DOUBLE ISSUE

The BEHAVIORAL MEASUREMENT *Letter*

Behavioral Measurement Database Services

Enriching the health and behavioral sciences by broadening instrument access

Vol. 8, No. 2
Winter 2005

Introduction to the January 2005 Issue of The Behavioral Measurement Letter

This issue of *The* Behavioral Measurement *Letter* includes four articles addressing a diverse range of measurement-related topics: Looking back fondly on the life of a legendary pioneer and measurement advocate in the field of nursing research; addressing test-takers' right to receive feedback about their test scores; helping researchers select tools and methods in cross-cultural research; and comparing and contrasting measurement in the behavioral and natural sciences. Note the multidimensional threads of measurement that run throughout these articles, weaving the different topics together into a rich and varied conceptual tapestry.

To begin this issue of *The* Behavioral Measurement *Letter*, Deidre Blank warmly remembers the life and times of Doris Bloch, a pillar in the field of nursing research. Dr. Bloch, who died in August 2003, was instrumental in building a formal structure for federal research support in nursing, and her contributions live on in the legacy she leaves behind in the health sciences. Doris was also a champion and ally of the Health and Psychosocial Instrument (HaPI) database, from its very inception. It is with deepest respect that we fondly dedicate this issue of *The* Behavioral Measurement *Letter* to the life, work, and memory of Doris Bloch.

Also in this issue of *The* Behavioral Measurement *Letter*, Robert Perloff makes a strong argument in support of test-takers' unalienable right to know the meaning and interpretation of their test scores, whenever test-takers willingly give their responses to a tester for scoring, analysis, and interpretation. Advocating a *quid pro quo* arrangement, he argues that anyone who uses test-takers' scores is ethically obligated to provide test-takers with feedback about the meaning of their test scores, unless an individual respondent voluntarily waives his or her right to receive such feedback. Dr. Perloff's clarion call for justice in ethical standards to guarantee the rights of individual test-takers suggests important extensions of informed consent procedures when conducting research with human participants. Readers interested in responding to Dr. Perloff's article should submit either a Letter to the Editor or a brief manuscript to the address provided below.

Reflecting the growing increase in cross-cultural research, Carolyn Waltz offers researchers studying intra-cultural (within culture) or inter-cultural (between culture) differences a framework for maximizing the validity and reliability of the measurements involved. Advocating a solid psychometric foundation, she highlights the preconditions necessary to optimize the validity and reliability of a particular measurement instrument in cross-cultural research. Dr. Waltz emphasizes the absolute necessity of (a) using an appropriate translation strategy when creating a new form of an original instrument in another language, and (b) demonstrating that the construct being measured is equivalent both within and across cultures by establishing psychometric equivalence intra- and inter-culturally.

Finally, in this issue of *The* Behavioral Measurement *Letter*, Fred B. Bryant compares

Introduction (*continued*)

and contrasts the process of measurement in the behavioral sciences (in which he includes the health and social sciences) and the natural sciences. Arguing that researchers in the behavioral sciences have much to learn about measurement from their counterparts in the natural sciences, Dr. Bryant presents excerpts from an interview of a research colleague in biochemistry that highlight basic similarities and differences between the two disciplines. Dr. Bryant uses the interview excerpts to identify a core set of concerns—involving instrumentation, measurement error, measurement validity, and graduate training in measurement—that interconnect and distinguish measurement in the behavioral and natural sciences.

We invite written responses from our readership. Please address comments, suggestions, letters, or ideas for topics to be covered in future issues of the journal to: The Editor, *The* Behavioral Measurement *Letter*, Behavioral Measurement Database Services, P.O. Box 110287, Pittsburgh, PA, 15232-0787. Email: bmdshapi@aol.com

We also consider short manuscripts for publication in *The* Behavioral Measurement *Letter*. Submit, at any time, a brief article, opinion piece, or book review on a topic related to behavioral measurement, to The Editor at the above address. Each submission will be given careful consideration for possible publication in a forthcoming issue of *The* Behavioral Measurement *Letter*.
HaPI reading…

The Editor

Email: bmdshapi@aol.com

> The purpose of computing is insight not numbers.
> R. Hamming

Doris Bloch Remembered

Deidre M. Blank, DrPH, RN, FAAN

On August 10, 2003, nursing lost a dear friend and colleague. Doris Bloch, DrPH, RN, FAAN, retired Special Assistant to the Director of the National Institute of Nursing Research (NINR), died of congestive heart failure at Suburban Hospital in Bethesda, MD. She was 75 years old.

Doris was a legend among nursing research leaders. Her fame was based largely on her early work developing federal programs for nursing research while at the Division of Nursing, Health Resources and Services Administration (HRSA). In 1986, Doris led the transfer of nursing research and research training staff from the Division of Nursing, HRSA, to the newly created National Center for Nursing Research (NCNR) at NIH, the forerunner of today's NINR. As a Branch Chief at the NCNR, Doris assumed primary responsibility for Program Planning and Evaluation before finishing her career as Special Assistant to the Director of the NINR. Always the faithful steward for nursing research, Doris did much of the planning for the National Nursing Research Agenda during the early years for nursing at NIH.

A native of Berlin, Doris and her sister escaped Nazi persecution in Germany and settled in the Netherlands. After completing high school in the Netherlands, Doris immigrated to the U.S. She received her bachelor's degree in zoology from Mount Holyoke College, a master's degree in nursing from Yale, and master's and doctoral degrees in public health from the University of California at Berkeley. She then served a distinguished career in public health, including work in Tanganyika (now Tanzania), Kenya, and the Philippines. After returning to the U.S., she joined the Division of Nursing, HRSA.

Doris was a dedicated professional who provided exemplary service to the nursing profession. In her quiet and unassuming way, she was a mentor to many nurses who today serve in pivotal leadership positions. As a young nurse who was assigned to work with Doris in her Section at the

Comparing Measurement in the Natural and Behavioral Sciences (*continued*)

DMF: *When I embarked in lithium research, the major tool for lithium detection was atomic absorption spectrophotometry. However, atomic absorption spectrophotometry can only be used to measure total lithium concentrations in biological samples, and cannot be used to discriminate between free and bound forms of lithium. It was proposed that a cell membrane abnormality was present in bipolar patients. If that is the case, one would anticipate that the distribution of lithium between free and bound forms in cell membrane preparations from bipolar patients and normal individuals would be different in disease and healthy states. My research group spent the first 8 years or so developing NMR spectroscopy of lithium-7 and demonstrated that this technique could indeed be used to discriminate free and bound forms of lithium and that this lithium distribution was indeed different in disease and healthy states.*

FBB: Are any formal resources available to help bioinorganic chemists like you identify alternative ways of measuring the phenomena you want to study? (The Health and Psychosocial Instruments database provides this measurement information for behavioral researchers.)

DMF: *No. In the natural sciences, finding the best method and experimental conditions to address a question requires creative insight. There are several regional laboratory facilities, which are funded by the Federal Government, across the United States. Although these facilities are well equipped with state-of-the-art instrumentation and are available for use by academic researchers, the traveling time, the limited availability of each instrument, and/or the relatively costly user fees preclude most researchers from having good access to every type of instrumentation for a specific research project. Any research laboratory or, for that matter, any academic department in the natural sciences cannot afford to purchase and maintain every conceivable type of sophisticated instrumentation. Most often, a given research laboratory specializes in the use of a relatively small number of physical methods and focuses on research problems that can be addressed by those tools.*

Reducing Error and Bias in Measurement

Behavioral scientists and biochemists alike strive to overcome the problems of error and bias in their research by using multiple, complimentary, and independent methods of measurement; and they both wrestle with situations in which the very act of measuring can alter the phenomenon under investigation. However, unlike behavioral scientists, biochemists can check to see if a measurement instrument is working properly by using it to assess a standard specimen with known properties. The following interview excerpt illustrates these points.

FBB: Researchers in the behavioral sciences are often concerned about potential sources of error or unreliability in their measurements. These are random influences on measurements that are unrelated to the underlying variables of interest. What are the potential sources of error in the measurements you make, and what steps do you take to reduce these spurious influences?

DMF: *Potential sources of error include purity of materials, viability of biological samples, and lack of a deep understanding of the physical methods used to obtain the measurements. To reduce these influences, we test for possible metal ion contamination in our biological samples. (It is ironic that some lithium researchers originally collected blood samples from lithium-treated samples in tubes containing the anticoagulant lithium heparin, which contaminated these samples. Because other nonlithium salts are commonly available as anticoagulants, the use of lithium heparin was clearly a poor choice.) We also ensure that the purified proteins and the cells that we use are biologically active and viable. Conducting measurements in solutions of known composition in terms of lithium concentration or other variables that we are trying to measure tests the proper use of the instrumentation.*

FBB: Researchers in the behavioral sciences are also concerned about sources of predictable bias in their measurements. These are systematic influences on measurements that have nothing to

Comparing Measurement in the Natural and Behavioral Sciences (*continued*)

do with the underlying variables of interest. For example, when a researcher knows the experimental hypothesis, he or she might unintentionally influence measurements in ways that confirm the expected results. In the behavioral sciences, researchers often keep the experimenter unaware or "blind" to hypotheses or experimental conditions, to control for this type of experimenter bias. Is this ever a concern in your research? Are there potential sources of systematic bias in the measurements you make?

DMF: This is an issue where, in my opinion, the natural sciences diverge the most from the behavioral sciences in terms of approaches to measurement. When we intentionally influence measurements in a way to test a hypothesis experimentally in the natural sciences, we can be sure that the observed results are due to our intentional influence. For example, we increase the lithium concentration holding everything else constant, and we see if this change has an effect on the magnesium concentration. This way we can make certain the effect is due to the variation in lithium concentration and not due to some other variable, i.e., change in sodium concentration, pH, etc. It seems harder to be sure that observed effects are due to intentional influences in the behavioral sciences.

FBB: What steps do you take to reduce these systematic influences on measurements?

DMF: Oftentimes, by measuring standardized solutions or conducting control experiments using independent methods, one can isolate the contribution of systematic errors. Even when the contribution of systematic errors cannot be quantified, we take great care in discussing in our publications the possible sources of error and the assumptions made in the calculations of parameters derived from direct measurements. In general, we test over and over again our measurements with freshly prepared biological samples to ensure that our measurements are accurate and reliable. We also make extensive use of statistical methods in our data analysis to quantify sampling error and conduct inferential hypothesis tests.

FBB: Although behavioral scientists rely heavily of inferential statistical tests to quantify the probability that chance produced observed effects, natural scientists seem to rely more on visual inspection of graphs in drawing research conclusions. Thus, your use of statistics is somewhat unique for a natural scientist. Why is this?

DMF: Having experienced transitions from fundamental research in inorganic chemistry to more applied research in complex systems in biochemistry, cell biology, and psychiatry during the course of my career, I notice that most basic chemists do not make as much use of statistical methods as they should. It is important to bear in mind, however, that in the basic sciences the elegance of an experiment is intimately related to the design and simplicity of the system being analyzed. Because the number of variables in these simpler systems is small, it is not surprising that any phenomenon observed tends to be huge compared to sometimes-tiny experimental effects observed in the applied physical and biological sciences.

FBB: Another common concern in the behavioral sciences is the possibility that the very act of measurement may well change the things we are studying. Is this "reactivity" issue ever a concern in your area of research? If so, how do researchers address this problem?

DMF: The "reactivity" issue is definitely a source of concern in our research. For instance, if we increase lithium concentrations in an experiment too much, we may run into problems of cell toxicity and have certain chemical reactions become prevalent, which are situations that are not generalizable to the therapeutically relevant concentration range of lithium. This problem could be safely avoided by conducting measurements within the pharmacologically relevant concentration of lithium. However, it is not always possible to avoid this problem because the physical methods used for lithium detection are sometimes not sensitive enough to pick up variations in lithium concentration within the therapeutic range.

Comparing Measurement in the Natural and Behavioral Sciences (*continued*)

Assessing Measurement Validity

Validating measurement instruments is vital in both the behavioral and natural sciences. Unlike natural scientists, however, behavioral scientists usually have no "gold standard" to use in validating a new measure. The best that behavioral scientists can do is correlate scores on the new measure with scores on a preexisting alternative measure of the same variable, to assess what is called "convergent validity." The following interview excerpt illustrates these points.

FBB: In the behavioral sciences, researchers devote a great deal of attention to assessing the validity of their measurement instruments. The key question in such cases is whether a particular instrument really measures what it's supposed to measure (i.e., the issue of construct validity). Do researchers in your field ever question whether their measurement instruments actually assess what these instruments are intended to assess? Can you provide any examples?

DMF: We always question whether the instrumentation used is actually measuring what it is supposed to measure. One of the laboratories with which we compete once reported that the total lithium concentrations in human red blood cells measured by NMR spectroscopy were less than those measured by atomic absorption spectrophotometry. They went on to conclude erroneously that the mechanism of action of lithium must take place outside the cell and not inside it. We later demonstrated that the total concentrations of lithium in human red blood cells were the same regardless of whether they were measured by NMR spectroscopy or atomic absorption spectrophotometry. The source of error for their NMR measurements was that these researchers failed to consider the long NMR relaxation properties of the lithium-7 nucleus.

FBB: In your field of specialty, are there agreed upon "gold standard" measures for assessing key research variables (e.g., a thermometer for assessing temperature)?

DMF: The use of "gold standards" is widespread in the natural sciences. For example, we test the performance of the NMR spectrometer by confirming the known value of the signal-to-noise ratio for the NMR spectrum of a standard solution.

FBB: How do your researchers in your field go about determining whether a measurement instrument actually measures what it is supposed to measure? What kinds of research strategies or procedures do they use to address this question of measurement validity?

DMF: We typically use another physical method to measure the same variable as well as conduct measurements of standard solutions.

FBB: Behavior scientists often use multiple ways of measuring things, in order to determine how much their results are due to their particular method of measurement. Do you ever use more than one method of measuring a phenomenon of interest?

DMF: We almost always use complimentary methods in our research. For example, if we want to test whether competition between lithium and magnesium ions occurs under a certain experimental condition, we can measure an increase (or a decrease) in free lithium concentration by NMR spectroscopy and a decrease (or an increase) in free magnesium concentration by fluorescence spectroscopy.

Training in Measurement and Instrument Development

Another major difference between the behavioral and natural sciences concerns the emphasis placed on measurement in advanced graduate training. Whereas measurement training is typically sporadic and informal in the behavioral sciences, such training is typically an integral part of the graduate curriculum in the natural sciences. This profound difference seems to reflect the complex nature of physical measurement in contrast to behavioral measurements. The following interview excerpt illustrates this point.

Comparing Measurement in the Natural and Behavioral Sciences (*continued*)

FBB: Many behavioral scientists have lamented the general lack of graduate training in measurement. For example, graduate students in psychology rarely receive formal training in how to use research instruments, how to select a suitable instrument for their research needs, or how to develop new instruments. Is this also the case in your area of specialty?

***DMF**: No. If anything, this is one of the primary focuses of graduate training in bioinorganic chemistry.*

FBB: Do you offer this type of training in your own graduate program or in your lab?

***DMF**: Yes. Both through lecture graduate courses in bioinorganic chemistry and in biological applications of NMR spectroscopy and advice during research lab meetings with my graduate students. In addition, in journal clubs, critiques of methodology used in a given research paper most often than not dwells on sensitivity of the methods used to address the questions raised in the publication.*

FBB: To what factor do you attribute this difference in educational emphasis?

***DMF**: Many of the measurement tools in bioinorganic chemistry are highly complex and require in-depth training to be used properly. Thus, by necessity we have made this training an integral part of the graduate curriculum.*

Conclusion

On the one hand, these interview excerpts highlight several important similarities in the nature of measurement in the natural and behavioral sciences. For example, both types of scientists often use multiple methods of assessment in making measurements; and when the size of the effect being investigated is small, both natural and behavioral scientists use inferential statistics to establish the presence of bivariate relationships. But when natural scientists study large effects, they typically rely on visual inspection of data to draw conclusions. Thus, the issue of sampling and measurement error is typically of less concern in the natural sciences.

On the other hand, the interview excerpts also reveal several critical differences in measurement across the two disciplines. For example, unlike natural scientists, behavioral scientists have neither "gold standards" with which to gauge measurement validity, nor "standard solutions" containing known levels of the variables they wish to measure, for use in calibrating and validating measurement tools. As a consequence, it is harder to establish measurement validity in the behavioral sciences than in the natural sciences. Moreover, the measurement tools of behavioral science are usually simpler, less expensive, and more readily available than those of natural science. As a consequence, there is less specialization in methods of measurement and less emphasis on formal measurement training in the behavioral sciences.

Clearly, much work is needed before the behavioral measurement instruments will be as reliable, accurately calibrated, and well established as the instruments of natural science. For some time to come, the complexities and random irrelevancies of human behavior will continue to make measurement an especially thorny concern for behavioral researchers. Indeed, behavioral scientists may never develop the level of accuracy and precision in measurement that natural scientists have achieved.

Fred B. Bryant received his Ph.D. in social psychology from Northwestern University in 1980, and currently teaches courses in social psychology, statistics, and research methodology at Loyola University Chicago. Dr. Bryant has over 100 publications in the fields of personality, cognition and emotion, psychometrics, and structural equation modeling. He has also served as a methodological consultant for the United States Government Accounting Office (GAO), as a statistical consultant for numerous medical centers and internationally prominent marketing firms, and as an expert social science witness in several major Federal Court cases. Email: fbryant@luc.edu

Search for measurement instruments
in the Health and Psychosocial Instruments (HaPI) database ONLINE or on CD-ROM!

HaPI

Produced by BMDS • **Behavioral Measurement Database Services**

SUBSCRIBE TO HaPI!

Call Ovid Technologies at 800-343-0064 for pricing and to order today!

Obtain copies of measurement instruments from BMDS Instrument Delivery!

Call BMDS at 412-687-6850, fax 412-687-5213, or e-mail bmdshapi@aol.com **($20/$30 processing/handling fee)**

HaPI Thoughts

DOUBLE ISSUE

The BEHAVIORAL MEASUREMENT Letter

Behavioral Measurement Database Services

In This Issue:

Now Indexed in CINAHL and Coming Soon in HaPI!

Doris Bloch Remembered (*continued*)

Division of Nursing (and later at the NCNR, NIH), I remember how dedicated Doris was to her work. Typically, she was in her office before her staff arrived in the morning and after they left for the day. However, despite her long hours, she still took time to mentor newcomers. I remember her giving me one of my first opportunities to be a manuscript reviewer, a role that I still enjoy today. I also remember that under Doris' leadership, the staff of her section at the Division of Nursing received the Assistant Secretary for Health's Award for Exceptional Achievement, a well-deserved honor for such a dedicated servant.

I also experienced a more sensitive side to Doris. I remember one day in particular, when I shared with her my plans for a trip with my husband to the Netherlands. Doris told me the story of how, as children, she and her sister had hidden in the Dutch woods to evade Nazi capture. She pinpointed the location on a map, as if the encounter had occurred yesterday. Making a determined effort to visit the area, I remember standing in those same woods feeling a deeper appreciation of this very strong, yet quiet, individual. Also, it was not unusual for Doris to work in our temporary quarters at the NIH on weekends. As a new mother, I would periodically make it a point on weekends to stop by the office with my daughter. Doris always seemed to enjoy the interaction, never failing to laugh and smile. She seemed to relish the innocence of youth.

One of the last things that Doris did before she passed away was to attend the motion picture, *Winged Migration*, with some of her friends. A film of great beauty, it explores the mystery of birds in flight, by following a variety of bird migrations through forty countries across seven continents. How fitting that Doris, who had started her life in flight from the Nazis, would later help give flight to nursing research and watch it rise to great heights, just as she watched the birds in the film soar high above the earth.

Dr. Deidre M. Blank, DSN, RN, FAAN, has served as Chief of the Health Promotion and Disease Prevention Branch at the National Center for Nursing Research (NCNR), NIH, as well as a Nurse Consultant to the Nursing Research Support Section in the Division of Nursing, HRSA; and has held positions on the nursing faculties of Rutgers, Emory, and Thomas Jefferson Universities. A Fellow of the American Academy of Nursing, Dr. Blank has received the Assistant Secretary for Health's Award for Exceptional Achievement and the Director's Award from the NCNR, NIH.
Email: dblank@netzero.com

Staff

Director	Evelyn Perloff, PhD
Associate Director	Linda Perloff, PhD
Record Analysts	Susan Walko, MLS Nathaniel Krumdick, MA Scott King, MA
Indexer	Jacqueline Reynolds, BA
Newsletter Editor	Fred B. Bryant, PhD
Reference Specialist	Barbara Brooks, MLS
Serials Manager	Sally Wilson, MLS
Computer Consultant	Alfred A. Cecchetti, MS
Measurement Consultant	Fred B. Bryant, PhD
Information Specialist Consultant	Ellen Detlefson, DLS
Database Services Manager	Diane Cadwell
Copy Assessor	Tracey Handlovic, BA
Clerical Assistant	Stephanie Jones
Record Checker	Betty Hubbard

The Test-Taker's Right to Know

Robert Perloff, PhD

It is my powerful intuition, reinforced subjectively by anecdotes and hearsay, that in many quarters there is oracular resistance to the test-taker's/examinee's right to know how he or she did, performed, stacked up, scored, or was assessed on a psychological test, attitude survey, personality checklist, diagnostic tool or other (generally) paper-and-pencil instrument designed to describe, measure, or evaluate how test-takers acquitted themselves on instruments of one sort or another.

Opposition to the Release of Test Scores

As a fellow of Division 5 (Evaluation, Measurement, and Statistics) of the American Psychological Association (APA), an erstwhile specialist in psychometrics, and a former APA President, but most of all as a fervid champion of civil rights and of the first Amendment to the United States constitution, I have long been persuaded that the *pas de deux* between an individual's personhood and his or her test scores or results (even if provided in a typical way) of a medical examination, e.g., blood pressure, electrocardiogram, or urinalysis, is binding, inviolate, never to be severed, never!

And yet, while I have not gathered empirical evidence supporting my contention, I am convinced—indeed, dead certain—that many aficionados in evaluation and measurement frown upon providing a person with a frank and unexpurgated, though properly explained to, and interpreted for the lay test-taker, an account of his or her test score. A likely reason offered for this resistance is that a test score is complex—but an electrocardiogram isn't complex?—and it would confuse the test-taker to be confronted with a perplexing array of "ifs, ands, or buts." Why throw test-takers into a deep funk by frightening them with the showy psychometrician's jargon of the likes of standard deviation, stanines, percentiles, probable error, and reliability?' In my judgment, these terms should be viewed by the tester as a challenge, not as a difficulty or as a reason to conceal test scores from the test-taker. If the tester is unable to describe a test score to a high school student, a college sophomore, a soccer mom, or to a guy who, at another time of the week is a patron of a sports bar watching Monday Night Football, then the tester should refrain from administering the test in the first place!

The only legitimate exception against providing test-takers with their test scores would be if test-takers themselves, of their own volition—without being prompted by the tester to refrain from seeing their test scores—opted not to be given their test scores. Of course, this mandate to provide test feedback applies only to tests of established validity, and not to tests of unknown validity. After all, if it is unclear what the test is measuring, then there is no trustworthy feedback to communicate to the test-taker.

The Case for Release of Test Scores

Test results should be unexceptionally (save for the above reason) given to test-takers, jargon-free and comprehensibly. But for the foregoing exception there is no condition, circumstance, or context in which test-takers should not be given their test results. "I'll let you take a snapshot of me, but you'll have to show me the picture"—otherwise, no deal. Either provide me with comprehensible test feedback, or I will not take the test. It's as simple as that.

After all, a test is invasive, puncturing the test-taker's behavioral or psychological epidermis; it probes for, measures, and reports intimate and significant aspects of one's core being. Would it not be objectionable, if the results of medical tests were withheld from the patient? Likewise, it is objectionable that psychological test scores would be withheld from the test-taker. An organization or institution afraid or reluctant to reveal test scores is exposing itself to the accusation that it doubts the validity, relevance, or dependability of the test score. If you're confident about the verisimilitude of the procedure, you should share your findings with the examinee. What's to hide? Let it all hang out, the good, the bad, and the indifferent. It's "show and tell" time: if test-takers are willing to show their responses, then the tester is obliged to"tell test-takers what scores their responses were transformed into, and what the scores mean.

The Test-Taker's Right to Know *(continued)*

In today's zeitgeist of transparency, accountability, and full disclosure—nothing hidden or kept back—it is the universal and unalienable right that test-takers must have guaranteed for them by those of us who poke around in every nook and cranny of the test-taker's imperfect and sometimes gnarled, wrinkled, and remorseful (or remorseless) psyches, their right to know what bits of information we uncovered about them.

Summary

In this paper, I have hoisted a flag on behalf of the proposition that unless test-takers distinctly specify that their test results not be shown to them, then, testers, test administrators, and organizations—colleges, universities, employing organizations, and other manner of assessors—responsible for administering tests are obliged to share test results with test-takers, test-takers (examinees), and college sophomores who are research participants.

The information provided to test-takers should be simple, clear, unambiguously interpreted, and impervious to misunderstanding. It is unmitigated arrogance, unforgivable imperiousness, despicably authoritarian, and palpably undemocratic for testers to presume to withhold test results from test-takers.

Robert Perloff, Ph.D., is a Distinguished Service Professor Emeritus of Business Administration and of Psychology in the Joseph M. Katz Graduate School of Business at the University of Pittsburgh. Dr. Perloff is a past President of the American Psychological Association and a recipient in 2000 of the American Psychological Foundation's Gold Medal Award for Lifetime Achievement in Psychology in the Public Interest. Email: rperloff@katz.pitt.edu

> There is no science which does not begin by requiring you to believe the incredible.
>
> Henry Adams

Editorial Afterthoughts

Fred B. Bryant, PhD

In his newsletter article, Dr. Perloff advocates a general principle that social and behavioral scientists have largely ignored—namely, the unalienable right of test-takers to receive feedback about their test scores. His article reflects an important and thought-provoking perspective that is bound to create controversy. Dr. Perloff's ideas also raise many questions that his article does not address (given the space limitations of this newsletter), questions that must be answered, however, if we are to implement effectively the policies he advocates. As Dr. Perloff argues, these technicalities should be seen not as insurmountable obstacles, but rather as challenges to be faced and overcome. To stimulate further discussion of his ideas, below we note six specific sets of questions that arise in response to Dr. Perloff's article. Although we offer no answers to these questions, we invite readers to respond with their own ideas.

1. <u>What specific information should be communicated to test-takers?</u> Dr. Perloff makes a strong case for the general rule that test administrators should report "test scores" or "test results" to provide "comprehensible feedback" to test-takers, but he does not specify exactly what form this report should take. Although he exhorts test administrators to avoid "showy psychometrician's jargon" and "let it all hang out, the good, the bad, and the indifferent," this test feedback could take many different forms with varying levels of specificity. For example, imagine a respondent completed the Thematic Apperception Test (TAT) in the context of a clinical interview. Should the test administrator tell the respondent the clinical interpretation of each response to each TAT stimulus card, or merely the overall score and what it indicates? Should clinicians tell test-takers where their scores fall in a normative sample, in terms of the percentage of the general population that scores higher or lower? What sort of qualifications, if any, should clinicians include regarding the reliability and validity of any one test-taker's score? Should test administrators report this

Editorial Afterthoughts *(continued)*

feedback verbally or in writing? Dr. Perloff's brief article is not intended to provide definitive answers to these questions, but rather to stimulate further thought about the issues that must be addressed directly if his ideas are to be implemented effectively.

2. Who is responsible for providing test feedback? Exactly whose responsibility is it to report test feedback to test-takers? Is this always the responsibility of the individual or organization administering the test? Imagine, for example, an industrial-organizational setting in which a business company retains the services of a professional test publishing company to analyze and interpret scores on a pre-employment screening test, and to make hire/no hire recommendations based on test scores. Whose responsibility is it to provide test feedback to test-takers: (a) the organization that administers the test, or (b) the test publisher who developed, validated, analyzed, and interpreted the test scores for the organization? In some cases, the organization may rely entirely on an outside test developer to score and interpret the test, and in these cases, no one in the organization may be qualified to provide test feedback. In such a case, whose responsibility is it—the tester's or the test developer's—to provide test feedback?

3. Who should receive test feedback? In situations in which the test-taker is a minor (e.g., a young child) or a person with impaired capacity to comprehend test feedback, then who should receive the feedback about test scores? Should test administrators be required to provide young children with feedback that these children can understand? Should parents or guardians receive this test feedback instead? Should both children and parents receive test feedback?

4. How can test administrators be sure that test-takers actually understand the test feedback provided? Just because test-takers receive test feedback does not guarantee that they understand it. In general, how can test administrators know that test-takers have obtained an adequate understanding of their test scores? Should test administrators be required to evaluate the level of comprehension of test feedback? Should test administrators be required to continue their efforts to provide feedback, if evidence indicates test-takers do not understand? Is it enough simply to provide test feedback, or must test administrators persist in their communications until each test-taker understands test results to some minimally acceptable degree? And how would one determine whether or not one has established this minimally acceptable degree of comprehension?

5. Are there possible exceptions in specific assessment situations? What about the proprietary rights associated with certain test interpretations (such as the Rorschach, TAT, or employee honesty tests), in which test administrators have at stake the loss of a "trade secret," if they were to reveal their interpretations of test responses to test-takers? Indeed, test developers may have spent a great deal of money to develop test items and scoring protocols that allow valid prediction of on-the-job performance.

Because providing feedback about test-score interpretation could potentially destroy the utility of the test (if test-takers can then coach others on how scores are interpreted), it might be argued that withholding such feedback from test-takers in such cases is a "business necessity." Don't test-takers give up their unalienable right to test feedback when they take such tests in the workplace, particularly if they are informed of this restriction in freedom of information beforehand? Can't an employer make the absence of test feedback to the test-taker a precondition to taking the test, and tell test-takers they are free to walk away and look for another job elsewhere, if they wish not to take the test under such conditions? The key here is that test-takers understand when they apply for the job that they are required to waive their right to test feedback, and that if they do not wish to waive this right, then they are free to look for work elsewhere.

6. Is there a possible compromise position? To extend Dr. Perloff's suggestions, a researcher or other test administrator who wants to withhold test feedback from respondents should be required to add a clause to the informed consent

Editorial Afterthoughts *(continued)*

form notifying potential respondents, before they decide whether to take the test, that they will not receive an interpretation of their test scores; and that if they object to this arrangement, then they are free to terminate participation without penalty. This might offer a reasonable compromise to the rule of always requiring feedback about test scores. It would also make for a more "informed" consent in which potential respondents are told, *before* deciding whether or not to take a test (not afterwards, as is currently done), that they either will or will not receive test feedback. To deny respondents the right to such feedback later, without explicitly informing them of this fact before they consent to take the test, is a breach of APA's ethical principles concerning the need for fully informed consent.

The Nazz, stompin' on a sweet, swingin' beat, goin' down the road. The Nazz talkin' about

How pretty the flowers,
how pretty the hours,
how pretty me, how pretty you,
how pretty he, how pretty she.

Nazz had them pretty eyes. He wanted everybody to see through his eyes so they could see how pretty it was.

Lord Buckley

Selecting Tools For Use in Cross-Cultural Measurement

Carolyn F. Waltz, RN, PhD, FAAN

Cross cultural measurement refers to the use of tools for comparing respondents from different cultures within one country, or comparing one cultural group to another in two or more countries (see Waltz, Strickland, & Lenz, in press). The demand for measures that can be employed cross-culturally has increased dramatically during the last decade, due to: (a) greater awareness of health problems with global impact; (b) the rise in education, research, and clinical collaborations across cultures; and (c) the attention focused on eliminating health disparities among population subgroups in the U.S. and other countries. To ensure that tools have the essential attributes for use across cultures, it is necessary to consider carefully the research concept of interest, specific attributes of the tool, and language translation strategies. Before selecting a tool for cross-cultural use, the following three conditions must prevail in order to minimize threats to reliability and validity.

1. *The concept the instrument is designed to measure should have the same meaning within each of the cultures in which the instrument will be employed.*

To determine whether this condition holds true, researchers should begin by conducting a review of the literature in each culture, to see if the concept is relevant, and if so, to determine the extent to which prior researchers have studied it and how they have measured it. Researchers should then conduct a pretest with respondents representative of the cultures of interest, instructing them to describe the concept, and specific traits or behaviors that characterize it (Serpell, 1993). Once they have selected or constructed a measurement tool, cross-cultural researchers should also examine similarities and differences in response patterns, by comparing factor structures of scores on the instrument across relevant subgroups within the same culture (Wilson, Hutchinson, & Holzemer, 1997) or performing differential item analyses (Allalouf, Hambleton, & Sireci, 1999). If the results reveal

Selecting Tools For Use in Cross-Cultural Measurement (*continued*)

structural equivalence across subgroups within each culture, then scores obtained from the same tool should be comparable when employed in the different subgroups within each culture.

2. *Researchers should use an appropriate translation strategy.*

Assuming measurement instruments are unavailable in the languages of the cultures under investigation, translators should be carefully selected. Foremost, translators should be ethnically and culturally representative of the population among whom the tool will be employed, fluent in both the original (source) language and the target language to which the tool is to be translated, familiar with both cultures, and knowledgeable about the concepts being measured and the tool being used. The probability of a successful translation is greater when: (a) translators use terms that refer to real experiences that are familiar in both cultures; (b) translators determine that each item describes the same phenomenon in both cultures and make necessary modifications when necessary; (c) translators recognize differences in meanings of idioms and take steps to assure these meanings are equivalent in both languages; (d) the original and translated versions of the tool are administered in the same manner, so that valid cross-cultural comparisons can be made; and (e) methods for assessing concepts are comparable between the two cultures. Cross-cultural researchers should pretest translated tools to determine that the resulting responses have distributional properties comparable to those for respondents in the source language version. If discrepancies are found, these differences should be analyzed for mistranslation, and the instrument should be modified if necessary. Before undertaking the actual research project, cross-cultural researchers should fully establish reliability and validity in each culture, and should conduct differential item analysis, for both the original and translated versions of the tool.

3. *Members of different cultures should respond in the same manner to the particular tool to be used.* Discrepancies in how people from different cultures respond to a specific tool may result from: (a) differences in the tendency to respond in ways that are socially desirable; (b) differences in the tendency to respond in an acquiescent style; (c) differences in familiarity with response procedures; (d) differences in physical conditions during administration; (e) noncomparable sampling with respect to such variables as age, gender, and educational background; (f) interviewer effects; or (g) interviewer-respondent effects such as communication problems (Van de Vijver & Poortinga, 1997). Strategies for minimizing such discrepancies include familiarizing respondents with the method of assessment before administering the instrument, and conducting a pilot study to investigate instrument reliability and validity and to explore individual differences among respondents that may influence responses to the instrument.

In summary, before using a tool in cross-cultural research, it is essential to demonstrate that the concept of interest has the same meaning in each culture, use appropriate translation strategies when appropriate, make sure that members of each cultural group respond in a similar manner to the particular tool, and strongly establish reliability and validity of the tool to be used.

References

Allouf, A., Hambleton, R.K., & Sireci, S.G. (1999). Identifying the causes of differences in translated verbal items. *Journal of Education Measurement, 36,* 185-198.

Serpell, R. (1993). *The significance of schooling: Life-journeys in an African society.* Cambridge, UK: Cambridge University Press.

Van de Vijver, F.J.R., & Poortinga, Y.H., (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 12,* 21-29.

Waltz, C. F., Strickland, O.S., & Lenz, E.R. (in press). *Measurement in nursing and health research* (3rd ed.). NY: Springer Publishing Company.

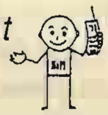
Selecting Tools For Use in Cross-Cultural Measurement (*continued*)

Wilson, H.S., Hutchinson, S., & Holzemer, W.L. (1997). Salvaging quality of life in ethnically diverse patients with advanced HIV/AIDS. *Qualitative Health Research, 75,* 75-97.

Carolyn F. Waltz, RN, PhD, FAAN, is Professor and Director of International Activities and Evaluation in the School of Nursing at the University of Maryland, Baltimore. She is a member of the American Academy of Nursing, and has received 11 American Journal of Nursing Book of the Year Awards since 1981. Dr. Waltz is best known for her work on the measurement of clinical and educational outcomes in nursing. Email: cwaltz@son.umaryland.edu

Remember me,
I'm HaPI at
BMDS!

Have a behavioral measurement question?
Can't find a measurement instrument?



Just call BMDS at 412 687-6850.
We are HaPI to Assist! (IT'S FREE!)

HaPI Advisory Board

Aaron T. Beck, MD
University of Pennsylvania School of Medicine

Timothy C. Brock, PhD
Ohio State University, Psychology

William C. Byham, PhD
Development Dimensions International

Nicholas A. Cummings, PhD
Foundation for Behavior Health

Donald Egolf, PhD
University of Pittsburgh, Communication

Sandra J. Frawley, PhD
Yale University School of Medicine, Medical Informatics

David F. Gillespie, PhD
George Warren Brown School of Social Work, Washington University

Robert C. Like, MD, MS
University of Medicine and Dentistry of New Jersey Robert Wood Johnson Medical School

Joseph D. Matarazzo, PhD
Oregon Health Sciences University

Vickie M. Mays, PhD
University of California at Los Angeles, Psychology

Kay Pool, President
Pool, Heller & Milne, Inc.

Ora Leo Strickland, PhD, RN, FAAN
Emory University Woodruff School of Nursing

Gerald Zaltman, PhD
Harvard University Graduate School of Business Administration

Stephen J. Zyzanski, PhD
Case Western Reserve University School of Medicine

Finnegan's Finagling Factor: That quantity which, when multiplied times, divided by, added to, or subtracted from the answer you got…gives you the answer you should have gotten.

Anonymous

Comparing Measurement in the Natural and Behavioral Sciences: An Interview With Biochemist Duarte Mota de Freitas, Ph.D.

Fred B. Bryant, PhD

Measurement—the quantifying of information—is crucial to all science. It is simply impossible to test and build knowledge efficiently without it. Yet, on the whole, behavioral scientists (by this term I mean also to include researchers in the health and social sciences) seem to know relatively little about how their counterparts in the natural sciences approach the process of measurement. This knowledge gap is regrettable, however, because a better understanding of measurement issues in the natural sciences may well shed light on these same issues in the behavioral sciences. The present article provides some insights into behavioral measurement by highlighting similarities and differences in comparison with measurement in biochemistry.

Clearly, the two sciences have some similarities in measurement concerns. Presumably, natural scientists wrestle with some of the same thorny measurement issues (e.g., instrumentation, reliability, error, validity, and bias) as do behavioral researchers. Yet, there are also fundamental differences in measurement concerns across the two sciences, reflecting the difference between measuring physical properties versus behavioral constructs. What exactly are the similarities and differences in approaches to measurement in the natural sciences, as compared to the behavioral sciences?

To begin to answer this question, I recently interviewed an experienced natural scientist, Duarte Mota de Freitas, who is actively engaged in experimental laboratory research in biochemistry. Having graduated from UCLA in 1984 with a Ph.D. in chemistry, Dr. de Freitas is currently a Professor of Chemistry at Loyola University Chicago, where he specializes in bioinorganic chemistry, an interdisciplinary area at the interface of biochemistry and inorganic chemistry. I have chosen to interview a scientist from this particular discipline because this specialty is representative of the natural sciences.

By way of background, bioinorganic chemists like Dr. de Freitas study the role of metal ions, such as those of lithium, magnesium, platinum, iron, copper, and sodium, in biology and medicine. During the past 18 years, the primary focus of Dr. de Freitas' research has been to seek an understanding at the molecular and cellular levels of how lithium salts work in the treatment of bipolar disorder. He is also applying this fundamental knowledge of lithium biochemistry to clinical questions, such as response to and toxicity of lithium treatment, by using blood samples from patients being treated for bipolar depression.

Below are excerpts from the interview of Dr. de Freitas (**DMF**), in which the author (**FBB**) posed questions about four broad issues related to measurement: (a) instrumentation, (b) error in measurement, (c) establishing measurement validity, and (d) training in measurement and instrument development. Where relevant, I highlight points of similarity and dissimilarity between the natural and behavioral sciences for each of these four measurement-related areas. Finally, I conclude by integrating the main points of convergence and divergence to form a broader perspective on measurement across the two sciences.

Instrumentation

In both behavioral science and biochemistry, scientists often specialize in particular methods of measurement. However, whereas this specialization is due primarily to specialized interests in the behavioral sciences, measurement specialization in the natural sciences is also determined by the prohibitive costs of acquiring additional measurement tools. The following interview excerpt illustrates these points.

FBB: In bioinorganic chemistry, is it ever the case that a new theory will require the development of a new measurement tool that currently does not exist, in order to test hypotheses derived from that theory? If so, can you provide any examples?