EVALUATING THE STATISTICAL SIGNIFICANCE OF HEALTH-RELATED QUALITY-OF-LIFE CHANGE IN INDIVIDUAL PATIENTS

RON D. HAYS
MARC BRODSKY
M. FRANCIS JOHNSTON
KAREN L. SPRITZER
KA-KIT HUI
UCLA Department of Medicine

Assessing individual change is feasible and potentially useful in clinical practice. This article provides an overview of the evaluation of statistically significant change in health-related quality of life (HRQOL) for individual patients. We review the standard error of measurement, standard error of prediction, and reliable change indices using a sample of 54 patients receiving care at the UCLA Center for East-West Medicine. The largest amount of change necessary for statistical significance was found for the reliable change index and the smallest change was needed for the standard error of measurement. The amount of change required for statistical significance was intermediate for the standard error of prediction. The median kappa for classifying change (declined, stayed the same, improved) by different indices was .82, indicating a high level of agreement. Future research is needed to determine if one index is most appropriate for evaluating the significance of individual change.

Keywords: *individual change; health-related quality of life; statistical significance; reliable change; standard error of measurement; standard error of prediction*

AUTHORS' NOTE: Address correspondence concerning this article to Ron D. Hays, Ph.D., UCLA Division of General Internal Medicine and Health Services Research, 911 Broxton Plaza, Room 110, Los Angeles, CA 90095-1736; e-mail: drhays@ucla.edu

EVALUATION & THE HEALTH PROFESSIONS, Vol. 28 No. 2, June 2005 160-171
DOI: 10.1177/0163278705275339

McHorney and Tarlov (1995) remarked, "Clinicians who use health status measures in their practice on an individual-patient level should report their experiences and findings to help move the field forward" (p. 301). Clinicians can use information about the statistical significance of individual change to help evaluate the health of their patients and to revise therapy as needed. Researchers can use data about the significance of individual change to supplement information about group mean change with the number of people who improve, stay the same, or decline in health-related quality of life (HRQOL). However, limited response to McHorney and Tarlov's call to action has occurred in the ensuing decade.

This article focuses on the evaluation of change in HRQOL for individual patients. We begin by providing a brief overview of statistical significance in a group of patients over time. Then, we describe methods of evaluating the statistical significance of individual change. A sample of patients receiving care at an integrative medicine clinic in Los Angeles was used for illustration because statistically significant group change in HRQOL was observed from baseline to approximately 6 weeks after treatment began.

It is widely acknowledged that a group difference can be small even if it is statistically significant when the sample size is large. Hence, it is important to evaluate whether statistically significant group change is also change that is minimally important or clinically significant (Bauer, Lambert, & Nielsen, 2004; Farivar, Liu, & Hays, 2004). With respect to individual change, one may also be interested in more than simply whether change was statistically significant. We note additional considerations in the Discussion section of the article.

METHOD

SAMPLE AND SURVEY

The UCLA Center for East-West Medicine is an organized unit within the Department of Medicine and has more than 12,000 patient visits a year. The clinical staff, trained in biomedicine and traditional Chinese medical methods, offers comprehensive care with emphasis on health promotion, disease prevention, treatment, and rehabilitation. UCLA physicians refer many of the patients to the Center after they have failed medical treatments that include potent medications

and surgeries. Patients suffer from a variety of illnesses, primarily chronic pain conditions such as neck and back pain, headaches, fibromyalgia, sports and occupational-related injuries, and cancer-related symptoms. Treatment plans are developed to meet each patient's needs. Therapeutic techniques include patient education, medication adjustments, trigger point injections, acupuncture, acupressure, therapeutic massage, dietary and herbal counseling, and mind-body exercises (Hui, Zylowska, Hui, Yu, & Li, 2002).

We estimated the indices of change discussed in this article in a consecutive sample of 54 patients receiving care at the UCLA Center for East-West Medicine. The average age of these individuals was 56 years, 84% were White, and 58% were women. Patients self-administered the paper-and-pencil version of the SF-36® Version 2 health survey at baseline and approximately 6 weeks after therapy began.

The SF-36 includes multi-item scales that evaluate physical and mental health: physical functioning (10 items), role limitations caused by physical health problems (4 items), pain (2 items), general health perceptions (5 items), energy and/or fatigue (4 items), social functioning (2 items), role limitations caused by emotional health problems (3 items), and emotional well-being (5 items). Furthermore, the SF-36 has physical and mental health summary scores that are weighted combinations of the eight scale scores (Ware, Kosinski, & Dewey, 2000). The eight scale scores and the two summary scores are scored on a *T*-score metric, where the mean is fixed at 50 and standard deviation is 10 in the U.S. general population.

ANALYTIC METHODS

We compute Student's (W. S. Gossett) *t* test to assess the significance of change for the sample. The within-group or dependent *t* test is computed as the average difference divided by the standard error (*SE*) of the difference (Table 1).

Internal consistency reliability is estimated for the eight SF-36 scales using a two-way mixed (fixed item effect) effects model by subtracting mean square error (interaction between respondents and items) from the mean square between, and dividing by the mean square between (Cronbach, 1951). Reliabilities are estimated for the SF-36 physical component summary (PCS) and mental component summary (MCS) scores using Mosier's (1943) formula (Table 1).

TABLE 1
Formula for Evaluating Significance of Differences

Statistic	*Formula*
Student's within-group *t* test	$X_D/(SD_d/N^{½})$
Internal consistency reliability	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS}}$
Reliability of a weighted composite	$Mosier = 1 - \frac{\sum(w^2_j)(S^2_j) - \sum(ww^2_j)(S^2_j)(\alpha_j)}{\sum(w^2_j)(S^2_j) + 2\sum(w_j)(w_k)(S_j)(S_k)(r_{jk})}$
Standard error of measurement (*SEM*)	$SD_b \times (1 - \text{reliability})^{½}$
Standard error of prediction (90% CI)	$1.64 * SD_b \times (1 - \text{reliability}^2)^{½}$
Standard error of prediction (95% CI)	$1.96 * SD_b \times (1 - \text{reliability}^2)^{½}$
Estimated true score	Mean + reliability (score – mean)
Reliable change index	$(X_2 - X_1)/SD_b(2 \times [1 - \text{reliability}])^{½}$ or $X_2 - X_1/\sqrt{2}SEM$.

NOTE: X_D = mean difference; SD_d= standard deviation of difference; MS_{BMS} = mean square between; MS_{EMS} = mean square for interaction between respondents and items; w_j = weight given to scale j; w_k = weight given to scale k; S_j= standard deviation of scale j; S_k = standard deviation of scale k; j = reliability of scale j; r_{jk} = correlation between scale j and scale k; SD_b = standard deviation at baseline.

Various approaches have been employed to assess the statistical significance of individual change (Table 1) including the standard error of measurement (*SEM*), the standard error of prediction (SE_P), and the reliable change index (RCI). A consensus about which approach is best does not exist. Ware, Bayliss, Rogers, Kosinski, and Tarlov (1996) categorized people in the Medical Outcomes Study as not changing (follow-up score was no different than what would be expected by chance from the baseline score), improved more than expected by chance, and declined more than expected by chance using two *SEM*s. If an individual's score at Time 2 fell within the confidence interval (CI) for the Time 1 score, then the change was deemed not significantly different.

Alternatives to the *SEM* include the SE_P and the RCI. The SE_P is equal to: a × *SD* × SQR (square root) (1 – reliability^2), where a = 1.64 and 1.96 for 90% and 95% CIs, respectively. The *SE* of the difference for an individual can be estimated by multiplying the *SEM* by the

TABLE 2
SF-36 Scale Scores at Baseline and Follow-Up for 54 Patients

	Reliability	*Time 1* M(SD)	*Time 2* M(SD)	*Change*	t *test*	*Probability*
Physical functioning	.94	42.3 (12.6)	44.0 (11.8)	1.7	2.38	.0208
Role–physical	.93	37.9 (11.6)	42.0 (10.3)	4.1	3.81	.0004
Bodily pain	.87	36.8 (10.3)	40.3 (9.0)	3.6	2.59	.0125
General health	.83	43.3 (11.5)	45.7 (10.3)	2.4	2.86	.0061
Energy	.77	43.0 (9.6)	48.0 (8.9)	5.1	4.33	.0001
Social functioning	.85	38.3 (12.9)	43.0 (10.4)	4.7	3.51	.0009
Role–emotional	.94	39.0 (13.7)	40.5 (11.4)	1.5	.96	.3400
Emotional well-being	.79	41.1 (10.6)	45.4 (9.4)	4.3	3.20	.0023
Physical component summary (PCS)	.94	40.6 (11.5)	43.3 (10.6)	2.8	3.23	.0021
Mental component summary (MCS)	.93	40.8 (13.2)	44.7 (10.8)	3.9	2.82	.0067

square root of 2 (Ferguson, Robinson, & Splaine, 2002; Jacobson, Follette, & Revenstorf, 1984). The RCI is a *z* test of change between baseline and follow-up, divided by the *SE* of the difference. If the RCI is 1.96 or larger, change is considered reliable (statistically significant at $p < .05$). Some have suggested that the RCI should use the *SD* of change instead of the *SD* at baseline as a basis for the *SE* of the difference (Collie, Maruff, McStephen, & Darby, 2003).

The approach used by Ware et al. (1996) and others has been criticized for not taking into account regression toward the mean (Bauer et al., 2004; Dudek, 1979; Hsu, 1989); that is, forming CIs around the estimated true score (ETS) rather than the observed score is recommended (Bland & Altman, 1994).

RESULTS

Table 2 presents SF-36 scale scores at baseline and follow-up for the sample, with the physical health scales toward the top and the mental health scores at the bottom. Internal consistency reliabilities for the SF-36 scales in this sample ranged from .77 (energy and/or fatigue) to .94 (physical functioning). The estimated reliabilities of the SF-36 PCS and MCS were .94 and .93, respectively. Mean changes in SF-36

scores ranged from 1.5 (role–emotional) to 5.1 (energy). Statistically significant ($p > .05$) group mean improvement over time was found for all of the SF-36 scales except for role limitations due to emotional problems (Table 2). The magnitude of the significant changes (effect size) ranged from .13 (physical functioning) to .53 (energy) of the baseline standard deviation.

Information relevant to estimating the statistical significance of change for a single individual in the sample of 54 patients is presented in Table 3. To increase the likelihood of demonstrating significant individual change on some SF-36 scales, we selected the case that had the largest observed change on the SF-36 PCS (i.e., PCS change of 18.3). This person's Time 2 and Time 1 observed scores, along with his or her estimated true score at Time 1, are shown. Also provided are the *SEM*, 95% CI based on the *SEM* (estimated baseline true score plus or minus 1.96 × *SEM*), and 90 and 95% CIs based on the SE_P. Each CI is centered on the baseline ETS. We also provide the difference between the Time 1 and Time 2 scores, the *SE* of the difference, and the RCI.

The observed scores for this person on role–physical, bodily pain, general health perceptions, social functioning, role–emotional, PCS, and MCS significantly exceeded the baseline score (Table 3) on each index of change (95% CI around the *SEM*, 90% CI around the SE_P, 95% CI around the SE_P, and RCI). The magnitude of observed change from Time 1 to Time 2 for these statistically significant changes ranged from 0.91 *SD* (MCS) to 2.55 *SD* (role–emotional). In contrast, the magnitude of change for the physical functioning and energy scales were .13 *SD* and .64 *SD*, respectively, and these changes were not statistically significant.

Table 4 provides the amount of change from the observed baseline and follow-up scores necessary to be classified as significant for the four indices of change for the example case with the largest change on the PCS and for the mean of all 54 cases. On average, the smallest to largest amount of change was required for the *SEM*, followed by the 90% SE_P, the 95% SE_P, and then the RCI. Accordingly, 11%, 8%, 6%, and 5% of the cases significantly declined, and 27%, 22%, 18%, and 18% of the cases significantly improved (averaged across SF-36 scales) according to the *SEM*, 90% SE_P, 95% SE_P, and RCI, respectively (Table 5).

TABLE 3
Evaluating Significance of Difference for Case With Largest Change in PCS

	T_2	T_1	T_1 *ETS*	SEM	*95% CI* SEM	*90% CI* SE_p	*95% CI* SE_p	$T_2 - T_1$	$\sqrt{2}$ *SEM*	*RCI*
Physical functioning	54.4	52.8	52.2	3.0	46.3 to 58.1 (5.9)	45.3 to 59.1 (6.9)	44.0 to 60.5 (8.2)	1.6	4.3	0.4
Role–physical	53.6	27.5	28.2	3.0	22.2 to 34.1 (6.0)	21.3 to 35.1 (6.9)	19.9 to 36.5 (8.3)	26.1	4.3	6.1[a]
Bodily pain	46.1	19.9	22.1	3.8	14.7 to 29.5 (7.4)	13.7 to 30.6 (8.4)	12.0 to 32.2 (10.1)	26.2	5.3	4.9[a]
General health	45.8	25.8	28.7	4.7	19.5 to 37.9 (9.2)	18.3 to 39.1 (10.4)	16.2 to 41.1 (12.5)	20.0	6.6	3.0[a]
Energy	52.1	45.8	45.2	4.6	36.1 to 54.2 (9.0)	35.1 to 55.3 (10.1)	33.2 to 57.2 (12.0)	6.2	6.5	1.0
Social functioning	35.0	13.2	16.9	5.0	7.2 to 26.7 (9.8)	5.8 to 28.1 (11.1)	3.7 to 30.2 (13.3)	21.8	7.0	3.1[a]
Role–emotional	44.2	9.2	11.2	3.5	4.3 to 18.0 (6.8)	3.2 to 19.1 (8.0)	1.6 to 20.7 (9.5)	35.0	4.9	7.1[a]
Emotional well-being	33.1	41.6	41.5	4.8	32.0 to 50.9 (9.5)	30.9 to 52.1 (10.6)	28.8 to 54.1 (12.7)	–8.4	6.8	–1.2
PCS	55.8	37.5	37.7	2.6	32.6 to 42.7 (5.0)	31.8 to 43.5 (5.9)	30.7 to 44.7 (7.0)	18.3	3.6	5.1[a]
MCS	34.7	22.8	24.0	3.5	17.2 to 30.9 (6.9)	16.1 to 32.0 (8.0)	14.5 to 33.6 (9.5)	12.0	5.0	2.4[a]

NOTE: T_2 = observed score at follow-up; T_1 = observed score at baseline; T_1 ETS = estimated true score; *SEM* = standard error of measurement; CI = confidence interval; SE_P = standard error of prediction; $T_2 - T_1$ = observed follow-up – baseline scores; RCI = reliable change index; PCS = physical component summary; MCS = mental component summary.
a. T_2 score significantly higher than estimated baseline true score on all four indices.

TABLE 4
Amount of Change in Observed Score Required for Significant Change in Example Case and Mean Across 54 Cases (Within Parentheses) by Index

	SEM	*90% CI* SE_p	*95% CI* SE_p	*RCI*[a]
Physical functioning	5.3 (5.9)	6.3 (6.9)	7.6 (8.2)	8.4
Role–physical	6.7 (6.0)	7.7 (6.9)	9.0 (8.3)	8.4
Bodily pain	9.7 (7.4)	10.7 (8.4)	12.4 (10.1)	10.4
General health	12.1 (9.2)	13.4 (10.4)	15.4 (12.5)	13.0
Energy	8.4 (9.0)	9.4 (10.1)	11.4 (12.0)	12.8
Social functioning	13.5 (9.8)	14.8 (11.1)	17.0 (13.3)	13.8
Role–emotional	8.7 (6.8)	9.9 (8.0)	11.4 (9.5)	9.7
Emotional well-being	9.4 (9.5)	10.5 (10.6)	12.6 (12.7)	13.4
PCS	5.2 (5.0)	6.0 (5.9)	7.2 (7.0)	7.1
MCS	8.1 (6.9)	9.2 (8.0)	10.8 (9.5)	9.7

NOTE: *SEM* = standard error of measurement; CI = confidence interval; SE_P = standard error of prediction; RCI = reliable change index; PCS = physical component summary; MCS = mental component summary.
SEM and SE_P estimates were all implemented around the estimated true scores at baseline.
a. RCI is a constant for every observation so change for case and mean across cases is the same.

Kappa statistics assessing the level of agreement between indices in classifying whether patients declined, stayed the same, or improved ranged from 0.50 to 1.00, with a median of 0.82 for the 10 scores (eight SF-36 scales, PCS, and MCS).

DISCUSSION

This article illustrates the well-known fact that changes for an individual need to be much larger than changes for a group to be statistically significant. It is important in clinical settings to understand that change is sometimes gradual and may not reach significance if the lag between assessments is not long enough. In addition, it has been recommended that more than two waves of data be collected to permit more precise assessment of individual change (Speer & Greenbaum, 1995).

A variety of similar approaches to determining the statistical significance of individual change were described in this article. Which is most appropriate? The data analyzed here were from a small sample for illustration, and definitive conclusions cannot be drawn. Preference for the SE_P over the *SEM* has been expressed (Dudek, 1979). The

TABLE 5
Number of People in Sample Significantly Declining or Improving on SF-36 Scale Scores by Index

	SEM –	SEM +	SE_p *90* –	SE_p *90* +	SE_p *95* –	SE_p *95* +	*RCI* –	*RCI* +
Physical functioning	5	10	4	8	1	7	1	7
Role–physical	4	19	3	17	1	16	1	17
Bodily pain	9	17	6	15	5	13	4	12
General health	2	5	0	4	0	4	0	4
Energy	7	13	2	9	1	6	1	5
Social functioning	10	16	6	11	3	9	2	9
Role–emotional	5	13	10	10	8	8	8	8
Emotional well-being	4	14	3	11	3	10	2	10
PCS	4	18	4	16	4	13	4	13
MCS	7	20	6	16	6	12	6	12

NOTE: *SEM* = standard error of measurement; SE_P = standard error of prediction; RCI = reliable change index; PCS = physical component summary; MCS = mental component summary; – = significantly declined; + = significantly improved.
Total sample is 54 cases.

RCI has the advantage of yielding a direct test of the significance of individual change. Moreover, some have advocated for the RCI over more sophisticated alternatives (such as methods that account for regression to the mean) because of generally similar results and the ease of computing it (Bauer et al., 2004). Future investigation is needed to help determine which approach is most appropriate.

In addition to evaluating whether change is statistically significant, another consideration is the relative position of the individual at the follow-up. In clinical practice, the main focus may be on bringing the patient to the normal range of a clinical parameter. For example, normal blood pressure is defined by a systolic > 120 mmHg and diastolic > 80 mmHg, prehypertension as a systolic of 120 to 139 mmHg or > diastolic 8 to 89 mmHg, and hypertension as systolic 140 to 159 mmHg or diastolic 90 to 99 mmHg (Chobanian et al., 2003). In this case, clinicians might focus on whether their therapy takes someone who starts out in prehypertension or hypertension to within the normal range. When examining HRQOL outcomes, one might be interested in whether the patient who is in the nonfunctional range at the beginning of the treatment ends up in the functional range at the end (Bauer et al., 2004). For example, guidelines published for the RAND-36 Health Status Inventory demarcated significant positive change as either (a) positive, but insufficient; (b) favorable; (c) very favorable; or (d) optimal (Hays, Prince-Embury, & Chen, 1998).

It is well known that group-level change can be statistically significant but trivial in magnitude if the group sample size is large enough. With individual-level change, the size of differences required to be statistically significance will not be trivial. Hence, using status at the follow-up as a refinement beyond statistical significance alone is potentially useful; however, it may be that any significant change is noteworthy and important.

A distinction has been drawn between anchor-based and distribution-based methods for determining whether a significant group difference is minimally important. Distribution based methods include the effect size (ES), standardized response mean (SRM), and the responsiveness statistic (RS). For all of these indices, the numerator is the mean change and the denominators are the standard deviation at baseline (ES), the standard deviation of change for the sample (SRM), and the standard deviation of change for people who are deemed to have not changed

according to an external standard (RS). In actuality, only anchor-based methods estimate whether group change is big enough to be regarded as minimally or clinically important. The so-called distribution-based indices are simply a way of expressing the observed change in a standardized metric (Hays, Farivar, & Liu, in press).

A reliability level of .90 has been advocated as a minimum standard for measurement that is designed for interpretation of scores at the individual level, because the *SEM* is about one third of the measure's standard deviation (Table 1), and CIs around the ETS are wide at reliabilities below this recommended cutoff point (Nunnally, 1994). In practice, the .90 reliability threshold for individual assessment may be too stringent as an absolute standard. Even if measures fall short of the .90 reliability level, which they often do, obtaining this information is preferred to not doing so. Although the CI around an individual patient's score may be wide, the interval is still preferred to no information at all. Indeed, the reliabilities of five of the SF-36 scale scores (Table 2) were below this .90 cutoff value. Moreover, reliabilities for standard clinical vital signs such as blood pressure have been found to be below the .90 reliability threshold. For example, Prisant, Carr, Bottini, Thompson, and Rhoades (1992) reported 24-hour test-retest reliabilities of .87 and .67 for systolic and diastolic readings, respectively. Clinicians need to be aware of the extent of unreliability in all of their measures and interpret them with appropriate caution (Hahn et al., 2004).

Finally, the reliability estimates used in the assessment of individual change are based on group data, and there is increasing recognition that measurement precision varies along the underlying continuum of the measured construct. A more appropriate strategy for the future would be to estimate scale information at a given level of underlying HRQOL (Embretson & Reise, 2000) in tandem with an indicator of person fit to the measurement model (Reise, 2000).

REFERENCES

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment*, *82*, 60-70.

Bland, J. M., & Altman, D. G. (1994). Regression towards the mean. *British Medical Journal*, *308*, 1499.

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo, J. L., et al. (2003). Seventh report of the Joint National Committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension*, *42*(6), 1206-1252.

Collie, A., Maruff, P., McStephen, M., & Darby, D. (2003). Are reliable change (RC) calculations appropriate for determining the extent of cognitive change in concussed athletes? *British Journal of Sports Medicine*, *37*, 370-372.

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

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, *86*, 335-337.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Farivar, S. S., Liu, H., & Hays, R. D. (2004). Another look at the half standard deviation estimate of the minimally important difference in health-related quality of life scores. *Research Expert Review of Pharmacoeconomics and Outcomes Research*, *4*(5), 515-523.

Ferguson, R. J., Robinson, A. B., & Splaine, M. (2002). Use of the reliable change index to evaluate clinical significance in SF-36 outcomes. *Quality of Life Research*, *11*, 509-516.

Hahn, E. A., Cella, D., Chassany, O., Fairclough, D., Wong, G., & Hays, R. D. (2004). *A guide for clinicians to compare the accuracy and precision of health-related quality of life data relative to other clinical measures*. Manuscript submitted for publication.

Hays, R. D., Farivar, S. S., & Liu, H. (in press). Approaches and recommendations for estimating minimally important differences for health-related quality of life measures. *COPD: Journal of Chronic Obstructive Pulmonary Disease.*

Hays, R. D., Prince-Embury, S., & Chen, H. (1998). *RAND-36 Health Status Inventory.* San Antonio, TX: Psychological Corporation.

Hsu, L. M. (1989). Reliable changes in psychotherapy: Taking into account regression toward the mean. *Behavioral Assessment*, *11*, 459-467.

Hui, K. K., Zylowska, L., Hui, E. K., Yu, J. L., & Li, J. J. (2002). Introducing integrative East-West medicine to medical students and residents. *Journal of Alternative and Complementary Medicine*, *8*(4), 507-515.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Toward a standard definition of clinically significant change. *Behavior Therapy*, *17*, 308-311.

McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Quality of Life Research*, *4*, 293-307.

Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrica*, *8*, 161-168.

Nunnally, J. C. (1994). *Psychometric theory.* New York: McGraw-Hill.

Prisant, L. M., Carr, A. A., Bottini, P. B., Thompson, W. O., & Rhoades, R. B. (1992). Repeatability of automated ambulatory blood pressure measurements. *Journal of Family Practice*, *34*, 569-574.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, *35*, 543-568.

Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology*, *63*, 1044-1048.

Ware, J. E., Bayliss, M. S., Rogers, W. H., Kosinski, M., & Tarlov, A. R. (1996). Differences in 4-year health outcomes for elderly and poor, chronically ill patients treated in HMO and fee-for-service systems: Results from the Medical Outcomes Study. *Journal of the American Medical Association*, *276*, 1039-1047.

Ware, J. E., Kosinski, M., & Dewey, J. E. (2000). *How to score version two of the SF-36® health survey.* Lincoln, RI: Quality Metric.