequential measurement development process

Measurement Theory

Every time a participant answers an item, the observed score may be thought of as consisting of both a “true” score and some random error. If one could remove all measurement error, that is, use a perfectly reliable measure, then the observed score would equal the true score. Of course, the more error (or noise) included in any measure, the more difficult it will be to find the true score (or signal) and the worse the reliability will be (Spector, 1992). Since errors are assumed to be random and normally distributed, the more items you include, the more their associated errors cancel each other out, leaving a better scale measuring something closer to the true score with less error, i.e., a more reliable scale. Just as summing across many items reduces error in a measure, using many respondents minimizes error in a sample as well. This is some of the theory underpinning the measurement development process. The goal is to select good manifest questions or items across reasonably large samples of people to measure a construct.

Response Bias

Of course, not all error is random. Sources of measurement error that are systematically related to participants’ attitudes or traits are called response biases. Biases are important since they influence observed scores systematically and yet, do not reflect the true score. Luckily, there are ways to measure and/or control most response biases. The main type of response bias is social desirability or a tendency for participants to portray themselves as more socially acceptable or “normative” than is the case. Many behavioral scientists measure social desirability as part of the scale validation process to demonstrate that social desirability does not “contaminate” their measure. This is a good idea, whenever possible. Also, by including a social desirability scale in the exploratory testing phase, one can evaluate individual item relationships to social desirability and choose items that are less influenced by it. Another response bias, acquiescence represents the tendency for people to agree with any item, regardless of its content. One can control for participants’ tendency to agree by simply balancing positively worded (e.g., I love to exercise.) and negatively worded (e.g., I hate to sweat.) items.

Reliability

Reliability is the proportion of the observed score that is due to the true score. Therefore, it also provides an estimate of the error assessed by a scale. Increasing the reliability of a scale improves the statistical power of a study because better reliability minimizes error in much the same way as a larger sample size reduces error (Lipsey, 1990). Reliability is important because it is the foundation for and literally defines the limits of any evaluation of validity. Table 1 provides a brief description of many different kinds of reliability estimates, and most of the terminology used in this article.

Table 1
Definitions of Types of Reliability

Type	Definition
Alternate-forms reliability	The correlation between two different versions of a measure.
Internal Consistency	The degree of homogeneity of scale items.
Inter-rater reliability	The correlation between two raters.
Split-half reliability	The correlation between two halves of a measure.
Temporal stability	The correlation between two administrations of a measure separated by some time period (usually days or weeks), that reflects all sources of error in addition to measurement error accompanying a dynamic measure over time.
Test-retest reliability	The correlation between two administrations of a measure separated by some time period (usually days or weeks).

Internal consistency is the most common type of reliability; it means that items in the scale hold together or intercorrelate highly. Cronbach's (1951) coefficient Alpha (α) is a commonly used measure of internal consistency. Alpha represents the true score or signal portion of variance in a measure. The average correlation between the items and the number of items in a scale jointly determine Alpha. Like a correlation coefficient, Alpha can range from 0 to 1, with smaller numbers reflecting less consistency and higher numbers reflecting more consistency. For most research purposes, an Alpha of at least .80 is recommended, although values as low as .70 can be acceptable (Clark & Watson, 1995; DeVellis, 1991; Nunnally & Bernstein, 1994). When Alpha is above .90, the scale may be too long and can be shortened without much loss of efficiency. Alternatively, a very high Alpha can indicate that the construct has been too narrowly defined and that construct breadth has been sacrificed for internal consistency.

There are several additional ways to assess reliability. All involve calculation of the correlation between at least two: different versions of a measure (alternate forms reliability); halves of a measure (split-half reliability); administrations of a measure separated by time (temporal stability or test-retest reliability); or sets of ratings (inter-rater reliability). When the construct of interest does not change over time (e.g., traits, personality, gender), then any changes in measurement over time can be attributed to error and test-retest reliability is appropriate to assess. However, most constructs of interest in the behavioral sciences are at least somewhat dynamic. So, this assumption of stability over time is most often false. When a construct is expected to change over time, then other types of reliability (aside from test-retest) that do not confound change over time and error of measurement provide clearer assessments. For many of the same reasons, the term temporal stability is recommended for correlations over time instead of test-retest reliability, since the former reflects the many potential sources of change over time in addition to measurement error (DeVellis, 1991).

Validity

How do you know your scale is measuring what you want it to measure? This is the question of validity. This is the most important and the most challenging aspect of measurement development. This step involves measuring the construct of interest along with other well-defined and understood constructs to see if the relationships among old and new constructs are as strong and in the same direction as one would expect, based on both theory and previous data. Constructs are accepted and validated because they appear to be useful within a theoretical and empirical context. Validity is inferred from the strength of the data supporting it, through converging lines of evidence; validity can never really be proven. Since it is so inherently related to theory, the interpretation of existing validity data can change when support for a theory broadens or the theory changes in some way to account for some new phenomena. Furthermore, unlike reliability, validity is not an inherent aspect of a measure — it depends critically on how the measure is used. A measure shown to be valid for one purpose is not necessarily valid for any other purpose.

There are three main kinds of validity relating to measurement development: content validity, criterion-related validity, and construct validity. Table 2 lists each type of validity with a short definition.

Table 2
Definitions of Types of Validity

Type	Definition
Content validity	The extent to which a measure fully assesses the breadth and depth of the intended construct.
Construct validity	The pattern of consistency with which a measure relates to other theoretically relevant variables and constructs.
Concurrent validity	The extent to which a scale relates meaningfully and significantly to similar measures collected simultaneously.
Convergent validity	The extent to which a scale relates meaningfully and significantly to similar measures.
Criterion-related validity	The extent to which a scale demonstrates meaningful and important relationships to an accepted criterion.
Divergent or Discriminant validity	The extent to which a scale demonstrates minimal or nonimportant relationships to unrelated constructs or variables.
Face validity	Clarity and relevance of the item or scale to the measured construct.
Known-Groups validity	The extent to which a scale differentiates meaningfully between groups demonstrated or known to be different.
Predictive validity	The extent to which a scale improves prediction of an accepted criterion variable.

Content validity reflects the fact that the scale has the appropriate breadth and nature of the items included in it. A scale has demonstrated content validity when either the content domain is very clear (e.g., sixth grade science words) or content experts agree that it contains the range of items necessary to represent the construct domain well. Another related term, face validity, refers to how clearly an item is phrased to reflect the intended construct. While face validity is not technically required from a psychometric standpoint, it is nevertheless highly desirable, since perceived relevance does enhance honest and accurate responding.

Criterion-related validity refers to the empirical association between the construct of interest and one or more "gold-standard" measures in the field. Criterion-related validity is often referred to as predictive validity, although this does not imply a causal relationship between variables. Predictive validity usually refers to the situation in which measurement of the criterion follows that of the construct of interest. However, measurement of the criterion can also be simultaneous (concurrent validity). What is important for criterion-related validity is how strongly related the construct of interest is to an important criterion. The term criterion-related validity is less often used compared to predictive and concurrent validity in the area of health and behavioral medicine. However, the term criterion-related validity is preferable since it is more neutral with respect to the issues of both time and causation (DeVellis, 1991).

A third main type of validity is construct validity (Cronbach & Meehl, 1955), the relationship between the construct of interest and other theoretically related constructs. For example, we would expect that two different measures of depression would show a substantial correlation since they measure the same construct. Such strong correlation between measures of related constructs is known as convergent validity. However, as Campbell and Fiske (1959) pointed out, it is also very important to show that measures of theoretically distinct constructs are not correlated (e.g., depression and intelligence). The lack of a substantive correlation between constructs that are expected to be different is known as discriminant (or divergent) validity.

The classic multitrait-multimethod matrix developed by Campbell and Fiske (1959) is an especially compelling procedure for demonstrating both convergent and discriminant validity.

Another less often used but strong method of assessing construct or criterion-related validity is known-groups validity. This procedure involves demonstrating that the construct of interest can meaningfully differentiate between groups that are expected to differ on the construct. For example, according to the Transtheoretical model, the confidence and temptations constructs are expected to vary across the stages of change in a specific and predictable fashion (Prochaska et al., 1991; 2002). Evidence for known-groups validity is obtained by demonstrating a significant relationship between these constructs and stages of change using, for example, an analysis of variance using stages as the independent variable and confidence and temptations as dependent variables. The magnitude of the relationship can be assessed using a measure of effect size, such as η^2. Rather than using defined or existing groups, known-groups validity can also be assessed through experimental manipulation. For example, in the development of a measure of self-efficacy for weight control, obese diabetic patients randomly assigned to an intensive weight management program showed increases in self-efficacy from baseline to post-treatment (Clark, Abrams, Niaura, Eaton, & Rossi, 1991). An even stronger approach would be to demonstrate that the sample assigned to the control condition did not show comparable increases in self-efficacy.

Criterion-related and construct validity are often confused since the same data can support both types of validity. The difference between them lies in both the intent of the investigator and the theoretical foundation for the hypotheses. When the theoretical foundation of a construct is strong, then construct validity is the appropriate term. When the study is focused more on empirical evidence, then criterion-related validity is a more appropriate term.

The Measurement Development Process

Construct Definition

As Figure 1 illustrates, the first step in the measurement development process is defining exactly what one intends to measure, operationalization. At this step, a good review and understanding of the previous theoretical and empirical literature is invaluable. What theories are important and how do they differ in their predictions or assumptions about how to measure the construct of interest? What alternative measures have already been developed and how clear, concise, valid, reliable, and internally consistent are they? How is this construct similar to and different from other constructs important within your perspective on health behavior? Finally, how specific or general is this construct? Being able to answer these questions about your construct means that you may be able to proceed to the next step in the process, that is, developing items to measure your construct.

The decision to go with an existing measure versus developing a new measure can be challenging. Often a measure exists which is close to the desired construct but suboptimal in some way. Conversely, developing a new measure is costly and time consuming. Once a literature review is conducted, several factors should be considered (Meyer et al., 1995). 1) Does a scale exist measuring the construct of interest? 2) What is the evidence for the reliability and validity of the existing scale? 3) What populations has this scale been tested on? 4) How many items are in the scale? 5) When was the scale developed? 6) Are administration and scoring directions clear? These are important questions to ask and answer when considering the development of a new scale. A new scale should only be developed when no other acceptable alternatives exist. If a partially acceptable scale does exist, several options can be considered. These include: using the scale as is; re-testing the scale in a new population, setting, or format; updating and testing the scale using more current or clear language or additional items to better assess the content domain; or, if the scale is too long, developing a short form of the scale (although the development of short forms is more complicated than commonly thought; see Smith, McCarthy, & Anderson, 2000). Whatever the ultimate decision, these methods provide a guide to maximizing the reliability and validity of the instrument.

Developing a Scale

Developing a scale involves writing items that reflect the construct, choosing a response format, and writing instructions for participants. All the writing in scales should be clear, concise, easy to understand, and written at an appropriate reading level for the target population. A reading level between fifth and seventh grade is probably appropriate for most general population surveys (DeVellis, 1991). Some redundancy is a good thing when generating pools of items. Good interitem correlations are the basis for internal consistency. The initial item pool should have at least double the number of items anticipated in the desired scale length. Since factor analytic methods usually serve as the basis for scale development, three items is the effective minimum scale length (Velicer & Fava, 1987, 1998).

Writing Items

Good items have face validity; they reflect the construct of interest clearly. An item should contain only one simple, straightforward idea. If several ideas are written into one item (a.k.a. item complexity), respondents cannot respond clearly. Complex items cannot perform well statistically. Jargon should be avoided as much as possible. Balancing positively worded and negatively worded items can be a good way to control for acquiescent response bias. Scoring, then, must be reversed on either the positively or negatively worded items, so that all items can be meaningfully summed into one scale score. Balancing positively and negatively worded scales can, however, create confusion, particularly among participants with low reading ability and/or in lengthy surveys as fatigue accumulates. Try to avoid using negatives such as “no” or “not” to balance the wording of items (e.g., use “I hate beans.” instead of “I do not like beans.”), since they are too easily missed as respondents read quickly. Finally, using grammar correctly avoids the common problems of misplaced modifiers and ambiguous pronouns. In short, good items are simple, clear and direct, while maintaining face validity.

Response Formats

There are several different response formats to choose from, the most common alternatives being frequency (How often?), evaluation (How much do you like it?), and agreement (How much do you agree?). Some items ask questions in a binary Yes/No or True/False format. In general, response formats that include more choices are better than those with only two choices, because they can explain more variance. If you must use a forced choice or binary format, then you’ll have to use correspondingly more items in order to explain comparable amounts of variance. Scales having 4 to 7 response options allow for a good deal of variability while keeping the differences in responses meaningful.

Expert Review, Pilot Testing, and Formative Research

Expert review, pilot testing, and formative research (focus groups, cognitive interviews, etc.) are conducted at this point. Expert reviews will provide some estimation of the content validity, clarity, conciseness, and face validity of individual items. If your scale has subscales or dimensions, content experts should be able to categorize items into appropriate subscale categories. Pilot testing and formative research can refine abstract ideas, ensure the content domain is fully understood, and find out how the target population talks about and/or thinks about the construct. It is also important during this phase to rewrite and correct any poorly worded or misleading items. There are many good texts on formative research available and readers are encouraged to pursue more detail elsewhere (Denzin & Lincoln, 2000; Maxwell, 1996; Miles & Huberman, 1994). Quina and colleagues (1999) provide a good applied description of the role of formative research in the measurement development process for HIV prevention research with at-risk women.

Field Testing and Exploratory Analyses

After human subjects review board approval, the questionnaire can be administered to respondents. Samples of convenience (e.g., college students, volunteers) can be fine for field testing item pools and exploratory analyses on a new measure, especially when they are very similar to the population being studied. However, since convenience samples often differ in important ways from other target groups (especially diverse or clinical samples), cross-validation with a sample more representative of the population of interest becomes even more important.

How many respondents will you need for your exploratory sample? In the past some have based sample size decisions on subject to variable ratios; however, this turns out to be a less important consideration than initially thought. Simulation studies have demonstrated that the size of item loadings (.60-.80) have a much larger effect on sample size needed than subject to variable ratios (Guadagnoli & Velicer, 1988; Marsh, Balla, & McDonald, 1988; Velicer & Fava, 1998). If lower item loadings (.40-.60) are expected, then many more participants may be required to find stable factor solutions. In practice, between 200-300 respondents will likely be sufficient for most measurement development situations (Clark & Watson, 1995; DeVellis, 1991). Practically, however, with only one scale extracted from about 20 items with good loadings, fewer participants (N = 150) would probably suffice. These sample size estimates are based on complete data.

Data Collection

How the data are collected can greatly influence data quality, so care should be taken to minimize demand characteristics of the survey administration and/or environment. Self-administered surveys are often used for measurement development, in part because they can be filled out quickly by many participants. Anonymous surveys often effectively reduce response bias, especially for sensitive content areas, such as sexual behavior or drug use. Generally, a research assistant's supervision during the survey administration will help to answer any respondents' questions, as well as to oversee the optimally quiet, nonreactive setting in which people fill out the survey. Self-administered surveys are limited by non-response bias and comprehension difficulties. Items may be left blank for many reasons including misunderstandings, non-applicable items and refusals. Response bias may also be present if surveys are mailed to respondents. Low return rates can greatly affect the representativeness of the final sample. In-person and telephone surveys can increase response rates and assist in clarifying items directly. However, these methods are more expensive and time-consuming than self-administered surveys. Careful consideration of data collection methods is necessary before developing any instrument.

Item Analysis

Once the data are cleaned and ready, an initial item analysis should be conducted. Since only a few items will be kept on the scale, choose those items that will discriminate maximally between respondents. First, the number of respondents to each item and the item means and standard deviations should be assessed. Items with very high or very low means, and items with very little variance should be deleted because they cannot discriminate among respondents. Similarly, items with lots of missing data should also be discarded. After these poor items are deleted, the corrected item-scale correlation for each item should be assessed. This tells us the correlation of the item to the rest of the scale, excluding itself.

The remaining items should then be entered into a factor or components analysis. Factor analysis or principal components analysis will assist the researcher in determining how many factors or dimensions are measured by the scale items. Rules for factor/component rotation and factor/component selection are important, complex and beyond the scope of this article. Several good articles and books address these concerns in more detail (Gorsuch, 1983; Velicer, Eaton, & Fava, 2000). Recommended methods for determining the number of factors or components to retain include the scree test (Cattell, 1966), parallel analysis (Horn, 1965; Lautenschlager, 1989), and the MAP criterion (Zwick & Velicer, 1986). The most common and simple rule for numbers of factors to retain, the Kaiser rule (i.e., the number of eigenvalues exceeding one), is not recommended, since it has been demonstrated to retain too many factors (Zwick & Velicer, 1986). When a measurement structure has a strong theoretical foundation, structural equation modeling can also be used in an exploratory manner to do measurement development; however, this can be challenging (for more on this, see Noar, 2003). After the dimensional analysis is clear, items that have low loadings (< .40) on any factor or have loadings > .40 on two or more factors (complexity) should be removed. The internal consistency of each factor can then be measured using coefficient Alpha. The optimal scale retains a broad measure of the construct without overburdening participants. Interested readers can consult DeVellis (1991) for more in-depth discussion.

Cross-Validation and Confirmatory Analyses

In research one study alone cannot conclusively demonstrate any phenomenon, since the results may have occurred by chance alone. Therefore, replication is required for any phenomenon to be accepted scientifically. This is also true for measurement development. Cross-validation, or administering the measure in another independent sample, provides much more certainty in the psychometric structure of a measure. Multiple samples reduce error across studies in much the same way multiple items do across a scale.

Cross-validation involves administering the same scale to another sample in order to conduct confirmatory analyses. Usually an independently collected sample provides the best cross-validation, however, some have attempted to cross-validate measures by splitting a very large sample in half or less often, by readministering the scale to the same sample at a new timepoint. Once the data are collected, one could run another (unrestricted) factor or components analysis comparable to that done in the exploratory phase. A slightly different approach would be to do a restricted factor analysis on the data, where items are constrained to load on the factor that they were intended for and loaded on before. This approach uses structural equation modeling (SEM), also known as confirmatory factor analysis (CFA) (Noar, 2003). SEM has many capabilities beyond measurement development that will not be reviewed here (Bollen & Long, 1993; Loehlin, 1992).

One useful technique is called split-half cross-validation. To use this procedure, a sample size is collected twice as large as that needed for exploratory analyses alone. Then, the sample is randomly split in half, and one half used for exploratory measurement analyses, while the other half is used later for confirmatory measurement analyses. Studies using this procedure can draw much more firm conclusions about the reliability and validity of their measures than studies relying on an exploratory dataset and analyses only. Large sample sizes (N = 400–600) supporting split-half cross-validation are not always feasible, however, in which case, exploratory analyses in a smaller sample (N = 200–300) work quite well.

Validation

After the measure has been demonstrated internally consistent and reliable, criterion-related and/or construct validity needs to be explored. This can be done creatively, but should include evidence for both convergent and divergent validity. If there is another well-accepted measure of a closely related construct in your field, it is a good idea to include it along with your scale in exploratory data collection. Then, analyses of variance and correlation coefficients can be used to assess its criterion-related validity. Regression analyses or longitudinal structural models can also provide good evidence of construct or criterion-related validity. Researchers should be cautioned, however, not to rely too heavily on the term ‘predictive validity,’ since, an analysis can be no more predictive than the data upon which it is based. So, for example, cross-sectional prediction analyses support much more tentative conclusions than longitudinal prediction analyses. Evidence for validity arises less from the method of analysis than from the strength of the research it is based on.

Summary

Even after all of these steps have been applied, good measures can be continually improved through the collection of additional validating evidence or through revision. Testing scales in new and more representative samples is a good idea before wide dissemination occurs. Structural equation modeling can be used to assess whether an instrument has the same psychometric structure for different genders, ethnic groups, or age ranges, an important procedure that is becoming more commonly known as structural or factorial invariance modeling. Continued careful testing of the instrument is the foundation for conducting high quality theoretical research.

Example #1: Development of an Alcohol Expectancy Measure

Construct Definition

In social cognitive theory, outcome expectancies play a central role in predicting health behavior. Outcome expectancies are defined as the value that a person places on a particular outcome (Baranowski, Perry & Parcel, 2002). In the field of alcohol use, the belief in the reinforcing effects of drinking have been shown to be related to alcohol consumption (Burden & Maisto, 2000). The relationship between expectancies and consumption has been conceptualized as the final common pathway in decisions about alcohol use (Cox & Klinger, 1990).

A comprehensive measure of alcohol expectancies should assess both positive and negative consequences of drinking (Fromme, Stroot & Kaplan, 1993). Also, the subjective evaluations of the outcome expectancies need to be measured. As Fromme and colleagues have shown, even negative outcomes such as irresponsibility and decreased motor control are often seen as desirable outcomes by college aged drinkers (Fromme, Marlatt, Baer & Kivlahan, 1994). Over the past 2 decades several inventories have been developed to measure alcohol expectancies (Leigh & Stacy, 1991). Very few of these measures examined the independent influence of outcome expectancies and subjective evaluations (Fromme et al., 1993). As an example, we will evaluate the methods used by Fromme and colleagues (1993) to develop a comprehensive effects of alcohol (CEOA) questionnaire.

Initial Instrument Development

To generate items for the initial instrument, 103 items were used from several existing expectancy questionnaires. In addition, the authors developed 36 new items to measure the biphasic and negative effects of drinking. All items were rated on face validity for inclusion. All of the 139 items began with the stem, “If I were under the influence from drinking alcohol…” (e.g. I would be friendly). Participants responded on a 4-point Likert scale (1 = disagree, 4 = agree). Subject evaluation of item face validity was rated on a 5-point Likert scale (1 = bad, 3 = neutral, 5 = good).

The method for the generation of items is strong. The use of existing items is a positive approach as long as the items have face validity. It appears that no expert panel of judges was used to rate the items. Use of expert raters is a recommended technique when possible. The large number of items is appropriate for the scale, allowing for elimination of low loading items. The use of four to five point Likert scales is recommended.

Field Testing and Exploratory Analyses

The instrument battery was completed by 344 participants recruited from college psychology courses. The sample was largely White, 57% female, and averaged 20 years old. Alcohol consumption varied with 14% abstainers, 14% light drinkers, 24% moderate drinkers, and 48% heavy drinkers. Participants completed a self-administered questionnaire containing the CEOA, a daily drinking questionnaire, and a demographic inventory.

Exploratory analysis of the inventory began with univariate statistics. Thirty-seven items that were not endorsed (M < 2.0) were removed. The remaining 102 items were then entered into an exploratory factor analysis, which failed to converge. The authors hypothesized that this was due to the low subject-to-variable ratio, so they divided the item pool into 46 positive (mean subjective rating > 3.0) and 56 negative (mean subjective rating < 3.0) expectancies. The authors also deleted 7 items with very low (< .20) corrected item total correlations. Principal components analyses (PCAs) were then conducted separately for the positive and negative expectancies, using scree plots and eigenvalues > 1.0 to decide on how many factors to retain. Five positive and four negative factors were found. Items were deleted if their largest factor loading was < .40 or if their deletion increased the Alpha for that factor (removing 17 items). PCAs were again conducted with items constrained to load on their previous factor. Items that loaded on more than one factor were removed. This produced a final exploratory factor structure with 4 positive factors (Sociability, Tension Reduction, Liquid Courage, & Sexuality) containing 22 items (55.9% of the positive variance) and 3 negative factors (Cognitive and Behavioral Impairment, Risk and Aggression, & Self-Perception) containing 19 items (46.3% of the negative variance). Gender differences in factor structure were then examined by conducting separate factor analyses by gender. Adequate coefficients of congruence were found for all factors except for tension reduction, indicating comparability of the factor structures across gender.

While college students are overused in psychological studies, they are an especially appropriate participant pool for studies of alcohol consumption. The sample size is adequate for a measurement development study. The sample is fairly homogeneous, which is advantageous in exploratory analyses. The preliminary item analysis removed several items with low endorsement. Both the reliance on the subject-to-variable ratio and the decision to run separate PCA’s for negative and positive expectancies were problematic. This methodology removes the possibility of having negative and positive items loading on the same factor, so it would have been preferable to run a PCA first. The use of eigenvalues > 1.0 rule for selecting the number of components often retains more factors than are necessary (Zwick & Velicer, 1986). The second round of PCAs might have reduced some of these problems if conducted on the entire measure. Cronbach Alpha’s for the scales should always be reported. Establishing factor structure comparability across gender is best accomplished through the use of structural invariance analysis rather than through the assessment of factor congruence of independent PCAs.

Confirmatory Model Testing

The confirmatory analyses were based on the responses of 485 college participants. This group was slightly younger (M = 19) and more likely to be female (66%). Drinking was also more common in this sample with 34% moderate drinkers and 48% heavy drinkers.

To test the factor structure developed in the exploratory phase, confirmatory factor analysis (CFA) was conducted using the original sample. As before, the positive and negative structures were tested separately. Both models needed small modifications to adequately fit the data, leading to the deletion of 2 positive items and 1 negative item. The confirmatory models were then tested in the second sample. Both the positive and negative models provided good fit to the data. Next the criterion-related validity of the scales was assessed across three measures of alcohol consumption. Regression analyses indicated significant relationships between positive and negative expectancies and alcohol consumption. In addition to the criterion-validity, temporal stability and dose-related expectancies were also examined. These analyses showed good temporal stability as well as the ability to differentiate between the number of drinks needed to experience positive effects (M = 4.4) compared to negative effects (M = 5.8).

The use of confirmatory factor analysis is an important step in measurement development. It is essential in judging the relationship of the individual factors to each other. Alternative plausible models were not tested. Also, the positive and negative models were not tested together. This would be recommended, at least to determine the pattern of correlations between the positive and negative factors. The regression showed a predictive relationship of the overall scales to alcohol consumption. Tests for differential effects of these scales on sub-groups of drinkers were not performed. Finally, again, Cronbach's Alpha's were not reported.

Future Directions

This article represents a good measurement development effort. While it was limited in some ways, it is stronger than many published articles. The use of split-half cross-validation, measurement of criterion-related validity, temporal stability and dose-response are all recommended. However, as is the case in any research, several studies could provide important additional evidence supporting this scale. First, internal consistencies (Alpha's) of the overall scales and subscales should be reported. Second, the scales could be tested to examine convergent and discriminant validity. An alcohol related problem index such as the College Alcohol Problems Scale could be used to assess the expectancies of different sub-groups of drinkers (Maddock, Laforge, Rossi & O'Hare, 2001). Third, higher order factors of the scales could be examined. How are the positive and negative scales interrelated? Fourth, the predictive validity of the scales could be evaluated. Do these scales predict alcohol consumption and problems prospectively? Finally, these scales could be examined in diverse populations using structural invariance modeling. Currently, the scale has only been tested on relatively homogeneous samples of White college students. Are the factor structures and scales valid for African-Americans, Asians, Hispanics, adolescents, and adult drinkers? These research suggestions would greatly strengthen the reliability and validity evidence supporting the utility of this promising scale.

B. Application #2 – The Self Efficacy Construct based on the Transtheoretical Model

Self efficacy is the most widely accepted and supported construct across various alternative theories of health behavior change (Bandura, 1977; 1986; Rossi & Redding, 2001; Strecher et al., 1986). The Transtheoretical model includes many variables in addition to self-efficacy and has a long history that has been well reviewed (Prochaska, Redding & Evers, 2002). The development and refinement of the self-efficacy construct within the Transtheoretical model was chosen as an illustrative example for this article, because the program of research is broad and demonstrates exceptional strength of both theory and evidence for reliability and validity. Finally, this measurement development has partially contributed to the development of effective, disseminable interventions for smoking cessation among adults (Prochaska, Velicer, DiClemente, & Rossi, 1993; Prochaska et al., 2001a, 2001b; 2004; 2005; Velicer et al., 1993, 1999). Comparable interventions are now being developed and evaluated among adolescents and across different risk behaviors (Hollis et al., 2005; Redding et al., 1999).

Initial measurement development of self-efficacy for smoking cessation was conducted on smokers and ex-smokers by DiClemente and colleagues (1981, 1982, 1985, 1986). The original scale was composed of 12 items (DiClemente, 1981), which was expanded into 31 items for both confidence in ability to avoid smoking across situations (self-efficacy) and temptation to smoke across situations (cue strength). These scales were first tested among 957 respondents at baseline and subsequently among 813 respondents at follow-up 3-5 months later (DiClemente, Prochaska, & Gibertini, 1985). Exploration of dimensionality in the item set using PCA's revealed "identifiable but not clearly interpretable subcomponents" so scales were analyzed as total scales, with very high internal consistencies (α = .97 and α = .98 for 31-item temptation and confidence scales respectively). Construct validity was examined using correlations of the self efficacy scale at baseline with: demographic variables; the endurance subscale of the Jackson Personality Inventory (Jackson, 1967); and life experiences survey (Sarason, Johnson, & Siegel, 1978) (divergent validity); and decisional balance, processes of change and other change-related constructs (convergent validity) among both smokers (n = 440) and ex-smokers (n = 300). Criterion-related (known-groups) validity was examined longitudinally by comparing self-efficacy scores for participants in each stage of change at baseline who did not progress over time to those who had progressed or regressed at follow-up (Prochaska et al., 1985). This series of studies established the internal consistency reliability, construct validity, and predictive utility of this self-efficacy scale.

The remaining 31-item set was refined, reanalyzed, and organized hierarchically using SEM by Velicer and colleagues (1990) using the same dataset as before. Importantly, Velicer and colleagues identified an extremity response bias in the data, which they eliminated by deleting any case that answered with either all 1's or all 5's on the scale items. As they point out, such a bias in the data would artificially create a general factor and could mask actual dimensions in the scale. PCA's were again conducted, and using the MAP rule (Velicer, 1976; Zwick & Velicer, 1986), three comparable components were identified for both scales, retaining 18 of the original 31 items. The three types of situations found to be important for smoking cessation were named: negative affective; habit/addictive strength; and positive social. An hierarchical SEM with each situation represented by one subscale of 5-7 items was found to describe the data best. This hierarchical structure allowed for scoring the scale at either the lower-order subscale level or at the higher-order construct level. More evidence for divergent validity was added by demonstrating a small relationship to social desirability in a subsample (n = 100). Finally, an independent sample of 421 smokers was collected, and several new items were added. Results confirmed the hierarchical three-factor model and resulted in a 17-item version of both scales. Coefficient Alpha's and one-month test-retest reliabilities (actually temporal stability) were reported for all three subscales.

Additional longitudinal predictive validity studies have been conducted using cross-sequential methods (Fitzgerald & Prochaska, 1990; Prochaska et al, 1991) and structural modeling techniques (Velicer, Rossi et al., 1996).

Cross-Validation Across Samples

Temptations was assessed in both a random digit dial representative sample of smokers (Fava et al., 1995) and among pregnant women (Ruggiero et al, 2000). The hierarchical model with many of the same items for Temptations to smoke was then re-tested among adolescent smokers and nonsmokers (Plummer et al., 2001), with the addition of one situational subscale for smoking (weight concerns) and three new subscales for nonsmokers (curiosity, peer pressure, and weight concerns). The relationship of temptations to smoke across stages of change was compared between adult and adolescent smokers (Pallonen, 1999) and was cross-validated in a sample of Bulgarian adolescent smokers and nonsmokers (Anatchkova et al., 2006, in press).

Validation of the Measurement Model Across Behaviors

Finally this measurement model has been adapted for other areas of behavior change, such as weight control (Clark et al., 1991), safer sex (Redding & Rossi, 1999), dietary fat reduction (Ounpuu, Wolcott, & S. Rossi, 1999; S. Rossi et al., 2001), exercise (Rossi, Benisovich et al., 2006), and binge drinking (Maddock et al., 2000). Of course, each content area has unique, content-specific items. Importantly, comparisons of these different efficacy scales reveal some subscales unique to the problem behavior (i.e., habit strength, partner pressure, situational cues), while other subscales appear to generalize across problem behaviors (i.e., negative affect, positive social). This understanding would not be possible if each problem behavior used measurement models that were not comparable. Most of these papers provide known-groups validation for the measures as well, by examining confidence and/or temptation as a function of stages of change.

Taken together, this substantial program of research provides excellent evidence for the reliability, temporal stability, and construct validity for the smoking confidence and temptation measures across an impressive range of smokers, as well as supporting the utility of this measurement model across new behavioral areas.

Future Directions for the Measurement of Health Behavior Constructs

These examples demonstrated well the ongoing process of measurement development and the accumulation of converging validity evidence so vital to the science of health behavior. The first example did an adequate job with the measurement development aspect of their scale and a very good job with their validity evaluation. The second example starts out with typical scale development procedures and through continued empirical and methodological work, produced a set of self-efficacy and temptation scales with broad evidence for validity, reliability, and psychometric stability that is impressive in both range and scope. There are many excellent measurement applications from our field. Until recently most measurement, like most theories of health behavior, has been focused at the individual level. However, theorists are moving beyond the individual level to include larger units of analysis, such as families, communities, worksites, and schools. The measurement of group-level constructs is more challenging in several ways, yet will hopefully prove to be enlightening for our field. Similarly, the development of so many excellent community-based programs as part of current public health initiatives (e.g., HIV prevention) is promising and measurement issues are obviously vital to program evaluation efforts.

Our population is fast changing, with recent census figures attesting to our increasing diversity and multiculturalism. Our measurement models must increasingly be tested within samples quite different from those in which measures were developed, potentially challenging our measurement models and psychometric skills. Calls to disseminate our intervention models and measures more widely place additional pressures on the science of health behavior to continue to do basic measurement work supporting our measures. Structural invariance of constructs across racial/ethnic groups, urban and suburban samples, socioeconomic groups, groups speaking different languages, and different educational levels is now a timely research agenda. Some excellent additional ideas for multicultural research can be found in Hishinuma et al. (2000), Matsumoto (1994), and Okazaki and Sue (1995).

The scientific value of using comparable measures across studies cannot be overstated. Health behavior research has been limited by a proliferation of non-comparable measures examining similar theoretical constructs. Such practices impede the accumulation of convergent evidence necessary for scientific progress and synergy. Some published articles fail to report entire scales or basic statistics making it difficult for other researchers to use them, to replicate results, or to conduct meta-analyses. Researchers often conduct measurement studies using their preferred theoretical constructs, yet fail to apply the same measurement standards to their outcome measures (Velicer et al, 1992). Substandard outcome measures hamper our ability to conduct meaningful outcomes research (not to mention increasing our required sample sizes and budgets).

Along these same lines, health behavior research programs could be greatly enhanced by the creation of a nationwide database of available measures organized within and across health behaviors. Such an effort might be a natural project for a crosscutting federal agency, such as the Office of Behavioral and Social Science Research. Such a database could be web-based, free to all researchers, contain all scale items and related reliability and validity information, and have links to all published articles using that scale. The development of such a database would greatly improve the comparability of studies done from Rhode Island to Hawaii. It would also provide an invaluable tool for developing and accelerating meaningful research programs across the science of health behavior.

Calls for wide dissemination of models and measures have suggested using commercial channels to partially fund this effort. However, commercial research settings may not report proprietary items and reliability/validity data that would be necessary for scientific evaluation. Maintaining the scientific integrity of our work by reporting complete psychometric work is a fundamental scientific standard that cannot be compromised. Commercial research should meet the same standards for being evidence-based and maintaining scientific integrity as research funded from other sources in order to be taken seriously. Establishing standards that would allow more clear evaluation of commercial research would allow federally funded researchers and commercially funded researchers to work together more closely, which could accelerate the science of health behavior and the dissemination of our best interventions.

Finally, we hope we have illustrated how important measurement is to the science of health behavior. We reiterate that scientific progress across fields of inquiry is linked to both validity and precision of measurement. In short, increasing our ability to detect the signal over and above the noise is fundamentally an issue of measurement. At the most basic level, if you can't measure it in a reliable and valid way, then you can't study it. Excellent measurement work must not only be the foundation for our science, but for our interventions as well so that our efforts to help individuals change their health behaviors can be truly evidence-based.

References

Anatchkova, M. D., Redding, C. A., & Rossi, J. S. (2006). Development and validation of decisional balance and temptations measures for Bulgarian adolescent smokers. Addictive Behaviors, 31(1), 155-161.

Anatchkova, M. D., Redding, C. A., & Rossi, J. S. (in press). Development and validation of transtheoretical model measures for Bulgarian adolescent nonsmokers. Substance Abuse and Misuse.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavior change. Psychological Review, 84, 191–215.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice Hall, Inc.

Baranowski, T., Perry C., & Parcel, G. (2002). How individuals, environments and health behavior interact: Social Cognitive Theory. In K. Glanz, F. M. Lewis & B. K. Rimer (Eds.), Health behavior and health education (3rd ed.). San Francisco, CA: Jossey-Bass.

Bollen, K. A., & Long, J. S. (Eds.) (1993). Testing structural equation models. Thousand Oaks, CA: SAGE Publications, Inc.

Burden, J. L., & Maisto S. A. (2000). Expectancies, evaluation and attitudes: Prediction of college student drinking behavior. Journal of Studies on Alcohol, 61, 323-331.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81–105.

Cattell, R. B. (1966). The Scree test for the number of factors. Multivariate Behavioral Research, 1, 245–276.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7, 309–319.

Clark, M. M., Abrams, D. B., Niaura, R. S., Eaton, C. A., & Rossi, J. S. (1991). Self-efficacy in weight management. Journal of Consulting and Clinical Psychology, 59, 739–744.

Comrey, A. L. (1988). Factor analytic methods of scale development in personality and clinical psychology. Journal of Consulting and Clinical Psychology, 56, 754–761.

Cox, W. M., & Klinger E. (1990). Incentive motivation, affective change, and alcohol use: A model. In W. M. Cox (Ed.), Why people drink (pp. 291-314). New York: Gardner Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281–302.

Denzin, N. K., & Lincoln, Y. S. (2000). Handbook of qualitative research (2ed.). Thousand Oaks, CA: SAGE Publications, Inc

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

DiClemente, C. C. (1981). Self-efficacy and smoking cessation maintenance: A preliminary report. Cognitive Therapy and Research, 5, 175–187.

DiClemente, C. C. (1986). Self-efficacy and the addictive behaviors. Journal of Social and Clinical Psychology, 4, 302–315.

DiClemente, C. C., Prochaska, J. O., Fairhurst, S. K., Velicer, W. F., Velasquez, M. M., & Rossi, J. S. (1991). The process of smoking cessation: An analysis of precontemplation, contemplation and preparation stages of change. Journal of Consulting and Clinical Psychology, 59, 295–304.

DiClemente, C. C., Prochaska, J. O., & Gibertini, M. (1985). Self-efficacy and the stages of self-change of smoking. Cognitive Therapy and Research, 9, 181–200.

Fava, J. L., Velicer, W. F., & Prochaska, J. O. (1995). Applying the transtheoretical model to a representative sample of smokers. Addictive Behaviors, 20, 189–203.

Fitzgerald, T. E., & Prochaska, J. O. (1990). Nonprogressing profiles in smoking cessation: What keeps people refractory to self-change? Journal of Substance Abuse, 2, 87–105.

Fromme, K. Marlatt, G. A., Baer, J. S. Kivlahan, D. R. (1994). Alcohol skills training program: A group intervention for young adult drinkers. Journal of Substance Abuse Treatment, 11, 143-154.

Fromme, K., Stroot, E., & Kaplan, D. (1993). Comprehensive effects of alcohol: Development and psychometric assessment of a new expectancy questionnaire. Psychological Assessment, 5, 19–26.

Gorsuch, R. L. (1983). Factor analysis (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Assoc., Inc.

Guadagnoli, E. & Velicer, W.F. (1988). The relationship of sample size to the stability of component patterns. Psychological Bulletin, 103, 265-275.

Hishinuma, E. S., Andrade, N. N., Johnson, R. C., McArdle, J. J., Miyamoto, R. H., Nahulu, L. B., Makini, G. K., Yuen, N. Y. C., Nishimura, S. T., McDermott, J. F., Waldron, J. A., Luke, K. N., & Yates, A. (2000). Psychometric properties of the Hawaiian culture scale — Adolescent version. Psychological Assessment, 12, 140–157.

Hollis, J. F., Polen, M. R., Whitlock, E. P., Lichtenstein, E., Mullooly, J., Velicer, W. F., & Redding, C. A. (2005). Teen REACH: Outcomes from a randomized controlled trial of a tobacco reduction program for teens seen in primary medical care. Pediatrics, 115(4), 981-989.

Horn, J. I. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179–185.

Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberger (Ed.), Current topics in clinical and community psychology, Vol. 2. (pp. 61–96). New York: Academic Press.

Kelly, J. R., & McGrath, J. E. (1988). On time and method. Beverly Hills, CA: SAGE Publications, Inc.

Lautenschlager, G. J. (1989). A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. Multivariate Behavioral Research, 24, 265–395.

Leigh, B. C., & Stacy, A.W. (1991). On the scope of alcohol expectancy research: Remaining issues of measurement and meaning. Psychological Bulletin, 110, 147-154.

Lipsey, M. W. (1990). Design sensitivity: Statistical power for experimental results. Newbury Park, CA: Sage.

Loehlin, J. C. (1992). Latent variable models: An introduction to factor, path, and structural analysis (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Maddock, J. E., Laforge, R. G., & Rossi, J. S. (2000). Short form of a situational temptation scale for heavy, episodic drinking. Journal of Substance Abuse, 11, 281–288.

Maddock, J. E., Laforge R., Rossi, J. S., & O'Hare T. (2001). The college alcohol problems scale. Addictive Behaviors, 26, 385-398.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391–410.

Matsumoto, D. (1994). Cultural influences on research methods and statistics. Pacific Grove, CA: Brooks/Cole.

Maxwell, J. A. (1996). Qualitative research design: An interactive approach. Thousand Oaks, CA: Sage Publications, Inc.

Meyer, E. C., Edwards, G. H., & Rossi, J. S. (1995). Evaluation and selection of standardized psychological instruments for research and clinical practice. Journal of Child and Adolescent Psychiatric and Mental Health Nursing, 8(3), 24–31.

Miles, M. B., & Huberman, A. M. (1994). Qualitative data analysis: An expanded sourcebook (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Noar, S. M. (2003). The role of structural equation modeling in scale development. Structural Equation Modeling: A Multidisciplinary Journal, 10(4), 622-647.

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

Okazaki, S., & Sue, S. (1995). Methodological issues in assessment research with ethnic minorities. Psychological Assessment, 7, 367375.

Ounpuu, S., Woolcott, D. M., & Rossi, S. R. (1999). Self-efficacy as an intermediate outcome variable in the transtheoretical model: Validation of a measurement model for applications to dietary fat reduction. Journal of Nutrition Education, 31, 16–22.

Pallonen, U. E. (1998). Transtheoretical measures for adolescents and adult smokers: Similarities and differences. Preventive Medicine, 27, A29–A38.

Plummer, B. A., Velicer, W. F., Redding, C. A., Prochaska, J. O., Rossi, J. S., Pallonen, U. E., & Meier, K. S. (2001). Stage of change, decisional balance, and temptations for smoking: Measurement and validation in a large, school-based population of adolescents. Addictive Behaviors, 26, 551-571.

Prochaska, J. O., Crimi, P., Lapsanski, D., Martel, L., & Reid, P. (1982). Self-change processes, self-efficacy and self-concept in relapse and maintenance of cessation of smoking. Psychological Reports, 51, 983990.

Prochaska, J. O., DiClemente, C. C., Velicer, W. F., Ginpil, S., & Norcross, J. C. (1985). Predicting change in smoking status for self-changers. Addictive Behaviors, 10, 395–406.

Prochaska, J. O., DiClemente, C. C., Velicer, W. F., & Rossi, J. S. (1993). Standardized, individualized, interactive and personalized self-help programs for smoking cessation. Health Psychology, 12, 399-405.

Prochaska, J. O., Redding, C. A., & Evers, K. (2002). The transtheoretical model and stages of change. Chapter 5 in K. Glanz, B. K. Rimer, & F. M. Lewis (Eds.), Health behavior and health education: Theory, research, and practice (3rd ed.)(pp. 99-120). San Francisco, CA: Jossey-Bass, Inc.

Prochaska, J. O., Velicer, W. F., Guadagnoli, E., Rossi, J. S., & DiClemente, C. C. (1991). Patterns of change: Dynamic typology applied to smoking cessation. Multivariate Behavioral Research, 26, 83–107.

Prochaska, J. O., Velicer, W. F., Fava, J. L., Rossi, J. S., & Tsoh, J. Y. (2001a). Evaluating a population-based recruitment approach and a stage-based expert system intervention for smoking. Addictive Behaviors, 26, 583-602.

Prochaska, J. O., Velicer, W. F., Fava, J. L., Ruggiero, L., Laforge, R. G., Rossi, J. S., Johnson, SS., & Lee, PA. (2001b). Counselor and stimulus control enhancements of a stage-matched expert system intervention for smokers in a managed care setting. Preventive Medicine, 32, 23-32.

Prochaska, J. O., Velicer, W. F., Redding, C. A., Rossi, J. S., Goldstein, M., DePue, J., Greene, G. W., Rossi, S. R., Sun, X., Fava, J. L., Laforge, R., Rakowski, W., & Plummer, B. A. (2005). Stage-based expert systems to guide a population of primary care patients to quit smoking, eat healthier, prevent skin cancer and receive regular mammograms. Preventive Medicine, 41, 406-416.

Prochaska, J. O., Velicer W. F., Rossi, J. S., Redding, C. A., Greene, G. W., Rossi, S. R., Sun, X., Fava, J. L., Laforge, R., & Plummer, B. A. (2004). Impact of simultaneous stage-matched expert system interventions for smoking, high fat diet and sun exposure in a population of parents. Health Psychology, 23(5), 503-516.

Quina, K., Rose, J. S., Harlow, L. L., Morokoff, P. J., Deiter, P. J., Whitmire, L. E., Lang, M. A., & Schnoll, R. A. (1999). Feminist process model for survey modification. Psychology of Women Quarterly, 23, 459–483.

Redding, C. A., Prochaska, J. O., Pallonen, U. E., Rossi, J. S., Velicer, W. F., Rossi, S. R., Greene, G. W., Meier, K. S., Evers, K. E., Plummer, B. A., & Maddock, J. E. (1999). Transtheoretical individualized multimedia expert systems targeting adolescents' health behaviors. Cognitive & Behavioral Practice, 6(2), 144-153.

Redding, C. A., & Rossi, J. S. (1999). Testing a model of situational self-efficacy for safer sex among college students: Stage of change and gender-based differences. Psychology and Health, 14, 467–486.

Redding, C. A., Rossi, J. S., Rossi, S. R., Velicer, W. F., & Prochaska, J. O. (1999). Health behavior models. In G. C. Hyner, K. W. Peterson, J. W. Travis, J. E. Dewey, J. J. Foerster, & E. M. Framer (Eds.), SPM handbook of health assessment tools (pp. 83–93). Pittsburgh, PA: The Society of Prospective Medicine & The Institute for Health and Productivity Management.

Rossi, J. S., Benisovich, S. V., Norman, G. J., & Nigg, C. R. (2006). Development of a hierarchical multidimensional measure of exercise self-efficacy. Manuscript in review.

Rossi, J. S., & Redding, C. A. (2001) Structure and function of self-efficacy across the stages of change for 10 health behaviors. Annals of Behavioral Medicine, 23, S094 (Abstract).

Rossi, J. S., Rossi, S. R., Velicer, W. F., & Prochaska, J. O. (1995). Motivational readiness to control weight. In D. B. Allison (Ed.), Handbook of assessment methods for eating behaviors and weight-related problems: Measures, theory, and research (pp. 387–430). Thousand Oaks, CA: Sage Publications, Inc.

Rossi, S. R., Greene, G. W., Rossi, J. S., Plummer, B. A., Benisovich, S. V., Keller, S., Velicer, W. F., Redding, C. A., Prochaska, J. O., Pallonen, U. E., & Meier, K. S. (2001). Validation of decisional balance and temptation measures for dietary fat reduction in a large school-based population of adolescents. Eating Behaviors, 2, 1-18.

Ruggiero, L., Tsoh, J. Y., Everett, K., Fava, J. L., & Guise, B. J. (2000). The transtheoretical model of smoking: Comparison of pregnant and nonpregnant smokers. Addictive Behaviors, 25, 239–251.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. Psychological Assessment, 12, 102–111.

Spector, P. E. (1992). Summated rating scale construction: An introduction. Thousand Oaks, CA: SAGE Publications, Inc.

Strecher, V. J., DeVellis, B. M., Becker, M. H., & Rosenstock, I. M. (1986). The role of self-efficacy in achieving health behavior change. Health Education Quarterly, 13, 73-91.

Velicer, W. F., DiClemente, C. C., Rossi, J. S., & Prochaska, J. O. (1990). Relapse situations and self efficacy: An integrative model. Addictive Behaviors, 15, 271–283.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), Problems and solutions in human assessment: Honoring Douglas Jackson at seventy (pp. 41–71). Boston: Kluwer.

Velicer, W. F., & Fava, J. L. (1987). An evaluation of the effects of variable sampling on component, image, and factor analysis. Multivariate Behavioral Research, 22, 193-209.

Velicer, W. F. & Fava, J. L. (1998). The effects of variable and subject sampling on factor pattern recovery. Psychological Methods, 3, 231-251.

Velicer, W. F., Prochaska, J. O., Bellis, J. M., DiClemente, C. C., Rossi, J. S., Fava, J. L., & Steiger, J. H. (1993). An expert system intervention for smoking cessation. Addictive Behaviors, 18, 269–290.

Velicer, W. F., Prochaska, J. O., Fava, J. L., Laforge, R. G., & Rossi, J. S. (1999). Interactive versus noninteractive interventions and dose-response relationships for stage-matched smoking cessation programs in a managed care setting. Health Psychology, 18, 21–28.

Velicer, W. F., Prochaska, J. O., Rossi, J. S., & Snow, M. G. (1992). Assessing outcome in smoking cessation studies. Psychological Bulletin, 111, 23–41.

Wilcox, N. S., Prochaska, J. O., Velicer, W. F., & DiClemente, C. C. (1985). Subject characteristics as predictors of self-change in smoking. Addictive Behaviors, 10, 407–412.

Zwick, W. R., & Velicer, W. F. (1986). A comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99, 432-442.

Acknowledgements
Authors wish to acknowledge the support of NIH grants #CA 50037, # CA 63745, # CA 85999 and #AR 43051 which made writing this paper possible, as well as the collegial intellectual environment at the Cancer Prevention Research Center, University of Rhode Island.

Author Information
Colleen A. Redding, Ph.D.*
Cancer Prevention Research Center
2 Chafee Road
University of Rhode Island
Kingston, RI 02881
Ph.: 401-874-4316
Fax.: 401-874-5562
E-Mail: credding@uri.edu

Jay E. Maddock, Ph.D.
Department of Public Health Sciences
University of Hawaii at Manoa
1960 East-West Rd.
Honolulu, HI 96822
Ph.: 808-956-5779
Fax.: 808-956-6041
E-Mail: maddock@hawaii.edu

Joseph S. Rossi, Ph.D.
Cancer Prevention Research Center
2 Chafee Road
University of Rhode Island
Kingston, RI 02881
Ph.: 401-874-5983
Fax.: 401-874-5562
E-Mail: jsrossi@uri.edu

* corresponding author