Numeracy
Advancing Education in Quantitative Literacy

Volume 2 | Issue 2 Article 3

2009

Quantitative Literacy Assessments: An Introduction to Testing Tests

Dorothy Wallace
Dartmouth College, dorothy.wallace@dartmouth.edu
Kim Rheinlander
Dartmouth College, kimr@math.dartmouth.edu
Steven Woloshin
Dartmouth Medical School, steven.woloshin@dartmouth.edu
Lisa Schwartz
Dartmouth Medical School, lisa.schwartz@dartmouth.edu

Recommended Citation

Wallace, Dorothy, Kim Rheinlander, Steven Woloshin, and Lisa Schwartz. "Quantitative Literacy Assessments: An Introduction to Testing Tests." *Numeracy* 2, Iss. 2 (2009): Article 3. DOI: http://dx.doi.org/10.5038/1936-4660.2.2.3

Quantitative Literacy Assessments: An Introduction to Testing Tests

Abstract

This paper describes how professional evaluators construct assessment instruments that work properly to measure the right thing. Constructing an assessment tool begins with getting feedback from relevant experts on the content of questions. The tool is developed and refined through comparison with existing instruments, focus groups and cognitive interviews. The final instrument is formally tested for content validity, usability, reliability and construct validity through a variety of statistical measures. This process of construction is illustrated by two examples relevant to quantitative literacy: the Medical Data Interpretation Test and the Math Attitudes Survey.

Keywords

Quantitative Literacy, Numeracy, assessment, measure validation, psychometrics, reliability, validity

Introduction: Evaluation Matters

Every instructor is adept at creating assessment instruments. We typically use them as ways to assign grades to students and decide on improvements in our courses. In these situations the audience for our evaluation is ourselves and, as consumers of our own assessment processes, we are particularly easy to please. However, the needs of the community of quantitative literacy (QL) course developers require assessment tools that are convincing to a potentially wide range of stakeholders. At the systemic level, assessments may decide how many students will be required to take an introductory QL course in a single institution or across multiple institutions, such as all of the two-year colleges in a given state. At the national level, assessments are used to show the efficacy of a particular curriculum or text, which then may be adopted by other institutions. Within a given institution, assessments may serve as an important justification for internal funding or as the responsible report to an external foundation.

Assessment is not merely an after-the-fact measurement. It is a force that drives curriculum and funding, generates scholarship, determines interactions among institutions (such as two- to four-year articulation agreements that determine credit transfer policies), and influences policy inside a given institution. The higher the stakes are, the more carefully the instrument should be designed. An assessment designed to convince a broad audience must be constructed far more rigorously than one whose main audience is its author. Furthermore, this rigor is a form of research justifying publication. This paper is designed to guide the construction and testing of such multi-constituency QL assessment instruments.

For smaller projects, an existing assessment tool may have already been created, although few of them have gone through the protocol described in this paper. For large-scale, high-stakes assessment, one may have to construct effective assessment tools from scratch. Even if one plans to hire a professional evaluator to construct these tools, a project director needs a good understanding of the process described in this paper to have productive conversations with the evaluator leading to an assessment tool that satisfies all stakeholders.

Basic Issues in Evaluation

No matter who the audience is, the designers of an assessment instrument must address the same three issues. First, they must be sure they are evaluating the right thing. With a slippery concept such as quantitative literacy, they need to know if their intended audience will agree that their instrument measures some aspect of quantitative literacy (instead of something else that might be related) and covers all the aspects they propose to test. The second question is whether

multiple instruments will be required to capture the information sought. In some cases subject knowledge is easily captured by a multiple-choice test. But other situations might require written pieces, interviews, or surveys. The third question is whether the instruments that have been constructed work properly. This requires a separate testing of the evaluation instrument and calibration of it well in advance of employing it as a diagnostic tool. If a baker is selling bread based on its weight, the buyer wants to be sure the scales are fair. Similarly, if a curriculum developer is promoting a new piece of curriculum based on its success, a potential adopter wants to know how that success was measured and whether one should trust the instrument used. In the case of multiple instruments, it is important to know whether they paint a consistent picture of the outcomes.

Assessment of gains in quantitative literacy presents some special problems. The context of most assessments will not generally permit a truly rigorous design incorporating random samples or clinical trials. It would be a mistake to conclude from this that the usual informal construction of test questions done in a classroom setting is good enough, unless the only goal of the assessment is to see if a particular group of students mastered a very specific set of skills. An assessment designed to test a particular text might require a control group of individuals learning similar material using a different (or no) text. In many cases, one may wish to compare broad interventions at more than one institution, so content items must draw from a wide set of materials and approaches. Quantitative literacy has been described as a habit of mind, so a deep approach to assessment would require subjects to approach new problems not previously mastered in a course. The mind is also an instrument of emotion, so attitudes surrounding quantitative topics may also be an important aspect of evaluating a particular program or course.

To illustrate how assessment instruments are developed and validated, we focus on two examples in quantitative literacy. The Medical Data Interpretation Test[1] (MDIT), developed by Lisa Schwartz and Steven Woloshin, is designed to measure the ability of patients to interpret quantitative medical data (Schwartz, Woloshin and Welch, 2005). Schwartz and Woloshin used this instrument to measure the effectiveness of a text they developed to teach people how to interpret health messages and statistics (Woloshin, Schwartz and Welch, 2008). The evaluation—a randomized controlled trial—compared MDIT scores among participants given their text versus those given a government pamphlet about preventive health practices (but no education about statistics) (Woloshin, Schwartz and Welch, 2007). The MDIT has also been translated and validated in Dutch (Smerecnik and Mesters, 2007). The MDIT is designed to measure patients' ability to interpret quantitative medical data at a particular point in time;

[1] http://www.vaoutcomes.org/research_tools.php (accessed June 25, 2009).

it is not meant to measure changes in patients' ability to do so before and after some intervention.

In contrast, the Math Attitudes Survey[2] (MAS), developed by Jane Korey, is designed to measure improvement in student attitudes towards mathematics. The MAS is not tied to particular mathematical content because it is intended to be used across a broad range of courses. The results of taking the MAS measurement of a collection of different kinds of courses led to Korey's conclusion that, for some students, the standard calculus sequence did not serve as well as interdisciplinary courses with a strong mathematics component (Korey, 2002). These two examples illustrate a range of approaches to creating and testing an assessment instrument.

Table 1. Guidance for developing and testing a new assessment instrument

Step 1. Be clear about what are you trying to measure.

Use your experience and what you learn from a review of the relevant literature to establish the concepts/skills for the assessment to capture.

Get feedback from experts to see if the list of relevant concepts is complete.

Step 2. Develop—and refine—the assessment.

Use or adapt previous questions where possible.

Generate new questions when necessary.

Edit new/revised questions based on feedback the target audience (i.e., people who will be "tested" using the instrument); specifically:

- Check comprehension by conducting focus groups or cognitive interviews to make sure questions and answer choices are meaningful.
- Reduce number of questions by eliminating redundant or poorly understood questions.

Step 3. Formally test the assessment.

Establish content validity

Establish basic attributes—usability (i.e., how often are questions left blank or answered nonsensically), adequate range of difficulty (i.e., proportion correct varies among individual questions), calculate scores* (sum responses and calculate mean, range, standard deviation).

Establish reliability (test-retest repeatability—stability of answers from same people 2 weeks apart (assuming nothing else has changed), internal consistency—calculate Cronbach's alpha of scale[s]).

If needed, establish subscales using factor analysis.

Establish construct validity.

*assuming questions all target the same concepts; this can be formally assessed using factor analysis, a method to see whether the assessment works as a single scale or as several subscales.

[2] http://www.math.dartmouth.edu/~matc/Evaluation/index.html (accessed June 25, 2009).

Table 1 summarizes the discussion presented in this paper. Understanding the steps used to develop/test an assessment instrument will give readers insight into selecting instruments for evaluating quantitative literacy programs.

What Are You Trying to Measure?

Evaluation begins by clarifying the knowledge sought. Most tests constructed for classroom use are designed to measure students against a benchmark or each other. For larger interventions one may wish to measure populations of students against each other (as in a revised course versus a control), or to measure student understanding before and after an intervention, or to measure the effect of completely different kinds of courses against each other. In any case it is useful to see if others have attempted a similar evaluation and to build on their work. A scholarly validation study of a new assessment tool for quantitative literacy of any sort is an asset upon which the entire community of curriculum developers, instructors, and evaluators can draw.

For example, the MDIT test assessed each of the concepts/skills in a curriculum designed to teach people how to make sense of health messages. Schwartz and Woloshin used messages in the form of direct-to-consumer prescription drug advertisements, news stories, and statements a physician might make to a patient. They modeled their approach on the quantitative and document literacy segments of National Adult Literacy Survey which simulates real-world information people routinely encounter (Kirsch 1993). They created hypothetical examples (instead of using existing materials) to ensure that participants had not encountered the information before. It is particularly important not to include examples that are part of the curriculum ultimately being evaluated with the assessment. To avoid problems related to the variability of grading, all the questions were closed ended (i.e., multiple choice). The test was used to measure the performance of two groups of people exposed to different educational materials.

A typical question on the test is:

In a new study, people either took pill X or placebo (a sugar pill). 3% of people taking placebo died; 1% of people taking pill X died. Which statement is correct about how pill X changes the chance of death?

a. Lowers by 66%
b. Lowers by 33%
c. Raises by 33%
d. Raises by 66%

https://scholarcommons.usf.edu/numeracy/vol2/iss2/art3
DOI: http://dx.doi.org/10.5038/1936-4660.2.2.3

The MAS survey was created to assess improvements in desired student attitudes towards mathematics as the result of taking a particular class. The survey questions are scored on a Likert scale (strongly disagree to strongly agree) with the intent to measure pre- and post-intervention scores, so that attitude change over a time period could be measured. Typical statements such as "Mathematics has been an important tool to help me learn other subjects" tested student beliefs about mathematics. The test was used to compare widely different courses, such as "math and music" versus "math and literature." After the first implementation of the MAS at Dartmouth, Korey tailored the MAS survey to other projects that had related but slightly different goals. Many of the core questions were the same, reducing the need to validate the survey as extensively as the first time.

A typical set of items on the survey, scored from 1 to 5 on a Likert scale and measured at least twice, pre- and post-intervention, is:

1. In mathematics I can be creative and discover things for myself.

2. Guessing (conjecturing) is an important part of doing mathematics.

3. Mathematics is essentially an accumulation of facts, rules, and formulas to be memorized and used. (Scored in reverse for comparison to positive statements.)

It is important to get feedback from experts in the field at an early stage in the design of the instrument, to informally establish "content validity"; that is, to make sure the fundamental concepts are included. An external group of experts is likely to notice omission of key concepts (or inclusion of extraneous ones). For example, Schwartz and Woloshin developed items for the MDIT based on their own experience and reviews of the medical literature. Then they sought feedback from individuals with expertise in statistics, education and cognitive psychology to be sure that they had captured the relevant concepts.

Develop—and Refine—the Assessment

When creating a new instrument it is very helpful to find existing, related instruments to see if there are questions that can be reused or adapted (acknowledging the source). Typically it will be necessary to generate new questions. It is very important to test new or adapted questions to ensure that they are understandable and answerable to the target audience. This process can be very time consuming, but it is time extremely well spent since poorly written questions will not yield reliable or useful answers. The basic approach to question writing has been summarized in countless texts (we like Peterson, 2000) and, for the most part, entails attention to detail (one idea per question, complete

and exhaustive answer choices), a good sense of the target audience (reading level, appropriate word choices), and most importantly, common sense (stay simple).

It is extremely useful to check explicitly that the questions and answer choices are understandable to the target audience. This can be done by convening focus groups (Fowler, 2008) typically involving 5−8 people. Some researchers prefer to conduct a series of one-on-one cognitive interviews. With either approach, people are asked to read and critique selected questions and the corresponding answer choices to make sure the language used is appropriate and unambiguous. This testing should replicate the conditions under which the instrument will actually be administered; this means that questions should appear with any corresponding instructions to ensure their clarity. Interview protocols (as opposed to written questionnaires) require additional testing. To ensure that all respondents are answering the same questions the same way, interviewer prompts need to be written out explicitly and tested (and later, interviewers need to be trained not to deviate from the protocol).

Focus groups and cognitive interviews are also a way to begin to reduce the number of questions asked by eliminating redundant or poorly understood questions For the MAS survey, Korey tested a large collection of statements on a convenience sample of subjects in order to eliminate ambiguously interpreted statements and reduce the size of the questionnaire by finding redundancies. It is important to emphasize, however, that some repetition in an assessment is good. To ensure that important concepts are not missed, good questionnaires will ask about key concepts "from a number of angles."

The reading level of subjects also requires careful consideration. Questions designed for doctors could be phrased differently from the same questions intended for a high school student. While reading-level formulas are easy to calculate, they are not nearly as helpful as cognitive interviews where respondents explain what the questions (and their answers) meant to them. When Korey used the MAS survey on Dartmouth courses, she also interviewed a sample of students in the course and conducted focus groups to verify that survey responses reflected the respondents experience with the course in question rather than some extraneous factor.

For richer data about subject experience, one could construct open-ended questions that allow for a wide range of response. Language has to be carefully chosen so answers cannot be yes or no, or “lead” the students to a “desired” response. Here, interview protocols allow for more flexibility than written questionnaires: protocols can be designed to allow the interviewer to gather a much more in-depth picture of things. The extra information that interviews yield can give an instructor not only the answer to whether something is working, but also why it is or isn’t working as expected. The cost of open-ended questions

(whether in a written survey or an interview protocol), of course, comes in analyzing the potentially large amount of unstructured data.

Korey's interviews led to a variety of course revisions that would not have been tried just on the basis of data from the MAS. For example one question that led to productive and useful answers was, "tell me something you learned in this course." Unable to give a yes or no answer to this, students would respond by recalling some topic from the course. Korey coded the interview responses according to whether the students mentioned traditional mathematics or one of the new interventions introduced in the course. Frequent mention of interdisciplinary aspects of these courses by multiple subjects corroborated survey evidence suggesting that attitude changes were a result of the intervention. This evidence was particularly important because many of these courses did not have counterparts in the traditional curriculum that could serve as controls.

Formally Test the Assessment

What follows in this section is a description of statistical approaches that allow an evaluator to have confidence in a set of measures. Technically, tests are not "validated", although the phrase "test validation" is commonly used to describe the psychometric properties to be established. Rather, the inferences made from the assessment are what require validation. These inferences occur in a context, and the tools described below attempt to establish basic relationships between the assessment and the context in which it occurs. Readers interested in a richer and more technical discussion of these methods are referred to Anastasi and Urbini (1997) or Thorndike and Hagen (1969).

Content Validity

A measure has content validity if it covers the relevant domains of the construct in question. Sensibility is a related concept, referring to the common sense impression that the measure "makes sense" to people with content area expertise. To formally assess content validity of the MDIT, Woloshin and Schwartz asked 20 Dartmouth Medical School faculty who teach evidence-based medicine (but were not involved in the study) to complete the data interpretation test and then formally rate its content validity using criteria derived from Feinstein's Index of Sensibility (Feinstein, 1987). Specifically, they were asked to rate the clarity of the test items, how well the data interpretation test covers the important concepts in the domain of critical reading skills, and whether a person scoring poorly on the test would have very limited ability to interpret medical data.

Basic Attributes

Before it is used as a measuring device on its test population, a good tool is first piloted on a sample to see if it performs as intended. It is important to establish the basic statistical properties of the measure, and to compare the results of the measure to any relevant outside indicators. For example (Table 2), to evaluate the MDIT, Schwartz and Woloshin recruited 178 people from advertisements in local newspapers, an outpatient clinic waiting area and a hospital open house.

Table 2. Some psychometric properties established by Schwartz et al. (2005) for the Medical Data Interpretation Test

Property	Assessment
Basic attributes	
Individual items	
usability	item non-response proportion answering item correctly
Aggregate score	Mean, range, standard deviation
Reliability	
Internal consistency reliability (i.e., extent to which items capture a single construct)	Cronbach's alpha (goal is alpha > 0.7)
Test-retest repeatability (i.e., whether answers are the same, assuming no change in underlying ability)	Correlation of scores assessed 2 weeks apart (goal is $r > 0.6$)
Construct validity (measure discriminates between people with different levels of skill)	Comparison of mean scores with hypothesized relationships (i.e., some groups of people will outperform others) people with higher > lower formal education people with higher > lower literacy people with higher > lower quantitative literacy Physician experts > other postgraduates

The usability of MDIT was measured by the item non-response rate. Usability here refers to the subject's ability to make sense out of the test items and to actually respond. If questions are usable, people will answer them (i.e., item

non-response will be low) in a meaningful way (i.e., there will be few nonsensical responses). The range of difficulty for individual questions was assessed and found to be broad: the percentage of correct answers to individual items ranged from 20% to 87%.

Finally responses to all questions were summed for each participant in order to calculate the mean, range and standard deviation of the overall MDIT score. Summing the responses is done because the MDIT was meant to work as a single scale. The individual questions all attempt to get at the same underlying concept: ability to make sense of health statistics. The total score is the single measure summarizing this ability.

Sometimes a set of questions can be sorted into distinct scales. Factor analysis (which is a variant of principal component analysis) can be used to determine, for example, that questions A–F don't work well together as a single scale, but questions A, C and D form one subscale and B, E and F form another. Conducting factor analysis can be useful to help researchers pare down a scale (by identifying redundant or marginally helpful items). Moreover, subscales identified are "purer" and so perform better (see reliability, below). In a recent study, investigators validating a Dutch version of MDIT (Smerecnik and Mesters, 2007) conducted a factor analysis which divided MDIT into four subscales; the usefulness of doing so was unclear, however, since the performance of each subscale was essentially the same as the overall scale. In contrast, factor analysis showed that the MAS really contained four scales (ability, interest, personal growth, utility) present in a list of about 20 survey items.

Reliability

Imagine using a ruler to measure the width of a laptop computer. If the ruler (and strictly speaking the measuring procedure, your eye sight, etc.) is reliable you will get the same result each time. Now imagine the ruler is made of a kind of wood that swells or shrinks a lot depending on the humidity. Repeatedly measuring your laptop will yield very different results on rainy vs. sunny days. This ruler could not be considered reliable. You can't trust that the measurement you get is right.

Conceptually, a single question—or an assessment instrument consisting of many questions—is no different than a ruler. To decide if you can trust its measurements you need to establish its reliability. There are many ways to do this. One common approach is to repeat an assessment and compare the results, typically by calculating the correlation coefficient between the two sets of measurements. This approach establishes "test-retest" reliability.

To assess the test-retest reliability of the MDIT, the test was given to the same people two weeks apart. The two-week time frame is arbitrary but is long enough to avoid practice effects (if the retest is too soon, people just remember

their answers rather than re-take the test) but not too long to have other things happen which change the subjects (if people take other courses or read more, their answers may change not because of the assessment's low reproducibility but because the people themselves have changed). The MDIT is considered reliable because the test-retest correlation coefficient was higher than an arbitrary standard, $r > 0.6$.

When an assessment consists of multiple questions there is another way to assess reliability. If the items in the assessment really get at the same underlying construct, the individual questions should also be internally consistent. If three survey questions are designed to discover the same information, then response to all three should be well correlated—that is, they should demonstrate internal consistency reliability. For example, if you were the food critic for your local newspaper and were reviewing a new restaurant, you would probably want to get at the question "is this a good restaurant" in multiple ways. You wouldn't eat just one breadstick, but you'd try an appetizer, main course, beverages and dessert; and you'd comment on the aesthetics of the place, the service, prices etc. All these elements should work together like a team in revealing how good the restaurant is by providing related but different information.

A statistic, called Cronbach's alpha, formally assesses the idea of "teamness" (i.e., how well a "team of questions" work together). Cronbach's alpha should only be calculated for a single team—so if factor analysis shows multiple subscales, they should be evaluated separately. Conceptually, Cronbach's alpha is simple: it is the ratio of variance shared by the questions divided by the total variance of all the questions. The statistic ranges from 0 (total unrelated items) to 1 (perfectly related information). If alpha is high (approaches 1) the items are all getting at the same thing. If alpha is low, the items are measuring different things (in the restaurant example, a question about parking might not really fit with the others). The desired alpha depends on the intended use of the tool. If the assessment is being used to judge group differences (e.g. how the skills of a group taught one curriculum compare to skills of a group taught another), then a Cronbach's alpha of 0.7 is adequate. But if the assessment is used to judge an individual (e.g., should this person have to take a remedial curriculum?), then a Cronbach's alpha of 0.9 or higher should be required.

In general, more items in a scale increase the scale's reliability. In concept, random bad luck (e.g., the soufflé fell) and random good luck (the oysters were exceptionally succulent) should balance out. The cost of the longer test, of course, is respondent burden (people lose concentration if asked too many items) and needless redundancy. Reassessing alpha—with and without individual items—can help test designers identify questions that can be omitted.

Construct Validity

While necessary, reliability is not sufficient to establish the usefulness of an assessment instrument. The reason is easy to understand: a perfectly reliable ruler could be consistently wrong. To show that an instrument is really useful entails showing that it really measures what it is supposed to measure. That means establishing validity.

There are many ways to establish validity. The easiest to establish, but the least compelling is content validity. As mentioned earlier, this means having individuals with appropriate expertise in the area review the items in the measure to see if they agree that the key concepts are being captured. The most compelling way to establish validity is to demonstrate criterion validity. This entails showing that the assessment agrees with some external gold standard measuring the same thing. For example, a person may claim to be able to guess the number of marbles in a jar. To validate the person's skill you could have them make their guess, then compare it with the actual number of marbles.

Since most of the assessments done in quantitative literacy will not have such a gold standard, a different kind of validity needs to be established. Construct validity is established by showing that a measure behaves in a predictable, logical way. To establish construct validity of the MDIT, Schwartz and Woloshin included secondary measures assessing quantitative literacy, taken from the quantitative and document literacy portions of the National Adult Literacy Survey. Schwartz and Woloshin also measured numeracy by a separate three-item scale used in prior work (Schwartz et al., 1997). They hypothesized that subjects with higher scores on numeracy instruments would perform well on the MDIT, and this was indeed so. Although the other instruments did not measure the specific knowledge that the MDIT tested, they clearly measured related areas (e.g., general quantitative literacy), which should track with medical data interpretation skills.

In addition, Schwartz and Woloshin hypothesized that MDIT scores would be higher for people with more vs. less educational attainment, and for people with more vs. less training in using medical evidence. The fact that each of these hypothesized relationships were observed helped establish the construct validity of the MDIT.

Conclusions

A good evaluation takes time and effort. If stakes are at all high, this is a multi-step process that takes at least six months of work to complete. Two years is a better time frame that allows for multiple iterations. It may take just as long to construct a good instrument as it does to create the curriculum being evaluated.

Depending on who needs to be convinced by the evaluation and how critical the decisions are that will be based on it, all the steps outlined in this paper may be necessary in order to understand and therefore trust an assessment instrument. The process outlined in this paper makes it clear why successful grant proposals build evaluation into the plan from the first day in a comprehensive manner. Reviewers are the first audience that needs to be convinced. At the end of the process, a good assessment tool produces quality data that can form the basis of future scholarship.

References

Anastasi, A., and S. Urbina. 1997. *Psychological Testing*. Upper Saddle River, NJ: Prentice Hall.

Feinstein, A. 1987. *Clinimetrics*. New Haven: Yale University Press.

Fowler, Floyd J. 2002. *Survey Research Methods* (2nd Ed.). Newbury Park, CA: Sage Publications.

Kirsch I.S., A. Jungeblut, L. Jenkins, and A. Kolstad. 1993. *Adult Literacy in America*. Washington, DC: National Center for Education Statistics.

Korey. J. 2002. The calculus trap. *PRIMUS: Problems, Resources, and issues in mathematics undergraduate education* XII(3): 209−218. West Point: United States Military Academy.

Peterson, Robert A. 2000. *Constructing Effective Questionnaires*. Thousand Oaks, CA: Sage Publications.

Schwartz L.M., S. Woloshin, and H.G. Welch. 2005. Can patients interpret health information? An assessment of the medical data interpretation test. *Medical Decision Making* 25: 290−300. http://dx.doi.org/10.1177/0272989X05276860

Schwartz L.M., S. Woloshin, W.C. Black, and H.G. Welch. 1997. The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine* 127: 966−972.

Smerecnik C.M.R, and I. Mesters. 2007. Validating the medical data interpretation test in a Dutch population. *Patient Education and Counseling* 68(3): 287-90. http://dx.doi.org/10.1016/j.pec.2007.06.013

Thorndike, R. L., and E.P. Hagen. 1969. *Measurement and evaluation in psychology and education.* New York: Wiley.

Woloshin S., L.M. Schwartz, and H.G. Welch. 2007. The effectiveness of a primer to help people understand risk: Two randomized trials in distinct populations. *Annals of Internal Medicine* 146: 256−265.

Woloshin S., L.M. Schwartz, and H.G. Welch, 2008. *Know Your Chances: Understanding Health Statistics*. Berkeley: University of California Press.