

Ten Recommendations for Advancing Patient-Centered Outcomes Measurement for Older Persons

Colleen A. McHorney, PhD

The past 50 years have seen great progress in the measurement of patient-based outcomes for older populations. Most of the measures now used were created under the umbrella of a set of assumptions and procedures known as *classical test theory*. A recent alternative for health status assessment is *item response theory*. Item response theory is superior to classical test theory because it can eliminate test dependency and achieve more precise measurement through computerized adaptive testing. Computerized adaptive testing reduces test administration times and

allows varied and precise estimates of ability. Several key challenges must be met before computerized adaptive testing becomes a productive reality. I discuss these challenges for the health assessment of older persons in the form of 10 “Ds”: things we need to deliberate, debate, decide, and do.

Ann Intern Med. 2003;139:403-409.

For author affiliation, see end of text.

www.annals.org

In the past 50 years, great progress has been made in the measurement of patient-based outcomes for older populations. More than 85 tools now measure basic and instrumental activities of daily living (1). Myriad depression measures exist (2), some of which are specific to older persons (3). Numerous measures of cognitive function (4) and almost two dozen generic quality-of-life instruments (5) have been developed, and hundreds of disease-specific instruments exist (6, 7). For patients with cancer alone, more than 75 quality-of-life measures are available. Generic and disease-specific quality-of-life measures have been widely used in older populations. Although such advances in measurement have served many useful purposes, a crucial drawback is that the various assessment tools cannot speak to one another. Data from one study cannot be compared with data from another study that assesses the same trait by using a different set of items. This state of affairs is called *test dependency*. Thus, despite five decades of measurement proliferation, each measure is still a separate yardstick—the measures are on different planes rather than at different spots on an underlying common continuum.

By far, most quality-of-life measures have been created under the umbrella of classical test theory, a set of assumptions and procedures that has been used to develop tests for much of the 20th century (8). Classical test theory has two major shortcomings. The first is test dependency: Different items assessing the same construct cannot be linked to a common metric. The second is group dependency: Different samples yield different item and scale parameter estimates, thereby limiting the usefulness of a given set of item parameters for different samples. These limitations of classical test theory have been discussed extensively (9–12).

In recent years, item response theory (IRT) has been increasingly used in health status assessment (5). Item response theory is a collection of quantitative techniques for constructing, scaling, and equating tests. It can also identify item bias and support computerized adaptive testing (CAT). If its assumptions are met, IRT can overcome some of the limitations of classical test theory by providing 1) item parameters that are invariant with respect to the

sample of examinees and 2) ability parameters that are invariant with respect to the set of items used (9).

Item response theory can achieve more efficient and precise measurement by using CAT (5), which uses a computer to administer items to respondents. Because each “test” is tailored to the unique ability level of each respondent, CAT is adaptive. It functions much like an experienced professor giving an oral examination. If a student cannot answer a question, the professor’s next question is easier. Conversely, if the student answers a question correctly, the next one is more difficult. In oral examinations, the students are asked different questions because their innate abilities and intelligence differ. Yet the erudite professor uses his or her superior knowledge to grade each student. The same process is used in CAT, but the professor is replaced by a computer and the teacher’s judgment is replaced by finely developed computer algorithms. In CAT, each person takes a different version of the test because questions are asked on the basis of the respondent’s previous answers. If a respondent cannot walk one block, the computer knows not to ask whether he can walk a mile. Instead, the computer asks whether he can walk across the room. By use of IRT, all of the different forms of a test can connect to each other on the same yardstick.

Computerized adaptive testing is used in educational assessment (for example, in the Graduate Record Examination, the Graduate Management Admission Test, the Computerized Placement Test, the Scholastic Aptitude Test, and the Test of English as a Foreign Language), professional credentialing (for example, in the National Boards for Professional Teaching Standards Assessment), and professional licensure (for example, by the National Board of Medical Examiners). The U.S. military now uses CAT (for example, in the Armed Services Vocational Aptitude Battery). Public schools are implementing CAT to assess aptitude. Computerized adaptive testing halves test administration time, and it allows ability to be estimated as precisely as desired.

There is great enthusiasm for moving health status assessment away from fixed-length tests and toward a more

Table. Ten “Ds” To Advance Patient-Centered Outcomes Measurement for Older Persons

Definitions of health status and quality of life
Discovery methods
Differential item functioning
Dimensionality
Item difficulty
Item discrimination
Dispute and divisiveness
Dynamic testing
Dangerous and detrimental
Debate

adaptive framework. However, several key challenges must be met before CAT can become a productive reality. I discuss these challenges for the health assessment of older persons in the form of 10 “Ds”: things that we need to deliberate, debate, decide, and do (Table).

DEFINITIONS OF HEALTH STATUS AND QUALITY OF LIFE

We have so many measurement tools in part because of a lack of shared agreement about what constitutes functional status, well-being, and quality of life (13, 14). Conceptual frameworks have generally been insubstantial, often blithely attributing conceptual foundations to the World Health Organization framework of physical, social, and mental health (15). Many instrument developers rely excessively on selecting items from existing measures, thus leading them to overlook exactly what they seek to measure: the patient point of view. As we make the transition from fixed-length, one-size-fits-all tests to more tailored assessments, it will be imperative to develop item banks that are driven by a clear conceptual framework. The strong unidimensionality assumption underlying IRT and item banking, discussed later in this paper, will force this conceptual hand.

DISCOVERY METHODS

Item banks can be assembled through expert opinion or by selecting items from existing instruments. However, many published items are themselves the byproducts of even earlier measures; thus, we have a large stock of items that may have lost their salience over time. We should be true to our stated intent—to measure patient-based outcomes—and use patients as active participants in generating, selecting, and pretesting items. We should generate items for item banks through patient focus groups, semi-structured interviews with patients, and ethnography, among other qualitative discovery methods. Item pretesting should be more patient centered, and it should more often use cognitive-testing methods (16). Through our use of qualitative discovery methods, patients can identify the need for new items to fill in gaps along the functioning and well-being continuum and can help purge redundant items. In short, the development process for item banks

needs to satisfy both patient-based and psychometric litmus tests.

This is not to imply that we need to start from scratch, paying no heed to the work that has been done in the past 50 years. However, deficiencies exist in some areas, most notably in the assessment of higher-order behavioral functioning and role functioning. The U.S. population is living longer, and our measurement tools need to reflect the objective and subjective states and perceptions of health that characterize older populations in the 21st century. We must capitalize on the strengths of measures established to date while recognizing their limitations and compensating for these limitations with primary data collection done using both qualitative and quantitative methods.

DIFFERENTIAL ITEM FUNCTIONING

An item functions differentially if two persons with equal ability (the same amount of the measured state or trait) do not have the same probability of item endorsement. For example, an item on crying in a mental health assessment would function differentially if men and women had the same underlying level of depression yet answered the item differently. Self-report measures of functioning and well-being can fall prey to differential item functioning (DIF) because human beings interpret items within the context of culturally and socially determined mindsets. In educational and psychological testing, item writers purify their items during item pretesting. In health outcomes assessment, instrument developers have ignored DIF or have identified it only long after a measure has been in use. Differential item functioning has been identified in studies of many outcome tools used in older populations, including measures of functional status (1, 17–19), cognitive status (20–23), and mental health (24–29). These studies have identified DIF by age, sex, race, ethnicity, socioeconomic status, language, and nationality. Differential item functioning has been great enough to cause meaningful shifts in group means or case rates when DIF items are removed from the scale (17, 25–28).

Culturally appropriate health outcomes tools (with “culture” defined broadly as comprising sex, age, race, ethnicity, socioeconomic status, geographic location, and language) are crucial to valid health status assessment across diverse groups. If items in a health assessment instrument are biased, detection rates can be overestimated or underestimated. This can lead to erroneous prevalence rates, overtreatment or undertreatment, and overuse or underuse of health services. Better identification and alleviation of DIF will help ensure that current and future assessment tools apply across diverse populations, elderly and nonelderly alike.

Because CAT efficiently tailors items to the respondent’s ability, it requires fewer items than traditional pencil-and-paper tests do (30, 31). Because fewer items are used, it is crucial that the items be unbiased. The conse-

quences of DIF become greater as the number of items used to estimate scores decreases. Thus, item-bank developers will need to assess and eliminate DIF in advance of use of such banks for CAT. If unidimensionality exists and sample size is sufficiently large, IRT provides some powerful methods for assessing DIF (32).

DIMENSIONALITY

Unidimensionality concerns the extent to which a measured attribute is a single unitary trait (for example, functional status) rather than a multidimensional attribute (for example, pain). Unidimensionality is an important underlying assumption for IRT analysis (10). It is a strong requisite for item banks and, thus, CAT (33, 34). Unidimensionality also plays an important role in the study of DIF because multidimensional items can be flagged as having DIF. Thus, for both old and new applications of measures, unidimensionality assessment should be a standardized and comprehensive aspect of instrument development and validation. Unidimensionality should be assessed for item banks before they are calibrated (that is, before item difficulty and discrimination are estimated). Ideally, investigators will use several different methods to assess unidimensionality (35). Principal components analysis, modified parallel analysis, and confirmatory factor analysis are some traditional staples. Other software and methods include DIMTEST (36), NOHARM (37), TESTFACT (38), and newer factor analysis programs (39).

ITEM DIFFICULTY

Easy items are those for which almost everyone endorses the rated behavior, feeling, or attitude. Difficult items are those for which few people endorse the rated behavior, feeling, or attitude. In the general population, an item assessing the ability to bathe would be easy; an item assessing the ability to run 5 miles nonstop would be difficult. For the past decade, the field of health status assessment has been entrenched in a paradigm of psychometric efficiency (5) that has emphasized construction of measures with as few items as possible. For such short-form measures, test developers often select items that are in the middle range of item difficulty and are alternate forms of one another. One major consequence of this protocol is that the end points of the health continuum are poorly defined, yielding skewed and imprecise score distributions (5). The most common reason for score imprecision is the selection of items whose difficulty is incongruent with the ability of the population of interest. Simply put, *ceiling effects*, and the less common *floor effects*, derive from a poor marriage between the difficulty of an item and the ability of the targeted population (6). Ceiling effects occur when easy items are administered to high-ability populations; floor effects are seen when difficult items are administered to low-ability groups.

Problems with precision pertain largely to measures of

physical, role, and social functioning (40). Most items assessing basic and instrumental activities of daily living are at the easy end of the item-difficulty continuum for elderly persons (1, 17, 18, 41–44), and this leads to substantial ceiling effects. Recent work on basic and instrumental activities of daily living (17, 18, 41–44) has shown obvious and manifold redundancies in measuring lower-level functioning and conspicuous gaps in measuring higher-order functioning. The challenge for future advances in measuring physical, role, and social functioning is to more effectively sample and distribute lower-level items while adding items to fill in gaps in the assessment of higher-order functioning, productive activities, executive functioning, leisure exercise, and physical fitness.

Qualitative discovery methods, as discussed earlier in this paper, should be used to glean from older persons themselves facets of contemporary physical, role, and social functioning. A combination of focus-group, diary, and time-use methods might yield useful insights into what types of basic, intermediate, and advanced activities are done regularly, as well as what types of activities have been abandoned and in what sequence. Qualitative methods could also be used to discover how older persons adapt to occult or incipient disability. Functional assessment might also benefit from the development of rating scales that tap functional compensation (45) rather than difficulty or dependence per se.

ITEM DISCRIMINATION

Discrimination is an item's ability to distinguish among persons who have different levels of the trait being measured. High-discrimination items yield more information than low-discrimination items (46). To be useful, an item bank must contain items that differ in difficulty. However, the bank must also have high-discrimination items to differentiate among persons close together in ability. One challenge for compilers of items banks is writing high-discrimination items and eliminating low-discrimination items. Ambiguity can degrade an item's discrimination (44). For example, ambiguity is increased and validity is compromised if respondents confuse functional capacity with functional performance (47). In one study (44), high-discriminating functional status items (for example, items asking about putting underclothes on, moving between rooms, taking pants off, getting into bed) were almost behavioral measures: They targeted daily activities in a way that was specific, explicit, and unequivocal. In short, they were questions that respondents could understand (because they were simple and concrete) and answer with respect to their range of function (because they were in the realm of daily experience). Item writing may be improved by scrutiny of low- and high-discrimination items (48, 49). In addition, adherence to conventional item-writing standards (for example, writing items that can be interpreted in only one way; using clear, simple, direct language; and avoiding

multiple attributions in a given item) may go far toward improving item discrimination.

DISPUTE AND DIVISIVENESS

In health status and outcomes assessment, two types of IRT practitioners have emerged: those who adhere solely to the Rasch (one-parameter) model and those who use whatever model (one-, two-, or three-parameter) best fits the data. The Rasch model estimates only item difficulty; items are assumed to have equal discrimination. The two-parameter model estimates both difficulty and discrimination. In studies of both physical and mental health status (1, 17, 20, 24, 44, 50, 51), item discrimination has been shown to vary greatly. Because items often have unequal discrimination, Rasch researchers could needlessly remove potentially informative items, perhaps compromising the content validity of the resulting item bank. Imbalanced content of item banks could have negative consequences for CAT applications.

It is time for the field to move beyond polemics—or religion, as some put it (52)—and toward head-to-head comparisons of different models. It may make little substantive, clinical, or policy difference if the one- or the two-parameter model is used to simply calibrate items. However, as the field moves toward item banking and CAT-based estimation of individual- and group-level ability, we need to understand the statistical and substantive consequences of using different models. For example, does one need more items in a bank if item discrimination is not modeled? Does adherence to the Rasch model negatively affect the spread of item difficulty parameters across the ability distribution? Is the standard error of individual ability estimates greater without an estimated discrimination parameter? Researchers in educational and psychological measurement have long moved beyond iconoclastic debate. We in outcomes assessment need to do the same, and we need to do so while publicly debating the relative merits of CAT.

DYNAMIC TESTING

Although IRT has been increasingly used in health status assessment, it has most often been used for simple item calibration. Fewer studies have appropriately used IRT for test equating and linking (1, 44, 53), which are cornerstones of item banking and CAT (otherwise known as dynamic testing). Much needs to be done before developers and users can confidently take the next step toward CAT. More work is needed on how many items should be in an item bank. Some researchers argue for as few as 30 items (46); others assert that hundreds of items are required (52). What should be the range of difficulty and discrimination estimates of banked items? Equally important, what should be the distribution of difficulty and discrimination estimates over theta (the unobserved ability distribution)? Polytomous items are known to be more

informative than dichotomous items (46, 54). Should we discourage banking with dichotomous items? We need more profound knowledge of the substantive and statistical implications of using CAT with the Rasch model and with the two- or three-parameter model.

Practically speaking, much needs to be done to establish the infrastructure for CAT. We need to develop theoretically driven item banks that are constructed, at least in part, from the patient point of view; meet unidimensionality requirements; and are free of DIF. We need more experience with different equating and linking designs, which are crucial to item banks and CAT (55). We need more research on item selection (46, 52) and trait estimation (54) procedures. Work on optimal stopping rules for different applications (for example, individual-level versus group-level ability estimation, or high-stakes versus low-stakes testing) and populations (for example, socioeconomically vulnerable groups and those in the digital divide) needs to be done. We will need to pay attention to item and test security and item theft (52) if item banking and CAT becomes proprietary.

DANGEROUS AND DETRIMENTAL

Throughout the history of aptitude and psychological testing, it has been assumed that testing is beneficial and that it is desirable to expand testing and then later assess the consequences of expansion. However, the U.S. experience with educational assessment has shown that testing is not always benign (56). We should not assume that CAT-based health status assessment will be beneficial across the board. Health services research generally, and health status and outcomes assessment specifically, has witnessed an increase in proprietary markets (and marketing) for the products of scientific labor (57). Any organization that owns the rights to a product, such as a measurement tool or system, may be more interested in promoting it than in disclosing its shortcomings as well as its virtues. As Shapiro argues, “the proprietary appropriation of instruments in health services research . . . involves introduction of the profit motive into the very marrow of the field” (57). Before CAT-based health and outcomes assessment becomes the rule, both developers and users need to seriously consider the following questions: 1) Will CAT be devoted to commercialism, promotion, profit, and institutional self-interest or to public service and disinterested research? 2) Will it simply market its tools, or will it ponder the merits, uses, and effectiveness of those tools? 3) Will it expand testing, or will it promote basic science and public good? 4) Will it promote secrecy or science? 5) Will it be monopolistic or cooperative? 6) Will it overpromote test strengths while minimizing weaknesses? 7) Will it suppress evidence that does not support testing and marketing? 8) Will it be unresponsive to its constituency (researchers, policymakers, funders, payers)? 9) To whom will it be accountable and

for what? 10) Will it lead to less objective and scientific research?

“Caveat emptor” has always guided consumers in the marketplace. It may soon guide potential users of CAT-based health status assessment if we remain on the path toward entrepreneurial and proprietary development. A for-profit CAT-based assessment industry could lead to a loss of openness between entrepreneurs and their constituency. Because CAT does not come in a standardized form, users cannot visually see it as one sees other instruments. In essence, CAT is a black box. Its virtue—a large repository of items—is also its downside— anonymity of items and procedures. How will the constituency know how many items are in the bank, how the items were derived, whether DIF has been assessed and eliminated, whether unidimensionality holds, the number of item parameters that have been calibrated, the number of item parameters that are actually used, the size and representativeness of the calibration sample, the equating procedures, and the item selection and stopping rules? Users should demand and receive complete and comprehensive disclosure on these, and other, crucial questions.

DEBATE

The intellectual and technical infrastructure for item banking and CAT under IRT is at hand. What is unclear is whether it is desirable, intellectually or scientifically, for health status and outcomes assessment to move in this direction without serious debate about the merits and shortcomings of this approach. Myriad tools exist for measuring the function and well-being of older persons. We have seen both advances and redundancies among these measures. The same, of course, could easily apply in the future to item banks, with investigators arguing about the extent to which “my bank is better than yours.” Further, the privatization (57) of CAT-based assessment enterprises could stifle scientific progress in measurement and assessment (and its many applications) because 1) methods may not be in the public domain, 2) results may not be in the public domain, and 3) replication may be frustrated, if not impeded entirely.

An item bank is the result of the work of hundreds of persons over decades. For the fairest and most productive use, health banks should be in the public domain since their constituent parts were most often developed with public monies. Health banks could reside with the National Center for Health Statistics, the Agency for Healthcare Research and Quality, or a nonprofit organization similar to the Educational Testing Service. We need to publicly debate the best way to administer and manage item banks and CAT.

In health status and outcomes assessment, which has more than two dozen generic measures (5) and hundreds of disease-specific measures (6, 7), dialogue among measurement specialists has sometimes resembled a childhood fight

with claims of “my tool is better than yours.” But at least these discussions have taken place within the context of peer-reviewed science. If measurement research moves from academia to the private sector, we may continue to hear the polemic of “my item bank is better than yours,” but without the safeguard of peer review. We have already witnessed some disadvantages of proprietary measurement tools, with developers eschewing peer-reviewed publication in favor of profit-yielding and non-peer-reviewed user’s manuals. In the absence of peer review, how will the constituency be able to differentiate between scientific truth and advertising? As health status and outcomes assessment begins to move toward the development and calibration of item banks and CAT (5), earnest and critical thought is needed about the extent to which the profit motive will corrupt developments that are on the rise.

CONCLUSION

In the past 50 years, we have accomplished much in the measurement of the function and well-being of older persons. Measurement specialists are at the threshold of a paradigm shift (5) away from classical test methods and toward broader use of IRT methods. There is reason for both excitement and caution as IRT methods are used for test construction, scaling, and score equating, as well as for the identification of DIF and the implementation of CAT. It may be desirable for stakeholders—methodologists, users, policymakers, payers, and funders—to reach consensus about the relative merits of the various courses that patient-based measurement could follow before any one road is definitively taken. I offer the 10 Ds as a platform for informing and stimulating discussion about how and where measurement advances might proceed in the years ahead.

From Richard L. Roudebush Veterans Affairs Medical Center; Indiana University School of Medicine; Regenstrief Institute; and Indiana University Center for Aging Research, Indianapolis, Indiana.

Grant Support: In part by the Department of Veterans Affairs (grants RR&D C-2488-R and RCS 02-066-1).

Potential Financial Conflicts of Interest: None disclosed.

Requests for Single Reprints: Colleen A. McHorney, PhD, Regenstrief Institute, RHC 6th Floor, 1050 Wishard Boulevard, Indianapolis, IN 46202; e-mail, cmchorney@regenstrief.org.

References

1. McHorney CA. Use of item response theory to link 3 modules of functional status items from the Asset and Health Dynamics Among the Oldest Old study. *Arch Phys Med Rehabil*. 2002;83:383-94. [PMID: 11887121]
2. Task Force for the Handbook of Psychiatric Measures. *Handbook of Psychiatric Measures*. Washington, DC: American Psychiatric Assoc; 2000.
3. Yesavage JA, Brink TL, Rose TL, Lum O, Huang V, Adey M, et al. Development and validation of a geriatric depression screening scale: a preliminary report. *J Psychiatr Res*. 1982;17:37-49. [PMID: 7183759]
4. Lorentz WJ, Scanlan JM, Borson S. Brief screening tests for dementia. *Can*

- J Psychiatry. 2002;47:723-33. [PMID: 12420650]
5. **McHorney CA.** Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. *Ann Intern Med.* 1997;127:743-50. [PMID: 9382391]
 6. **McHorney CA.** Health status assessment methods for adults: past accomplishments and future challenges. *Annu Rev Public Health.* 1999;20:309-35. [PMID: 10352861]
 7. **Bowling A.** *Measuring Disease: A Review of Disease-Specific Quality of Life Measurement Scales.* Buckingham: Open Univ Pr; 2001.
 8. **Gulliksen H.** *Theory of Mental Tests.* New York: J Wiley; 1950.
 9. **Hambleton R, Swaminathan H.** *Item Response Theory: Principles and Applications.* Boston: Kluwer Nijhoff; 1985.
 10. **Hambleton R.** Principles and Selected Applications of Item Response Theory. In: Linn R, ed. *Educational Measurement.* Phoenix: Oryx Pr; 1993:147-200.
 11. **Hambleton RK, Jones RW.** Comparison of classical test theory and item response theory and their applications to test development. *Education Measurement: Issues and Practice.* 1993;38-47.
 12. **Hambleton R, Slater S.** Item response theory models and testing practices: current international status and future directions. *European Journal of Psychological Assessment.* 1997;13:21-8.
 13. **Gill TM, Feinstein AR.** A critical appraisal of the quality of quality-of-life measurements. *JAMA.* 1994;272:619-26. [PMID: 7726894]
 14. **Leplege A, Hunt S.** The problem of quality of life in medicine. *JAMA.* 1997;278:47-50. [PMID: 9207338]
 15. *Constitution of the World Health Organization.* Geneva, Switzerland: World Health Organization; 1948.
 16. **Sudman S, Bradburn N, Schwarz N.** *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology.* San Francisco: Jossey-Bass; 1996.
 17. **Teresi JA, Cross PS, Golden RR.** Some applications of latent trait analysis to the measurement of ADL. *J Gerontol.* 1989;44:S196-204. [PMID: 2768780]
 18. **Spector WD, Fleishman JA.** Combining activities of daily living with instrumental activities of daily living to measure functional disability. *J Gerontol B Psychol Sci Soc Sci.* 1998;53:S46-57. [PMID: 9469179]
 19. **Fleishman JA, Spector WD, Altman BM.** Impact of differential item functioning on age and gender differences in functional disability. *J Gerontol B Psychol Sci Soc Sci.* 2002;57:S275-84. [PMID: 12198107]
 20. **Teresi JA, Golden RR, Cross P, Gurland B, Kleinman M, Wilder D.** Item bias in cognitive screening measures: comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. *J Clin Epidemiol.* 1995;48:473-83. [PMID: 7722601]
 21. **Teresi J, Kleinman M, Ocepek-Welikson K, Ramirez M, Gurland B, Lantigua R, et al.** Applications of item response theory to the examination of the psychometric properties and differential item functioning of the comprehensive assessment and referral evaluation dementia diagnostic scale among samples of Latino, African American, and white non-Latino elderly. *Research on Aging.* 2000;22:738-73.
 22. **Teresi J, Holmes D, Ramirez M, Gurland B, Lantigua R.** Performance of cognitive tests among different racial/ethnic and education groups: findings of differential item functioning and possible item bias. *Journal of Mental Health and Aging.* 2001;7:79-89.
 23. **Jones RN, Gallo JJ.** Education and sex differences in the mini-mental state examination: effects of differential item functioning. *J Gerontol B Psychol Sci Soc Sci.* 2002;57:P548-58. [PMID: 12426438]
 24. **Schaeffer N.** An Application of Item Response Theory to the Measurement of Depression. In: Clogg C, ed. *Sociological Methodology.* San Francisco: Jossey-Bass; 1988:271-307.
 25. **Stommel M, Given BA, Given CW, Kalaian HA, Schulz R, McCorkle R.** Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). *Psychiatry Res.* 1993;49:239-50. [PMID: 8177918]
 26. **Christensen H, Jorm AF, Mackinnon AJ, Korten AE, Jacomb PA, Henderson AS, et al.** Age differences in depression and anxiety symptoms: a structural equation modelling analysis of data from a general population sample. *Psychol Med.* 1999;29:325-39. [PMID: 10218924]
 27. **Grayson DA, Mackinnon A, Jorm AF, Creasey H, Broe GA.** Item bias in the Center for Epidemiologic Studies Depression Scale: effects of physical disorders and disability in an elderly community sample. *J Gerontol B Psychol Sci Soc Sci.* 2000;55:P273-82. [PMID: 10985292]
 28. **Azocar F, Arean P, Miranda J, Munoz RF.** Differential item functioning in a Spanish translation of the Beck Depression Inventory. *J Clin Psychol.* 2001;57:355-65. [PMID: 11241365]
 29. **Kim Y, Pilkonis PA, Frank E, Thase ME, Reynolds CF.** Differential functioning of the Beck depression inventory in late-life patients: use of item response theory. *Psychol Aging.* 2002;17:379-91. [PMID: 12243380]
 30. **De Ayala R.** A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. *Educational and Psychological Measurement.* 1989;49:789-805.
 31. **Eggen T, Straetmans G.** Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement.* 2000;60:713-34.
 32. **Holland P, Wainer H.** *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates; 1993.
 33. **Green B, Bock R, Humphreys L, Linn R, Reckase M.** Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement.* 1984;21:347-60.
 34. **Roznowski M, Tucker L, Humphreys L.** Three approaches to determining the dimensionality of binary items. *Applied Psychological Measurement.* 1991;15:109-27.
 35. **Drasgow F, Lissak R.** Modified parallel analysis: a procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology.* 1983;68:363-73.
 36. **Stout W.** A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika.* 1987;52:589-617.
 37. **Fraser C, McDonald R.** NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research.* 1988;23:267-9.
 38. **Wood R, Bock D, Gibbons R, Schilling S, Muraki E, Wilson D.** TESTFACT. Lincolnwood, IL: Scientific Software International; 2003.
 39. **Hattie J.** Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement.* 1985;9:139-64.
 40. **McHorney CA, Ware JE Jr, Rogers W, Raczek AE, Lu JF.** The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts. Results from the Medical Outcomes Study. *Med Care.* 1992;30:MS253-65. [PMID: 1583937]
 41. **Haley SM, McHorney CA, Ware JE Jr.** Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. *J Clin Epidemiol.* 1994;47:671-84. [PMID: 7722580]
 42. **Jette AM.** How measurement techniques influence estimates of disability in older populations. *Soc Sci Med.* 1994;38:937-42. [PMID: 8202742]
 43. **Kempen GI, Myers AM, Powell LE.** Hierarchical structure in ADL and IADL: analytical assumptions and applications for clinicians and researchers. *J Clin Epidemiol.* 1995;48:1299-305. [PMID: 7490592]
 44. **McHorney CA, Cohen AS.** Equating health status measures with item response theory: illustrations with functional status items. *Med Care.* 2000;38:II43-59. [PMID: 10982089]
 45. **Hazuda HP, Gerety MB, Lee S, Mulrow CD, Lichtenstein MJ.** Measuring subclinical disability in older Mexican Americans. *Psychosom Med.* 2002;64:520-30. [PMID: 12021426]
 46. **Dodd B, Koch W, De Ayala R.** Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement.* 1989;13:129-43.
 47. **Keller D, Kovar M, Jobe J, Branch L.** Problems eliciting elders' reports of functional status. *Journal of Aging and Health.* 1993;5:306-18.
 48. **Steinberg L, Thissen D.** Item Response Theory in Personality Research. In: Shrout P, Fiske S, eds. *Personality Research Methods.* Hillsdale, NJ: Lawrence Erlbaum Associates; 1995:161-81.
 49. **Gray-Little B, Williams V, Hancock T.** An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin.* 1997;23:443-51.
 50. **Gibbons RD, Clark DC, VonAmmon Cavanaugh S, Davis JM.** Application of modern psychometric theory in psychiatric research. *J Psychiatr Res.* 1985;19:43-55. [PMID: 3989737]
 51. **Steinberg L.** Context and serial-order effects in personality measurement:

limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*. 1994;66:341-9.

52. **Wise SL, Kingsbury GG.** Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica*. 2000;21:135-55.

53. **Orlando M, Sherbourne CD, Thissen D.** Summed-score linking using item response theory: application to depression measurement. *Psychol Assess*. 2000;12:354-9. [PMID: 11021160]

54. **Dodd B, De Ayala R, Koch W.** Computerized adaptive testing with poly-

tomous items. *Applied Psychological Measurement*. 1995;19:5-22.

55. **Dorans N.** Scaling and equating. In: Wainer H, ed. *Computerized Adaptive Testing: A Primer*. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.

56. **Lemann N.** *The Big Test: The Secret History of the American Meritocracy*. New York: Farrar, Straus & Giroux; 1999.

57. **Shapiro MF.** Is the spirit of capitalism undermining the ethics of health services research? *Health Serv Res*. 1994;28:661-72; discussion 678-87. [PMID: 8113051]