The BEHAVIORAL MEASUREMENT *Letter*

Behavioral Measurement Database Services

Enriching the health and behavioral sciences by broadening instrument access

Vol. 6, No. 2
Fall 1999

Introduction to This Issue

I begin my introduction to this issue of *The Behavioral Measurement Letter* with an apology to both its readers and contributors. This issue is very much behind schedule due mostly to me, more specifically, a series of illnesses I suffered and the resultant accumulating backlog of various items of business, The BML included. I want to thank all of you for your patience and understanding, and especially our regular contributor, Fred Bryant, for his expert assistance in editing this issue.

One of the most interesting yet least investigated topics at the intersection of health care and social science is the relationship of spirituality to health and illness. In this issue, Bruce Frey and Timothy Daaleman discuss their current work to define and measure the construct spirituality. Their review of prior work found that existing studies of spirituality in health have two fundamental flaws: (a) these studies mostly examine religiosity rather than spirituality, or they otherwise confound these two different constructs, and/or (b) the studies examine spirituality from the perspective of the health care provider or researcher rather than that of the patient/subject in whom the spirituality of interest resides and from whom such is expressed. Given the lack of valid studies and measures of spirituality, Drs. Frey and Daaleman have initiated qualitative studies to define the construct of spirituality so that it then may be operationalized.

Understanding a culture, one's own or a foreign culture, requires knowledge of how persons in that culture view their world. This, in turn, requires tools to obtain such knowledge. John Gatewood, a cognitive anthropologist, discusses three measurement methods for measuring similarities, differences, and relationships among products of a culture or items in its environment -- free listing, pile sorting, and triadic comparison. Using examples from our own culture and environment, he shows how these tools may be used to begin to gain an understanding of a culture. One drawback of using these tools, however, is that data analysis can be cumbersome and time-consuming. Thus, at the end of his column, Dr. Gatewood provides an Internet website address of a manufacturer of data analysis software to be used in processing data obtained by these methods.

Once again, Fred Bryant writes about measurement modeling. In this installment of a continuing series on the topic, he discusses the use of measurement modeling to determine if what is measured by the Affect Intensity Measure (AIM) is one construct or more than one construct, that is, if the AIM taps one factor or multiple factors. The work Dr. Bryant reports indicates that the AIM actually measures three factors, and that the construct "affect intensity," as measured by the AIM, therefore consists of three factors. As in preceding columns in the series, this piece clearly demonstrates how measurement modeling is useful in defining the "structure" of a measurement instrument and thereby contributes to our understanding of instruments and of constructs as operationalized by instruments.

Address comments and suggestions to The Editor, *The Behavioral Measurement Letter*, Behavioral Measurement Database Services, PO

Introduction *(continued from page 1)*

Box 110287, Pittsburgh, PA 15232-0787. If warranted and as space permits, your communication may appear as a letter to the editor. Whether published or not, your feedback will be attended to and appreciated.

We also accept short manuscripts for The BML. Submit, at any time, a brief article, opinion piece or book review on a BML-relevant topic to The Editor at the above address. Each submission will be given careful consideration for possible publication.

HaPI reading . . .

Al K. DeRoy, Editor

> The claims of habit are too weak to be felt until they are too strong to be broken.
>
> Samuel Johnson

Toward a Patient-Centered Measure of Spirituality

Bruce B. Frey and Timothy P. Daaleman

Spirituality has become an accepted, even fashionable part of the contemporary American social scene. It should come as little surprise, then, that spirituality has an increasingly larger presence in health care, as evidenced by recent discussions of this ill-defined construct in both clinical and research settings (Levin, Lavson, & Puchalski, 1997). However, development of operational and clinically useful definitions, a classification scheme, and methods to assess spirituality in health care contexts remain in their infancy (Daaleman, 1999).

Most research on relationships of spirituality to health and well-being has concentrated on "religiosity" or religion-based constructs, rather than the broader construct, "spirituality." Within this limited context of "religiosity," researchers from the fields of psychology, sociology and, to a lesser extent, theology and pastoral care have produced some measures and templates for instrument development. The relatively few studies of *spirituality* in health and illness have been compromised by the lack of precision and clarity in operationally defining "spirituality" (Sloan, Bagiella, & Powell, 1999).

In response to this need for a clear, precise definition of spirituality in health care, two recent national conferences -- one sponsored jointly by the National Institute on Aging (NIA) and the Fetzer Institute (Ory & Lipman, 1998), the other by the National Institute for Healthcare Research (NIHR) (Larson, Swyers, & McCullough, 1998) -- focused on challenges in defining and measuring religiousness and spirituality. Products of these conferences have provided investigators some direction and guidelines for researching spirituality within health care settings.

Although the fields of theology and pastoral care have long examined spirituality and religion, there has been little empiric testing or validation of instruments used in these disciplines. Nonetheless, these faith-based sources can provide ample materials and resources for developing instruments to measure spirituality as well as religiosity. One such resource is *Measures of Religiosity*, a text edited by Peter Hill and Ralph Hood, to be published by Religious Education Press.

In choosing or developing an appropriate scale, researchers must first examine their overall project goals and objectives to determine which spiritual dimension (i.e. practice or belief system) is most applicable. For example, studies of chronic illness would benefit from scales that emphasize spiritual coping or support (practice), while projects that focus on health-related behaviors would benefit from a measure of belief systems. In any case, there are, at present, no parsimonious yet comprehensive scales that are applicable in all research settings. What follows is a brief overview highlighting some of the pitfalls and challenges in producing an operational definition of spirituality, a necessary step in developing an instrument to measure spirituality in health and illness.

Our review of previous work revealed the primacy of investigator-based constructs of spirituality and the virtual exclusion of the patient's perspective. However, much of the current interest in the spiritual aspects of health care stems from two highly patient-centered movements within medicine: end-of-life care and alternative medicine. Given these movements toward patient-centered and relationship-centered care, and given the reasonable assumption that the agent for any spirituality-related outcome substantially resides in the patient, we chose to begin with a patient-centered approach. Use of a patient-centered approach will permit us to define the presently ill-defined construct of health-related spirituality from the perspective of patients.

Measurement theory and practice provide quantitative methods for illuminating constructs and exploring validity arguments linked to hypothesized definitions of constructs. By its nature, the process requires that scores be obtained for some group on a set of items or tasks, and then the scores examined, correlated or compared in order to shed light on the nature or components of the construct of interest. The validity of those scores is a grand summary of all existing evidence for meaningful interpretation of the scores and the potential consequences of those interpretations (Messick, 1989). Strictly speaking, it refers to the meaningfulness of scores, not the value of the constructs employed in developing the items used to obtain the scores. But clearly, the meaningfulness of item responses is dependent on the extent to which the dimension to be measured has been appropriately well-defined (Foddy, 1993).

In order to develop quantitative measures, it is necessary to create operational definitions of the constructs to be measured, a task that is best begun using qualitative methods. Qualitative tools, such as directed interviews, open-ended questionnaires and focus groups, can generate both ideas and understanding. Also, in developing instruments, we are mindful that the definition of a construct is meaningful to the extent that it authentically captures *subjects'* understanding of the construct. Thus, we have begun to develop patient-centered measures of spirituality and its relationships to health and well-being using qualitative techniques.

One qualitative research tool we have chosen to use is focus groups. Methodological descriptions of focus groups are beyond the scope of this discussion and are presented elsewhere (for example, Maxwell, 1996; Strauss & Corbin, 1998). The operational definition of "focus group" that we employ is "a small group of people with specific, targeted characteristics who are asked a series of directed questions in a forum designed to promote a contextual understanding." Focus groups are a type of purposeful sampling where members are selected because of a specific attribute. The researcher does not, at least at this stage of the instrument development process, wish to generalize to a larger population, but through an analysis of the transcribed focus group sessions, seeks to identify and understand the themes generated by the groups in response to the directed questions. Size considerations are not based on issues of statistical power, but rather on saturation which seeks a total sample size that provides differing perspectives until no new themes emerge from the sessions.

The focus groups that we formed were recruited within a university medical center. In order to define spirituality in the contexts of both illness and wellness, one set consisted of patients with an identified medical condition (type 2 diabetes mellitus), a second, independent set of focus groups consisted of persons with no self-identified acute or chronic illness. Some core questions that we used to frame the discussion include, "When you hear the word 'well-being,' what comes to mind? How would you define it?" Then, to explore the link between subjective well-being and spirituality, the following was posed to the groups: "There has been a lot of talk about spirituality recently. What is spirituality? What is the connection between spirituality and your understanding of well-being?"

Transcripts of the sessions are then coded for analysis, the goal being to reflect accurately the output or "common voice" of the focus group. We have been using the coding methods suggested by Strauss and Corbin (1998) in order to generate categories, dimensions and theoretical relationships. In the coding process, the transcripts are reviewed sentence by sentence, phrase by phrase, sometimes even word by word, and the thoughts expressed are categorized (i.e., coded). Once compiled, the list of categories -- often quite long -- is then reanalyzed and edited to combine, refine, and reduce the number of categories, often with input from collaborators or expert reviewers who provide additional interpretative perspectives. The process is complete when consensus is reached among the investigators that further analysis will add very little or nothing of value. Points of theoretical connection among categories, or nodes, are then identified, expanded or merged, and preliminary models may emerge.

Once we produce a patient-centered conceptual definition of spirituality via the methods described above, we will move into the second stage, i.e., creating items for a quantitative scale to measure health-related spirituality. Next the items will be piloted and then revised with reference to theory as well as psychometrics.

If spirituality is an aspect of human nature that affects well-being, it is vital that the definition one chooses for it reflects the views of those affected. To date, however, measures of spirituality within health care settings have been developed from the investigator's perspective exclusively, yet its effect on patients is what is of interest. Measures generated from a patient-centered model, as we have outlined, could authentically capture spirituality in health and illness.

References

Daaleman, T.P. (1999). Science of the spirit. How researchers quantify the value of belief. *The Park Ridge Center Bulletin, 7,* 13-14.

Ferris, L.E., & Tudiver, F. (1992). Creating an original measure. In Stewart, M.A. et al. (Eds). *Tools for Primary Care Research.* Newbury Park, CA: Sage.

Foddy, W. (1993). *Constructing questions for interviews and questionnaires.* Cambridge, UK: Cambridge University Press.

Larson, D.B., Swyers, J.P., & McCullough, M.E. (Eds). (1998). *Scientific research on spirituality and health: a consensus report.* National Institute for Healthcare Research, Rockville, MD.

Levin, J.S., Larson, D.B., & Puchalski, C.M. (1997). Religion and spirituality in medicine: Research and education. *Journal of the American Medical Association,* 278, 792-793.

Maxwell, J.A. (1996). *Qualitative research design: An interpretive approach.* Thousand Oaks, CA: Sage.

Messick, S. (1989). Validity. In R.L. Linn (Ed.). *Educational measurement.* New York: American Council on Education.

Ory, M., & Lipman, P.D. (1998) *Religion and spirituality in aging and health.* National Institute on Aging, Behavioral and Social Research Program, Bethesda, MD.

Sloan, R.P., Bagiella, E., & Powell T. (1999). Religion, spirituality, and medicine. Lancet, 353, 664-667.

Strauss, A., & Corbin, A. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory.* (2nd ed.) Thousand Oaks, CA: Sage.

Bruce Frey is an Assistant Professor at the University of Kansas, School of Education, Department of Psychology and Research in Education, where he teaches courses in measurement, research design, and statistics. His research interests include applied measurement, survey construction, and program evaluation. His work in the past has included risk and resiliency factors in predicting drug abuse, and physician's attitudes towards religion.

Timothy P. Daaleman, DO, is an Assistant Professor in the Departments of Family Medicine and the History and Philosophy of Medicine at the University of Kansas Medical Center School of Medicine. He also holds an appointment as Assistant Scientist at the University of Kansas Center on Health in Aging, and is a Robert Wood Johnson Foundation Generalist Faculty Scholar.

Staff

Director . . . . Evelyn Perloff, PhD
Document Delivery
Coordinator . . . . Johanna Kehler, D Phil
Newsletter Editor . . . . Al K. DeRoy, PhD
Computer Consultant . . . . Alfred A. Cecchetti, MS
Measurement Consultant . . . . . Fred B. Bryant, PhD
Information Specialist
Consultant . . . . Ellen Detlefson, PhD
Record Analysis Manager . . . . . Daniel Moore, MA
Record Reviewer/Indexer . . . . . Jacqueline A. Reynolds, BA
File Processor . . . . Betty Hubbard
Business Services Manager . . . . Diane Cadwell

Phone: (412) 687-6850 Fax: (412) 687-5213
E-mail: bmdshapi@aol.com

Culture . . . One Step at a Time

John B. Gatewood

There is a cross-disciplinary joke in fisheries management circles. A commercial fishery is in crisis, and its managers commission research to determine the causes and possible solutions. The *biologist* evaluates a handful of variables pertaining to fish stock growth and mortality, and writes a 3-page report concluding with a recommended policy action. The *economist* considers a dozen or so variables concerning the costs and benefits of the fishery based on alternative future scenarios, and submits a 15-page report with an either/or recommendation. The *anthropologist* considers the full multitude of factors contributing to the fishery's problem, and hands in a 200-page report with no recommendation.

Contrary to such folk humor about the idiosyncratic nature of anthropological research, in this column I review some ways in which anthropologists can and do use explicit, replicable methods to make headway understanding the cultural part of social life. I will focus on a few of the rather specialized data collection techniques being used in cognitive anthropology -- specifically, free-listing, pile-sorts, and triadic comparisons. Each of these topics is well presented in several recent anthropological methods books (Bernard, 1994, 1998; Borgatti, 1996; D'Andrade, 1995; Weller & Romney, 1988).

Free-Listing

One of the initial problems facing anthropologists interested in studying aspects of culture, or cultural domains, is how to ask questions in ways that are meaningful to natives, or people living within that culture. In particular, researchers need to phrase questions using natives' own cognitive categories (Frake, 1962, 1964). But what are these categories, what are the elements of the cultural domain, and how can a non-native discover them? The free-listing task is a good way to explore native vocabularies, and it provides interesting information about research informants.

A free-listing task is virtually the same as a free-recall task in psychology, except that the period of learning is not controlled by the researcher but rather consists of the previous (and variable) life experiences of the informants. Informants are simply asked to name (if non-literate) or write (if literate) all the items they can think of that match a given category. Examples include:

"Please make a list of all the contagious diseases you can think of."

"Please name all the parts of a human body you can think of."

"Please write down all the phases of the human life cycle you can think of."

Conceptually, such tasks are quite simple and are generally well understood by the informant, they do not impose preconceived response categories, and they work well with both non-literate and literate informants. Furthermore, although the ideal is to work with informants one at a time, the same free-listing task can be given to multiple literate informants at the same time or even as a communal task for focus groups. Still, there are a few choices and problems researchers should be aware of beforehand.

First, because the usual free-listing task (such as in the instructions above) is virtually unconstrained, it is not always clear when informants are finished. Typically, informants generate the first several items rather quickly, but then slow down dramatically. Depending on the cultural domain in question, informants' knowledge of the domain, and their motivation, the task can take from a few seconds to ten minutes before they run out of steam. Related to this problem of having no clear endpoint, informants seldom enjoy free-listing tasks. The instructions sound easy, and often informants are initially eager. But once they begin and realize the full open-endedness of the task, many begin to feel apologetic for the brevity of their lists or resentful at being made to encounter their own memory limitations.

There are two effective ways to bring closure to the free-listing task, each of which makes the task less onerous and more standardized across different informants. First, the instructions can include an explicit time-limit, e.g., "Please make a list of all the kinds of diseases you can think of *in the next 3 minutes*." Alternatively, the task can specify a maximum number of items to be identified, e.g., "Can you remember any sponsors of last year's Super Bowl? If yes, please name up to three of them." Of course, the specific limitation chosen should be guided by a pilot test of the unconstrained version of the task to observe when informants bog down, by the research objectives, and by the size of the sample of informants. Specifying a maximum number of items to list works well enough, if the domain itself is the primary research objective and the task is given to hundreds of informants. On the other hand, I prefer the time-limit approach when working with small numbers of informants (e.g., 40 or fewer), and especially if I am interested in informant-level attributes. When working with US college students, 90 seconds seems to be a reasonable limit for many domains, such as kinds of kin, trees, mammals, fish, mixed drinks, hand tools, fabrics, or musical instruments. Less indoctrinated informants might prefer a somewhat longer time-limit.

The second main problem with free-listing arises after collecting the data and beginning the analysis. The aggregated findings of free-listing tasks are displayed as a table in which rows contain the name of an item followed by the number of lists in which the item appeared, and usually the table items are sorted by decreasing frequency of mention. Many cultural domains, however, are organized taxonomically (e.g., a red oak is a kind of oak, and an oak is a kind of deciduous tree; a Neon is a kind of Dodge, and a Dodge is a kind of car; etc.), and hierarchical relations among items in a domain are not captured by a listing task. Also, some items may be synonymous with one another. These problems can make it hard to determine how many *different* items appear in the sample's lists. Suppose, for instance, that you asked four informants (A--D) to list 'five kinds of American cars' and obtained the lists in Table 1. How many different kinds of American cars appear in the four lists?

Table 1.
Four Free-Lists of Five Kinds of American Cars

List A	List B	List C	List D
Ford	Ford Taurus	Taurus	minivan
Chevy	Chevrolet Corvette	Explorer	sport utility vehicle
Dodge	Jeep Cherokee	Corvette	sedan
Cadillac	Jeep Wrangler	Neon	convertible
Jeep	Dodge Neon	Seville	station wagon

Because we are very familiar with the cultural domain of American automobiles, we can see patterns in these responses. Apparently, Informant A interpreted the instructions to mean "car companies"; Informant B likes binomial nomenclature but, like Informant C, has interpreted the task as asking for "car models"; and Informant D has interpreted the question functionally rather than along brand lines. (This sort of diversity in response indicates that our original expression, "kinds of American cars," is ambiguous and needs refining.) On the other hand, if we did not know this domain well, then our initial tally would have to rely on linguistic differences to identify different items. And on this basis, there are 20 "differently named" items in the four lists.

There are two customary ways to deal with these sorts of item identification problems. Immediately after completing the task, each informant can be asked to list alternative names for each item in his or her list. Second, the researcher can compile an initial aggregate table, then ask several of the more knowledgeable natives to judge the distinctiveness of the items. Either way, the idea is to enlist native experts to eliminate redundancy, but even these potential solutions are often ineffective, as illustrated in Table 2.

These problems aside, free-listing tasks is an excellent research tool to explore cultural domains (Gate-wood, 1983), with the added benefit that their results enable interesting compar-isons both across domains and across informants (Gatewood, 1984). If a single sample

Culture . . . One Step at a Time *(continued from page 6)*

of informants is used, domains can e compared in terms of various indices, such as median length of list and number of different items generated. At the same time, informants -- either as individuals or grouped by age, gender, expertise, ethnicity, etc. -- can also be compared.

Table 2.
Two Alternative Aggregations of Free-Lists A, B, C, and D

Item	Freq.	Item	Freq.
1. Ford Taurus : Taurus	2	1. Ford : Ford Taurus :Taurus : Explorer	4
2. Chevrolet Corvette : Corvette	2	2. Chevy : Chevrolet Corvette : Corvette	3
3. Dodge Neon : Neon	2	3. Dodge : Dodge Neon : Neon	3
4. Ford	1	4. Jeep : Jeep Cherokee : Jeep Wrangler	3
5. Chevy	1	5. Cadillac : Seville	2
6. Dodge	1	6. minivan	1
7. Cadillac	1	7. sport utility vehicle	1
8. Jeep	1	8. sedan	1
9. Jeep Cherokee	1	9. convertible	1
10. Jeep Wrangler	1	10. station wagon	1
11. Explorer	1		
12. Seville	1		
13. minivan	1		
14. sport utility vehicle	1		
15. sedan	1		
16. convertible	1		
17. station wagon	1		

Pile-Sorts

After identifying the items in a cultural domain, we can begin to examine their meanings and interrelationships, or their similarities and differences. In the 1950s and 1960s, the preferred technique was componential or semantic feature analysis (Goodenough, 1956; Lounsbury, 1956). But because different feature analyses can be formulated to explain the same data (the so-called "psychological reality" problem), research interest has shifted to studying the most salient or important semantic features in a given domain (see D'Andrade, 1995, pp. 31-91). Pile-sorts are an easy way to collect data toward this end.

The most commonly used variant of the pile-sort task is the *single* pile-sort, in which the informant is asked to group items based on their *overall similarity*. Informants are given a collection of stimuli -- either the items themselves, cards with names of items, or pictures of items. The basic task is to group the stimuli such that "similar" items are in the same pile. Informants are free to define similarity in their own terms, to make as many piles as they want, and to place very unusual items in piles by themselves. Indeed, the only constraints are (a) there must be more than one pile (extreme "lumping" is disallowed), and (b) every item cannot be in a pile by itself (extreme "splitting" is disallowed). The researcher then records each informant's pile-sort and asks why items were placed in their piles. For example, Fred's and Janine's sortings of 19 kinds of fish might be recorded as shown in Table 3.

Singleton items -- such as marlin and shad in Fred's sorting -- end up being scored as dissimilar from every other item. Given his rationales, Fred's two singleton items are properly separated. By contrast, Janine's fifth pile is problematic. She has lumped carp, marlin, and shad together *because* she doesn't know anything about the three of them, i.e., the only thing they have in common is that Janine doesn't know anything about them. The proper way to handle such "unknown" items is to treat each as a singleton; hence, when recording Janine's pile-sort, the researcher should split her residual fifth category into three singleton piles. Only by asking informants for brief rationales, however, can we catch such bogus groupings and prevent informants' lack of knowledge from distorting the results.

In doing a single pile-sort, each informant is essentially judging the similarity of every item vis-a-vis every other item, using a dichotomous scale -- any two items are either in the same pile (similar) or they are in different piles (not similar). Thus, when preparing these data for analysis, an informant's similarity judgments are represented as an item-by-item matrix in which each cell contains either 1 (items appear in the same pile) or 0 (items appear in different piles), and each informant produces one such matrix. The aggregate judged similarity for any two items, $item_i$ and $item_j$, is calculated by adding the values in $cell_{i,j}$ across all *N* informants and dividing by *N*. The resulting number is the proportion of informants in the sample who placed $item_i$ and $item_j$ in the same pile. The researcher can then examine the most salient similarities and differences among items in the domain by using such multivariate statistical techniques as multidimensional scaling or hierarchical clustering to analyze the aggregate similarity matrix or consensus analysis (Romney, Weller, & Batchelder, 1986) to analyze inter-informant agreement.

Table 3.
Two Informants' Single Pile-Sorts of 19 Kinds of Fish

Fred's Piles
1. barracuda, piranha, shark
2. bass, carp, pike, sunfish, trout
3. marlin
4. catfish, cod, flounder, herring, salmon, swordfish, tuna
5. goldfish, minnow
6. shad

Janine's Piles
1. piranha, shark
2. barracuda, cod, flounder, swordfish, tuna
3. bass, catfish, pike, sunfish, trout, salmon
4. goldfish, herring, minnow
5. carp, marlin, shad

Fred's Rationales
1. dangerous to humans
2. freshwater (mostly sport) fish
3. ocean sport fish
4. grocery store fish
5. weird, little fish compared to the rest
6. don't know what this is

Janine's Rationales
1. dangerous fish
2. ocean fish
3. fish found in lakes and streams
4. very small fish
5. don't know what these are

The principal advantages of the single pile-sort are that the task is easy to administer, informants enjoy doing it, and it can be done with a relatively large number of items (i.e., as many as 30-50). Also, one can achieve reliability coefficients of .90 or higher with respect to the aggregate similarity matrix using as few as 30-40 informants (Weller & Romney, 1988, p. 25). On the other hand, since informants are free to come up with different numbers of piles, single pile-sorts have limited utility for comparing individuals. In particular, the "lumper" versus "splitter" variation tends to overwhelm other characteristics that might differentiate informants. Other variants of the pile-sort task, such as multiple sorts and successive pile-sorts, obtain more information per informant and, by imposing uniform constraints, are better suited to comparing informants. Weller and Romney (1988, pp. 20-31) provide an excellent discussion of the strengths and weaknesses of various pile-sort tasks.

Triadic Comparisons

Another way to obtain overall similarity judgments is to present three items at a time, and to ask informants to pick the one that is most different from the other two. The procedure is repeated until each item has been presented in a triad with every pair of other items. To avoid uncontrolled order effects and response biases, however, it is important to randomize the presentations, both among and within the triadic sets. Using this method to obtain similarity judgment among four kinds of fish, for example, we might obtain results like those in Table 4.

Table 4.
One Informant's Triadic Similarity Judgments Among Four Kinds of Fish (Capitalization indicates the informant's "odd item out" judgments)

Triad 1:	BASS	salmon	trout
Triad 2:	tuna	BASS	salmon
Triad 3:	trout	bass	TUNA
Triad 4:	salmon	tuna	TROUT

Like the single pile-sort, such data can be represented as an item-by-item matrix. Each triad involves three pairwise comparisons, i.e., ABC breaks down into three pairs: AB, AC, and BC. Thus, for each triad, there are three similarity scores: the pair of items not chosen is judged similar (scored 1), and the two other pairings are judged not similar (scored 0). Following this scoring procedure, Table 5 shows the matrix representation of the information contained in Table 4. (Note: Because similarity data are symmetric, we display only the lower half of the matrix.)

Table 5.
Matrix Representation of Data from Table 4

	Bass	Salmon	Trout	Tuna
Bass	....	....	....	....
Salmon	0	....	....	....
Trout	1	1	....	....
Tuna	0	2	0	....

Informants are quick to understand this triadic comparison task, whether it is administered in written or oral form. Another advantage of the method is that, because all informants perform the same task, the resulting data enables comparisons among individuals. And as long as there are only a few items in the domain, informants find the task mildly amusing. Unfortunately, as the number of items increases, the number of triads required rises dramatically -- the combination of n things taken three at a time, or $n! / [3!(n-3)!]$ -- and informants quickly lose patience when confronting a large number of triads. For example, 8 items require 56 triads, 10 items require 120 triads, and 19 items require 969 triads.

Although one can use a balanced incomplete block design (BIB) to reduce the number of triads given each informant, this forces the researcher to decide whether to focus on differences among items within the domain or on informant differences with respect to the domain. If the objective is to compare informants, then the same subset of triads should be given to each informant; whereas if the focus is more on the items themselves, then each informant should get a different randomly selected instance from the BIB (see Borgatti, 1996, for a fuller discussion of this point). Thus, practical concerns dictate that the number of items be no more than 8-10 for complete designs, and no more than about 25 for BIB designs (Weller & Romney, 1988, p. 37).

In a complete design, such as the example in Table 4, each pair of items occurs in n-2 triadic sets. For example, salmon and tuna were judged similar in both of the triads in which they occurred, whereas salmon and trout were judged similar in only one of their two co-occurrences. If we had eight items, then each pair would co-occur in six triadic sets, and so forth. Following this logic (and modified appropriately for balanced incomplete block designs), the aggregate similarity for any two items can be expressed as a proportion, i.e., the number of triads in which the two items are judged similar by all informants divided by the total number of triads in which the two items are presented to all informants. Thus, aggregated data from triadic comparisons are similar in form to the results from pile-sorts and amenable to the same kinds of statistical analyses.

Conclusion

My main purpose in this essay has been to explain enough about free-listing, pile-sorting, and triadic comparisons to motivate readers to learn more. I might also mention that medical anthropologists and other applied researchers are increasingly using these techniques (e.g., Boster & Weller, 1990; Garro, 1986; Mathews, 1983; Ryan, Martinez, & Pelto, 1996; Weller, 1984; Weller & Mann, 1997). Although I have not covered current approaches to the statistical analysis of these kinds of data, I will do so in a column to appear in the next issue of the newsletter.

In closing, let me note that these approaches to data collection and analysis are much less laborious than they once were thanks to ANTHROPAC (Borgatti, 1998), a PC-based software package. Whereas tabulating a free-list task with 40 informants used to take hours or days of uninterrupted work, ANTHROPAC reads in text files, tallies items, and computes the appropriate descriptive statistics in seconds. What's more, ANTHROPAC not only reads and analyzes pile-sort and triadic comparison data, but can also generate randomized triads questionnaires using many different BIB

Culture . . . One Step at a Time *(continued from page 9)*

designs. It even includes a variety of analytical tools, such as consensus analysis, multidimensional scaling, and hierarchical clustering. To learn more about this worthwhile software package, including pricing, see Analytic Technologies, web page: http://www.analytictech.com/.

References

Bernard, H.R. (1994). *Research methods in anthropology, (2nd Ed.).* Thousands Oaks, CA: Sage.

Bernard, H.R. (Ed.) (1998). *Handbook of methods in cultural anthropology.* Walnut Creek, CA: Altamira Press.

Borgatti, S.P. (1996). *ANTHROPAC 4.0 methods guide.* Natick, MA: Analytic Technologies.

Borgatti, S.P. (1998). *ANTHROPAC, Version 4.95X.* Natick, MA: Analytic Technologies.

Boster, J.S., & Weller, S.C. (1990). Cognitive and contextual variation in hot-cold classification. *American Anthropologist,* 92, 171-179.

D'Andrade, R. (1995). *The Development of cognitive anthropology.* New York: Cambridge University Press.

Frake, C.O. (1962). The ethnographic study of cognitive systems. In T. Gladwin and W.C. Sturtevant (Eds.), *Anthropology and human behavior* (pp. 72-85). Washington, DC: Anthropological Society of Washington.

Frake, C.O. (1964). Notes on queries in ethnography. *American Anthropologist,* 66 (3, Part 2), 132-145.

Garro, L.C. (1986). Intracultural variation in folk medical knowledge: A comparison between curers and noncurers. *American Anthropologist,* 88, 351-370.

Gatewood, J.B. (1983). Loose talk: Linguistic competence and recognition ability. *American Anthropologist,* 85, 378-387.

Gatewood, J.B. (1984). Familiarity, vocabulary size, and recognition ability in four semantic domains. *American Ethnologist,* 11, 507-527.

Goodenough, W.H. (1956). Componential analysis and the study of meaning. *Language,* 32, 195-216.

Lounsbury, F.G. (1956). A semantic analysis of Pawnee kinship usage. *Language,* 32, 158-194.

Mathews, H.F. (1983). Context-specific variation in humoral classification. *American Anthropologist,* 85, 826-847.

Romney, A.K., Weller, S.C., & Batchelder, W.H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist,* 88, 313-338.

Ryan, G., Martinez, H., & Pelto, G. (1996). Methodological issues for eliciting local signs/symptoms/ illness terms associated with acute respiratory illnesses. *Archives of Medical Research,* 27, 359-365.

Weller, S.C. (1984). Consistency and consensus among informants: Disease concepts in a rural Mexican town. *American Anthropologist,* 86, 966-975.

Weller, S.C. & Romney, A.K. (1988). *Systematic data collection*: Qualitative research methods, Volume 10. Newbury Park, CA: Sage.

John B. Gatewood is Professor of Anthropology at Lehigh University. He has roughly 35 professional publications in the areas of cognitive anthropology, fisheries ethnography, tourism studies, and linguistic anthropology. In addition, Dr. Gatewood has considerable experience doing marketing and advertising research for a variety of firms. His current research interests include distributive models of culture, network analysis of organizations, motivations of heritage tourists, and theoretical issues concerning the 'units' of culture.

HaPI Advisory Board

What Does It Measure?: Using Measurement Modeling to Clarify the Construct(s) Underlying the Affect Intensity Measure (AIM)

Fred B. Bryant

In previous columns (Bryant, 1997, 1998, 1999), I described a versatile, new data-analytic approach to construct validation known as "measurement modeling." With this approach, one systematically compares alternative ways of conceptualizing the construct or constructs that a particular instrument taps using powerful, state-of-the-art, multivariate statistical tools, in order to clarify what these instruments actually measure. Measurement modeling provides researchers with a host of invaluable benefits. But perhaps its most important psychometric contribution is its ability to improve construct validity by distinguishing instruments that measure a single, unitary construct from instruments that tap multidimensional constructs, and by further decomposing the latter into their constituent parts. This increased conceptual precision not only helps researchers choose appropriate instruments for a given purpose, but also reveals how to score responses to these instruments so as to assess the underlying construct(s) with maximum reliability.

To review the basics as covered in these previous columns (Bryant, 1997, 1998, 1999), measurement modeling, also known as confirmatory factor analysis, is a specific form of structural equation modeling that examines the "structure" of people's responses to a set of questions. "Structure" refers to the relationships among the responses to the individual questions, and to the underlying construct or constructs (called "factors") that these interrelationships define. A cluster of questions that produce similar responses (i.e., that intercorrelate) is considered to reflect or define a single common factor. A set of measures may have any number (i.e., zero or more) of such clusters of interrelated questions, i.e., underlying factors. Questions that strongly define a particular factor are said to "load" highly on that factor or to have a strong "factor loading." In other words, each question's loading on a particular factor indicates how strongly responses to that question define the underlying construct that the given factor taps.

To impose a formal "measurement model" on a set of questions, one specifies: (a) the number of constructs or factors underlying responses to the set of questions; (b) the specific questions that reflect each of these factors; (c) whether or not multiple factors, if they exist, are intercorrelated; and (d) whether the unique variance in responses to each question (i.e., the variance that is unrelated to the underlying factor) is independent or intercorrelated across the questions. By contrasting how well alternative measurement models explain responses to a set of questions (e.g., one-factor versus two-factor versus three-factor models), researchers can determine whether a particular instrument measures more than one construct, and if so, what these multiple constructs are and how they relate to each other. This work not only improves how we use measurement instruments, but also refines our conceptual understanding of what these instruments actually assess. (For further details about measurement modeling, see Kline, 1998.)

The following summary of research on the Affect Intensity Measure (AIM; Larsen, 1984; Larsen & Diener, 1987) illustrates concretely how measurement modeling can be used to achieve these important benefits. Larsen (1984) developed the AIM to assess the personality trait of affect intensity, or the characteristic strength with which people experience emotions. Analogous to a kind of "emotional thermostat," affect intensity reflects one's emotional temperament. Larsen (1984) originally conceptualized affect intensity as a unidimensional construct – that is, people are assumed to have a trait-like tendency to feel a particular level of emotion, regardless of whether this emotion is positive or negative. The underlying theoretical model is based on the assumption that high intensity individuals actually experience *lower* levels of emotional arousal than do low intensity individuals, but that they express *higher* levels of emotion to try to achieve an optimal level of internal arousal (Larsen & Diener, 1987).

The AIM consists of 40 statements that are intended to reflect one's characteristic level of emotion, both in general and in response to specific situations. These items cover a wide range of both positive emotions (calmness, contentment, delight, ecstasy, elation, enthusiasm, euphoria, excitement, exuberance, joy, jubilation, peacefulness, relaxation, zest) and negative emotions (anger, anxiety, guilt, nervousness, sadness, shame, tension). Respondents are instructed to indicate on a five-point scale (1 = never; 2 = almost never; 3 = occasionally; 4 = usually; 5 = almost always) how characteristic each statement is of them.

Virtually all research using the AIM has followed Larsen's (1984) original theoretical model, which regards affect intensity as a unidimensional construct. Accordingly, researchers have typically summed responses to the 40 AIM items to obtain a global total score. A great deal of empirical evidence suggests that individual differences in total AIM score are temporally and situationally stable and are related to personality in a conceptually meaningful way. For example, total AIM score has been found to correlate with the extremity of daily moods and the frequency of emotional swings, parental reports of early childhood behaviors indicative of temperamental intensity, the strength of physiological and expressive changes associated with emotion, scores on psychosomatic symptom checklists, measures of risk for cyclothymia and bipolar affective disorder, and many important personality characteristics, social behaviors, and emotional responses. Moreover, researchers have found that 13-14% of the variance in total AIM score is linked to genetic factors.

However, there is also evidence that the AIM is multidimensional. Discussing Larsen's (1984) original work, for example, Diener, Sandvik, and Larsen (1985) stated that the AIM assesses at least five underlying factors: positive affect intensity, negative affect intensity, preference for arousal, general emotional intensity, and visceral reactivity to emotional events. Similarly, Williams (1989) reported an exploratory factor analysis of the AIM that revealed four underlying factors, two affectively-positive (which correlated with extraversion) and two affectively-negative (which correlated with neuroticism). Whether the AIM is unidimensional or multidimensional is an important issue, because a total score that collapses across multiple constructs could distort hypothesized relationships between different aspects of affect intensity and other constructs.

About six years ago, my colleagues and I embarked on a program of research using measurement modeling to investigate more carefully whether the AIM measures a single, unitary construct or multiple sub-constructs. This work resulted in two articles (Bryant, Yarnold, & Grimm, 1996; Weinfurt, Bryant, & Yarnold, 1994) that illustrate the use of measurement modeling to: (a) determine what an instrument actually measures, (b) clarify how best to score responses to an instrument, and (c) refine our conceptual understanding of the constructs involved.

We began our first study (Weinfurt et al., 1994) by administering the AIM to 673 undergraduates and then using measurement modeling to impose both Larsen's original one-factor model and William's four-factor model on these data. Contrary to the notion that the AIM measures a single, unitary construct, the four-factor model (which explained 80% of the variance in responses to the AIM) fit the data significantly better ($p < 0.0001$) than did the one-factor model (which explained only 62% of the variance in responses). Although these results strongly suggest that affect intensity is multidimensional, the four-factor model fails to explain 90% or more of the variance in responses to the AIM, the standard by which a measurement model is considered adequate (Bentler & Bonett, 1980). Accordingly, we continued to search for a better-fitting measurement model for the AIM.

Scrutinizing the multiple dimensions of affect intensity more carefully, we realized that they incorporated two critical distinctions: positive versus negative valence, and intensity (strength) versus reactivity (responsiveness). By explicitly crossing these two distinctions, a four-factor model emerges that consists of self-evaluations of one's predisposition to experience: (a) positive intensity, (b) negative intensity, (c) positive reactivity, and (d) negative reactivity. We termed this structure the AIR model, for Affect Intensity and Reactivity. To develop a measurement model that explicitly embodied

What Does It Measure? *(continued from page 12)*

these dimensions, we (Bryant et al., 1996, Study 1) began by sorting the 40 AIM items into a subset of 27 items that could be judged *a priori* as indicative of either the characteristic intensity or reactivity of either positive or negative emotion. We categorized seven AIM items as reflecting positive intensity (e.g., item 2: "When I feel happy it is a strong type of exuberance"); six as reflecting negative intensity (e.g., item 30: "When I do feel anxiety it is normally very strong"); eight as reflecting positive reactivity (e.g., item 23: "When I receive an award I become overjoyed"); and six as reflecting negative reactivity (e.g., item 11: "Sad movies deeply touch me"). We discarded 13 AIM items because we could not unequivocally classify them as reflecting one of these four dimensions of affective experience (e.g., item 3: "I enjoy being with other people").

Which measurement model better explains responses to these 27 AIM items -- Larsen's original one-factor model that assumes affect intensity is a unitary construct, or a four-factor AIR model that assumes people report separate experiences of positive and negative intensity and reactivity? To answer this key question, we administered the 40-item AIM to an independent sample of 631 undergraduates. We then used measurement modeling to compare how well the one-factor "total score" model and the four-factor AIR model explained responses to the 27 items for both this new sample and Weinfurt et al.'s (1994) earlier sample as well.

For both samples, the four-factor AIR model explained responses to the 27 AIM items significantly better ($p < 0.0001$) than did the one-factor model. Moreover, the four-factor model explained 83% and 85% of the variation in the responses in the two samples, respectively, whereas the one-factor model explained only 62% and 66%, respectively. While inspecting the relationships among the four AIR factors, however, we noticed that the positive intensity and positive reactivity factors were highly intercorrelated -- .92 for sample 1 and .90 for sample 2, whereas the negative intensity and negative reactivity factors intercorrelated at only .55 for both samples. (Evidently, affect intensity and reactivity are different in relation to positive versus negative emotions, with the distinction between feeling and expression being much more relevant for negative affect. This may be because negative emotions often have more harmful social consequences then do positive emotions, and thus negative emotions are more likely to be repressed or inhibited.) With the high correlation between positive intensity and positive reactivity, we combined these two factors to produce a three-factor AIR model (positive affectivity, negative intensity, negative reactivity) that achieved the same degree of goodness-of-fit as did the four-factor model. Because this three-factor AIR model provides equivalent statistical precision but greater parsimony than the four-factor model, it is currently the best measurement model for the AIM.

In the final phase of this research, we (Bryant et al., 1996, Study 2) investigated whether the three AIR factors contribute more, both conceptually and statistically, to understanding and predicting an important personality characteristic than does total AIM score. To accomplish this, we examined the relationship between affect intensity and dispositional empathy as measured by the Interpersonal Reactivity Index (IRI; Davis, 1983). First, we tested the hypothesis that the three AIR factors in combination would do a better job of predicting dimensions of dispositional empathy -- empathic concern, perspective taking, personal distress, and fantasy -- than would total AIM score. Second, we assessed the discriminant validity of the three AIR factors, relative to that of the unidimensional total AIM score, in predicting dimensions of dispositional empathy. For example, would positive affectivity, relative to negative intensity or negative reactivity, show a different pattern of relationships with the empathy dimensions?

We began by administering the AIM and the IRI to a new sample of 218 undergraduates. Once again, for this new sample, the three-factor AIR model provided a significantly better ($p < 0.0001$) measurement model for the 27 AIM items than did the one-factor "total score" model. We next used regression analyses to predict each of the four IRI factors using total AIM score first and then the three AIR factors. If affect intensity is truly unidimensional, as Larsen (1984) argued, then using the three AIR

What Does It Measure? *(continued from page 13)*

factors together as predictors should explain no more variance in dispositional empathy than using total AIM score as a global predictor. However, for each of the four empathy dimensions, the three AIR factors together explained more variance than did total AIM score, with the difference in r^2 ranging from a low of 8% (for fantasy) to a high of 125% (for perspective taking). These results clearly show the greater predictive utility of the three-factor AIR model.

Investigation of discriminant validity found that no two AIR factors showed the same pattern of relationships with the four IRI factors, and none of the IRI factors showed the same pattern of relationships with the three AIR factors. For example, positive affectivity predicted greater empathic concern and greater empathic fantasy, but not personal distress and perspective taking. Negative affect intensity predicted greater personal distress and empathic fantasy, but not perspective taking and empathic concern. Negative reactivity predicted greater empathic concern, greater personal distress, and greater perspective taking, but was unrelated to empathic fantasy. The multidimensional (three-factor) model of affect intensity thus demonstrated superior conceptual and predictive precision relative to the unidimensional total AIM score. Thus, affect intensity, as operationalized by the AIM, is multidimensional rather than unidimensional.

Further supporting the discriminant validity of the three AIR factors, women reported *higher* levels of negative reactivity than of positive affectivity ($p < 0.0001$), whereas men reported *lower* levels of negative reactivity than of positive affectivity ($p < 0.0001$). Thus, women say they are more emotionally reactive to negative events than to positive, whereas men say they are more emotionally reactive to positive events than to negative. (This pattern of results may reflect sex differences in socialization that encourage females to express or exaggerate negative feelings, and encourage males to suppress or deny negative feelings, in response to aversive events.)

Our research has an important implication for anyone who uses the AIM to measure affect intensity. Specifically, using total AIM score to operationalize affect intensity may well obscure findings that would otherwise emerge using the three AIR factors. Clearly, total AIM score does not provide the same picture of affect intensity as the three-factor model.

Does this mean that researchers should not use a global measure of affect intensity? The answer is that it depends on one's purpose. If, for example, one wants to obtain a global assessment of "affect intensity" as a general personality trait, then the unidimensional model is appropriate. If, however, one wants to use "affect intensity" to predict other traits or outcomes, then the multidimensional model is more appropriate. With regard to the unidimensional approach, however, in the case of "affect intensity," our research suggests that summing the 27 items from the AIR model into a single, total score provides a better global measure of general affect intensity than does the 40-item total AIM score.

As a more general point, this work on the Affect Intensity Measure illustrates how measurement modeling can be used to refine our understanding of the constructs that instruments measure, determine the most reliable and informative methods of scoring instruments to improve conceptual clarity and statistical precision, and enhance the effectiveness with which we use instruments. These important benefits make measurement modeling an invaluable psychometric tool in the behavioral sciences.

References

Bentler, P.M., & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588-606.

Bryant, F.B. (1997). The comparative anatomy of related instruments: An emerging specialty. *The Behavioral Measurements Letter*, 4 (2), 7-9.

Bryant, F.B. (1998). Measurement modeling: A tool for investigating the comparative anatomy of related instruments. *The Behavioral Measurements Letter*, 5 (2), 14-17.

Bryant, F.B. (1999). Measurement modeling: Identifying the constructs underlying the Center for Epidemiologic Studies Depression Scale (CES-D). *The Behavioral Measurements Letter*, 6 (1), 6-9.

What Does It Measure? *(continued from page 14)*

Bryant, F.B., Yarnold, P.R., & Grimm, L.G. (1996). Toward a measurement model for the Affect Intensity Measure: A three-factor structure. *Journal of Research in Personality*, 30, 223-247.

Davis, M.H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44, 113-126.

Diener, E., Sandvik, E., & Larsen, R.J. (1985). Age and sex effects for emotional intensity. *Developmental Psychology*, 21, 542-546.

Kline, R.G. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Larsen, R.J. (1984). Theory and measurement of affect intensity as an individual difference characteristic. *Dissertation Abstracts International*, 85, 2297B. (University Microfilms No. 84-22112).

Larsen, R.J., & Diener, E. (1987). Affect intensity as an individual difference characteristic: A review. *Journal of Research in Personality*, 21, 1-39.

Weinfurt, K.P., Bryant, F.B., & Yarnold, P.R. (1994). The factor structure of the Affect Intensity Measure: In search of a measurement model. *Journal of Research in Personality*, 28, 314-331.

Williams, D.G. (1989). Neuroticism and extraversion in different factors of the Affect Intensity Measure. *Personality and Individual Differences*, 10, 1095-1100.

Fred Bryant is Professor of Psychology at Loyola University, Chicago. He has roughly 80 professional publications in the areas of social psychology, personality psychology, measurement, and behavioral medicine. In addition, he has coedited 5 books, including **Methodological Issues in Applied Social Psychology** *(New York, Plenum Press, 1993). Dr. Bryant has extensive consulting experience in a wide variety of applied settings, including work as a research consultant for numerous marketing firms, medical schools, and public school systems; a methodological expert for the U.S. Government Accounting Office; and an expert witness in several federal court cases involving social science research evidence. He is currently on the Editorial Board of the journal* **Basic and Applied Social Psychology**. *His current research interests include happiness, psychological well-being, Type A behavior, the measurement of cognition and emotion, and structural equation modeling.*

The BEHAVIORAL MEASUREMENT Letter

Behavioral Measurement Database Services

In This Issue:

The Behavioral Measurement *Letter*

Behavioral
Measurement
Database
Services

BMDS

PO Box 110287 • Pittsburgh, PA 15232-0787

DATED MATERIAL

Bulk Rate
U.S. Postage
PAID
Permit No. 1235
Pittsburgh, PA