Volume 4, Number 4, 2000/2001 ISSN 1090-655X

Journal of Outcome Measurement®

Dedicated to Health, Education, and Social Science

REHABILITATION
FOUNDATION
INC. EST. 1993
Research & Education

This issue of the
Journal of Outcome Measurement
was generously donated by
William P. Fisher, Jr.

JOURNAL OF OUTCOME MEASUREMENT®

Volume 4, Number 4 2000/2001

Articles

REVIEWER ACKNOWLEDGEMENT

The Editor would like to thank the members of the Editorial Board who provided manuscript reviews for the Journal of Outcome Measurement, Volume 4, Number 4.

JOURNAL OF OUTCOME MEASUREMENT®, 4(4), 721- 739

Weighted Health Status in the Medicare Population: Development of the Weighted Health Index for the Medicare Current Beneficiary Survey (WHIMCBS)

Jason N. Doctor, Ph.D.

Department of Rehabilitation Medicine, University of Washington School of Medicine, Seattle, Washington.

Department of Health Services, University of Washington School of Public Health, Seattle Washington.

Leighton Chan, M.D., M.P.H.

Department of Rehabilitation Medicine, University of Washington School of Medicine, Seattle, Washington.

Division of Clinical Standards and Quality, Health Care Financing Administration, Region 10, Seattle, Washington

Richard F. MacLehose, M.S.

Division of Clinical Standards and Quality, Health Care Financing Administration, Region 10, Seattle, Washington

Donald L. Patrick, Ph.D., M.S.P.H.

Department of Health Services, University of Washington School of Public Health, Seattle Washington.

Department of Rehabilitation Medicine, University of Washington School of Medicine, Seattle, Washington.

Requests for reprints should be sent to Jason N. Doctor, Department of Rehabilitation Medicine, University of Washington School of Medicine, Box # 35690, 1959 NE Pacific St., Seattle, WA 98195-6490.

We present an approach to constructing an aggregate index of health at the population level with data from Medicare beneficiaries using the 1991 (N = 12,667), 1995 (N = 15,590), and 1997 (N=17,058) Medicare Current Beneficiary Survey (MCBS). Similar to other work with survey data, we develop a weighted health status index from which one can calculate a point in time health status score for any beneficiary. Scores range from 1.0, representing "excellent health and no activity limitation", to 0.0, representing deceased. Sequences of numerically weighted health states experienced over time can be summed to calculate years of healthy life for beneficiaries. We test both the stability of the scoring system when developed on independent samples, as well as the sensitivity of years of healthy life calculations to changes in scoring assumptions. Findings suggest that, in addition to mortality, morbidity appears to play a significant role in the years of healthy life accrued by Medicare beneficiaries since entry into the Medicare program. Further, the index scoring system is highly stable when derived on independent samples. Finally, calculations of years of healthy life are robust to changes in scoring assumptions. The weighted health index for Medicare current beneficiaries (WHIMCBS) is a stable overall index of health and may be a useful ongoing indicator of health within the Medicare population.

INTRODUCTION

There is growing interest among health policy analysts in population level health indicators that consider health status. Traditional population health indicators (i.e., infant mortality and life expectancy) offer only limited information with regard to health (i.e., living vs. dead). Little information exists as to the health of a population while they are living. Yet a considerable amount of funded health services have the primary or secondary aim of improving health status.

In a recent report, entitled *Summarizing Population Health*, the National Academy of Science, Institute of Medicine, recommended that the health of a population be used to evaluate public programs and to assist in policy decision making (Field & Gold, 1998). In Canada, researchers have initiated the development of a weighted health status index that can be used at the population level to assist in policy decisions for their publicly funded health programs (Wolfson, 1996). In the United States, weighted health status scores were derived from survey responses to the National Health Interview Survey (NHIS) and applied to monitoring the goals of Healthy People 2000 (Erickson, 1998; Erickson et al., 1995).

The method employed with the NHIS has potential application with other large populations, such as the Medicare population. Medicare, the largest insurer in the United States, serves over 40 million persons. Medicare policy decisions effect the lives of millions of U.S. citizens. Clearly, there is value in being able to assess the impact of these decisions on the health of Medicare recipients. The development of one or more population level health status indices, to be used within Medicare, may provide useful information as to the benefit of programmatic decisions, the changes in health of beneficiaries over time, and the changes in the health of the Medicare population as enrollment criteria shift.

The Medicare Current Beneficiary Survey (MCBS) collects information on health status and health perceptions on a representative national sample of beneficiaries. However, because of differences in the NHIS and MCBS samples and because these two surveys differ in the questions that they ask, direct application of scores

developed on the NHIS sample to the MCBS sample is ill-advised. A careful development of scores from MCBS responses is necessary. This paper represents the first effort to develop a weighted health status index within the Medicare population. To our knowledge, no such health indicator is currently in existence.

What is Weighted Health Status?

Not unlike some economic indicators that are designed to measure economic productivity, weighted health status indices at the population level seek to gauge "health productivity", or the output of health services. Weighted health status represents health somewhere along the continuum between 0 (death) and 1.0 for (optimum health). Where persons in poor states of health achieve lower scores than persons in good or excellent health (Patrick & Erickson, 1993).

The following example illustrates the value of weighted health status measurement in health services research. Suppose Mrs. A is a person who smokes, dies at age 75, but had a life expectancy of 85 years. We may say this person has lost 10 years of life. But, death is not the only outcome for smokers. Smokers who live out their full life expectancy often do so with considerable disability. Persons who smoke are at high risk for cardiovascular disease, pulmonary diseases and cancer. Morbidity associated with these diseases can lead to diminished health status during one's later years of life. Thus, an analysis of <u>only</u> years lost to smoking does not fully capture the impact of smoking on a person's health.

Weighted health status indices are better able to capture population health, because they permit gradations of disability to be compared to one another. When weighted health status is studied over time, one can calculate years of healthy life (YHL) achieved. Suppose Mrs. A. enrolls in Medicare at age 65 and lives to age 75. Assume during these 10 years that Mrs. A has some health problems and that she perceives her health as less than perfect. Suppose also that her weighted health status score is 0.5 and she lives in this state of health without change. Over a 10 year period, Mrs. A has achieved 5 years of healthy life (i.e., 10 years of survival x 0.5 health status =

5 YHL). If there are 1,000 people in the Medicare program, with the same health status as Mrs. A, each living over the same 10 year period, then the program has achieved 5,000 YHL for these individuals (1,000 persons x 5 YHL = 5,000 YHL). Improving the health of Mrs. A, and those like her, will result in an increase in the number of years of healthy life accrued by beneficiaries. Policy changes that adversely impact the health status of Mrs. A and others in the Medicare program will result in a decrease in the number of years of healthy life accrued by Medicare recipients. The appeal of this approach is that it offers an index describing the outcome of health services, and it strengthens the emphasis on accountability of those providing health services to Medicare recipients.

As is clear from the previous example, the purpose of the weighted health status approach is to adjust for the quality of health over time. Under such a model, a health status score is assigned to each unit of survival duration (typically each year of life over a specified time period). The score is weighted to reflect severity of disability, as well as, each participant's perception of the quality of their health (Patrick & Erickson, 1993). When studied over time, accrued scores can be used as an estimate of years of healthy life (YHL) achieved. The YHL technique is a close cousin to the popular quality-adjusted life years (QALY) model. However, this latter model employs the expected utility of health states over time and is often implemented in smaller studies with assessment of utility on a case by case basis (Drummond, O'Brien, Stoddart, & Torrance, 1997).

In this paper, we present a general health state classification system for use with Medicare survey data. We also empirically derive health status weights using methods previously validated on the National Health Interview Survey (Erickson, 1998; Erickson et al., 1995; Torrance et al., 1995). The NCHS monitors the nations health with the Health and Activity Limitation Index (HALex) (Erickson, 1998; Erickson et al., 1995). Thus, our methodology is consistent with the procedures developed at the National Center for Health Statistics (NCHS) for monitoring the Nations health goals. We then test the stability of the scoring system, by deriving health status weights for 6 independent Medicare samples and comparing the consistency

of the values. Finally, we conduct a sensitivity analysis on scoring assumptions by examining the impact of varying these assumptions on the calculation of years of healthy life remaining for Medicare beneficiaries in a life table.

METHODS

Data Source

The MCBS is a longitudinal survey of a nationally representative sample of Medicare beneficiaries. The Health Care Financing Administration (HCFA) sponsors the survey. According to the MCBS protocol, between 12,000 and 17,100 individuals are interviewed three times during the year. Responses are recorded to a wide range of questions regarding demographics, health care utilization, access to and satisfaction with care. The MCBS uses a multistage, probability sampling design based on 107 geographic primary sampling units. The oldest-old participants (age 85 years and older) and people with disabilities (age younger than 65 years) are over sampled. Due to the nature of the sampling design, sample weights and Taylor series estimates of the standard error were used (Adler, 1994). We utilized the 1991 (N = 12,667), 1995 (N = 15,590), and 1997 (N=17,058) survey data for the present study.

Health State Construction

The first step in developing a weighted health status measure involves defining the health states under study. Health states within the WHIMCBS, like the HALex, are characterized by two dimensions of outcome (Erickson, 1998; Erickson et al., 1995; Torrance et al., 1995). The first dimension, Activity Limitation consists of three ordered categories describing increasing disability: 1) 'no activity limitation', 2) 'limitation in at least one instrumental activity of daily living (I-ADL)', and 3) 'limitation in at least one activity of daily living (ADL)' (Patrick & Erickson, 1993). Questions on the MCBS regarding activity limitation are similar, but not identical to those

given to persons over 65 years of age in the NHIS sample (see Table 1). The second dimension, Perceived Health, consists of ratings in five ordered categories: 1) 'Excellent', 2) 'Very Good', 3) 'Good', 4) 'Fair', and 5) 'Poor'. The perceived health question is also part of the National Health Interview Survey and is identical to that used for perceived health classification with the HALex. By crossing the activity limitation dimension with the perceived health dimension, living Medicare current beneficiaries each achieve 1 of 15 possible health states. Each of these 15 states will be assigned weighted value.

ANALYSES

Correspondence Analysis

Correspondence Analysis (CA) is a statistical modeling technique for describing the relationship between categorical data in a contingency table. Of interest in the present study is a 3 x 5 activity limitation by perceived health contingency table. With CA, relative frequencies within table cells and marginal proportions (row and column proportions) are calculated. A computation of the chi-square distances between categories of the same variable is carried out. That is, distances between activity limitation categories are computed separately from distances between perceived health categories. Then, the best-fitting two-dimensional space is calculated via a principal components analysis. The technique requires no distributional assumptions to measure chi-square distances. A chi-square statistic, is used to test the null hypothesis that the distribution of responses with respect to perceived health and activity limitation are mutually independent. A statistically significant chi-square test along with evidence that one dimension in the two dimension space explains most of the covariation in the table, supports both the hypothesis that perceived health and activity limitation measure the same underlying construct, and the use of principal component weighted perceived health and activity limitation categories. Detailed information on the algorithm used for this analysis is available elsewhere (Clausen,

Table 1 **Comparison of the MCBS vs. NHIS questions used in the classification of activity limitation health-related quality of life dimension.**

Activity Limitation	MCBS	NHIS
I-ADL	Person responded 'yes' to any of the following and 'no' to ADL questions: For health reasons … 172. Receive help with Phone? 173. Receive help with light housework? 174. Receive help with heavy housework? 175. Receive help with making meals? 176. Receive help with shopping? 177. Receive help with managing money?	Person responded 'yes' to the following question (14a) and 'no' to 14b: Limited in activities of daily living 14a. Because of any impairment or health problem, does ___ need the help of other persons in handling ___ routine needs, such as everyday household chores, doing necessary business, shopping, or getting around for other purposes?
ADL	Person responded 'yes' to any of the following: For health reasons … 208. Receive help bathing/showering? 209. Receive help dressing? 210. Receive help eating? 211. Receive help getting in/out of chairs? 212. Receive help walking? 213. Receive help using the toilet?	Person responded 'yes to the following: Limited in activities of daily living 14b. Because of any impairment or health problem, does ___ need the help of other persons with ___ personal care needs, such as eating, bathing, dressing, or getting around this home?
Limitation in Other Activities	Not Applicable	Limited in other activities 6a. Is ___ limited in ANYWAY in any activities because of an impairment or health problem?

1998; Greenacre, 1984).

Correspondence Analysis is appropriate for our purposes since we are interested in perceived health and activity limitation to the extent that they help explain the same underlying health status. In the present study, two-way perceived health by activity limitation tables for male and female Medicare beneficiaries in 1991, 1995, and 1997 were estimated using the WesVarPC variance estimation program (Westat, 1997). Correspondence analysis of these tables were completed in SAS (SAS Institute, 1990).

It is important to note that CA is able to quantify distances between different levels of health within each of the two dimensions, but does not offer a way to combine these into one value. One value is needed because weighted health status reflects a person's point-in-time overall health. Ultimately, health status score is based on a composite score of each individual's perceived health and level of activity limitation score. Assigning rules for composite scoring requires further analytical work. This work is discussed below.

Multiattribute Modeling

In essence, a multiattribute model, establishes a rule for combining activity limitation and perceived health weights obtained through CA into a single number representing weights for health states achieved in the model. In the WHIMCBS outcomes model, participants can achieve one of 15 different health states. Each of these health states represents some combination of perceived health and activity limitation. Put another way, each health state is represented by an activity limitation and perceived health "pair". A multiattribute model was employed to produce health status weights for each activity limitation-perceived health pair (Keeney & Raiffa, 1993). Multiattribute model assumptions are equivalent to those described by Torrance et al. 1995 and were used with the development of NCHS' weighted health status index for monitoring the goals of healthy people 2000 (Erickson et al., 1995). These assumptions for forming a composite score are as follows: 1) the perceived health and activity limitation dimensions are given equal weight in the model, 2) the health state

"ADL Limitation and Excellent Health" receives a scale weight of 0.47, and 3) persons with an "ADL Limitation and Poor Health" receive a weight of 0.10. Assumption 1 is prescriptive, in that we have chosen to place equal importance on participants' perception of their health as we do on behavioral indicators of their health. The value used in Assumption 2, is associated with the weight used for a person having an ADL limitation in the Health Utilities Index Mark I (Drummond et al., 1997). This weight was obtained from studies of people's preferences for different health outcomes. Assumption 2 is needed to facilitate a solution to other health state pairs in the model. Assumption 3 is also needed to facilitate a solution to other health states in the model. This third assumption is most arbitrary. We tested the sensitivity of our calculations to changes in this assumption and found our model was robust to such changes. Findings from this sensitivity analysis are presented later in the paper.

All model assumptions and statistical techniques used for scale development were identical to those employed by the NCHS in developing a weighted health status index to monitor national health goals. This was done to establish an index that would allow comparisons between the national sample and the Medicare sample. Since our assumptions are stated explicitly, future work can be done to test their plausibility and or the effect of modifying them.

Stability of the Scoring System when Developed on Independent Samples

We assessed the stability of the independently derived weighting structure over three time points (i.e., 1991, 1995, and 1997), and across male and female beneficiaries. Using standard methods (Shavelson, Webb, & Rowley, 1989), an intraclass correlation coefficient (ICC) was computed that expresses the ratio of the of the "universe-score" variance, or the variance component due to health-state across the universe of observations studied (i.e., across time and gender), to the expected observed-score variance. The ICC for health state weighting is represented as follows:

$$\text{ICC} = \frac{\sigma^2_{HS}}{E\sigma^2_x} = \frac{\sigma^2_{HS}}{\sigma^2_{HS} + \sigma^2_{error}}$$

where σ^2_{HS} is the universe-score variance, and $E\sigma^2_x$, is the expected observed-score variance, which is comprised of the universe-score variance (σ^2_{HS}) plus an error term (σ^2_{error}) All variance components associated with a health state value, except for that attributable to the health state itself, are defined as error. Thus, in the present investigation, error consists of the variance component due to time (T), gender (G), health state by time interaction (HS x T), health state by gender interaction (HS x G), time by gender interaction (G x T) and residual (HS x T x G, confounded with error).

Sample estimates of the parameters in the ICC are used in the computation (Shavelson et al., 1989). We employed the VARCOMP procedure in SAS (SAS Institute, 1990) to estimate the necessary parameters for this computation. In our analysis, we utilized only the 11 health state values allowed to vary in the multiattribute model so as not to artificially inflate the coefficient.

Sensitivity Tests of Multiattribute Model Assumptions

We tested the sensitivity of our scoring model to changes in Assumption 3, mentioned above. Specifically, we varied the worst health state cell between 0.2 and 0.01 and tested the effect on years of healthy life calculations for beneficiaries.

Years of Healthy Life Remaining

We calculated YHLs remaining for older male and female Medicare beneficiaries without disability benefits from WHIMCBS scores and life table data provided by the Social Security Administration (Social Security Administration, 1998) using standard methods (Patrick & Erickson, 1993; Shryock, Siegel, & al., 1971). We utilized the weighting structure in Table 2, applied to the 1995 beneficiary data in combination with life table data based on mortality rates in 1996. Thus, the procedure insured continuity between health status and mortality data.

RESULTS

Weighting Results

Using 1991 data, correspondence analysis indicated a significant association between the perceived health and activity limitation dimensions ($\chi^2(8) = 1388.29$, $p < .001$). As expected, increased disability most often corresponded to endorsements of lower perceived health. This analysis was subsequently replicated across years (1991, 1995, and 1997) and derived separately by gender. In all derivations, statistically significant and highly similar results were observed. A single principal component explained greater than 98% of the covariation in the contingency table.

Multiattribute model computations resulted in a weighting structure that was both highly stable across time and across gender. Table 2 illustrates the health state scoring averaged over men and women in 1991. As is discussed below, scoring systems derived separately by gender and year produced highly similar results.

Stability of Scoring System on Independently Derived Samples

Table 3 displays estimated components of variance contributing to measurement of health state and the intraclass correlation coefficient. The WHIMCBS health state weighting appears highly stable across time and gender (ICC = 0.98). Thus, when weighting structures were derived independently for males and for females within each year studied (1991, 1995, and 1997) the resulting structures are close to equivalent.

Years of Healthy Life Remaining

Results from the YHL calculations are presented in Table 4. Here we see expected years of life remaining, adjacent to years of healthy life remaining for male and female Medicare recipients of different ages. The last two columns reflect the difference between life years and years of healthy life for each age group, and the percent differ-

Table 2 **Multivariate health quality scores for all Medicare Beneficiaries**

Activity Limitation	Perceived Health Status					
	Excellent	Very Good	Good	Fair	Poor	Dead
No Limitation	1.0	0.95	0.84	0.65	0.47	
Limited in I-ADL	0.63	0.58	0.50	0.36	0.21	
Limited in ADL	0.47	0.43	0.36	0.23	0.10	
Dead						0.0

Table 3 **Estimated components of variance for different sources of variation and intraclass correlation (ICC) for health state weighting.**

Source of Variation	Estimated Variance Component
Health State (HS)	0.05613
Gender (G)	0.00014
Time (T)	0.00000
HS x G	0.00000
HS x T	0.00000
G x T	0.00044
HS x T x G	0.00065
Intraclass Correlation Coefficient (ICC) = 0.98	

ICC was computed on 11 health states, for the two genders, over three time points (1991, 1995, & 1997).

ence, respectively. The calculations suggest that YHL estimation over the last 30+ years of life (ages 65 ≥ 95), depart significantly from estimated life years. This is particularly true for women who as a group live longer and thus are susceptible to decrements in quality of life after age 85 years.

Table 4 **Years of Health Life Remaining by age, for Male and Female non end stage renal disease and non-disability eligible Medicare Beneficiaries in 1995.**

AGE	Life years Remaining	Years of Health Life Remaining	Difference Years	Difference %
Men				
65 -69	15.4	11.7	3.7	24
70 - 74	12.4	9.0	3.4	27
75 - 79	9.6	6.6	3.0	31
80 - 84	7.3	4.7	2.6	36
85 - 89	5.4	3.2	2.2	41
90 - 94	4.0	2.1	1.9	48
95 & over	2.9	1.5	1.4	52
Women				
65 -69	19.0	13.1	5.9	31
70 - 74	15.4	10.2	5.2	34
75 - 79	12.0	7.5	4.5	38
80 - 84	9.1	5.1	4.0	43
85 - 89	6.6	3.3	3.3	50
90 - 94	4.7	2.1	2.6	55
95 & over	3.1	1.2	1.9	61

Sensitivity Analysis

Sensitivity tests varying the arbitrary assignment of 0.10 to the worst health state ("poor health", "ADL Limitation"), indicated that YHL calculations were fairly robust to changes in this assumption. As mentioned above, under this assumption male beneficiaries age 65 can expect 11.7 YHL remaining, whereas female beneficiaries can expect 13.1 YHL. Changing the worst health state value, from 0.1 to 0.2, resulted in 12.3 YHL for males (5% increase) and 14.0 YHL for females (7% increase) age 65 years. Varying the value for the worst health state in the other direction, from 0.1 to 0.01, resulted in 11.6 YHL for males (o.9% decrease) and 12.9 for females (2% decrease) age 65 years. Persons age 65 are most sensitive to changes in the model assumptions, because they stand to live the greatest number of years. Sensitivity analyses suggest that changes in this model assumption result in only minor changes in the total accrual of YHL.

DISCUSSION

Development of a weighted health status index from responses to health-related questions in large health survey databases is highly desirable. The WHIMCBS appears to show promise as a measure of health among Medicare recipients. This method can be used to track the health of a population over time and under changing health service conditions. One potential application is to track the health of Medicare recipients through the fluctuation of health maintenance organizations enrollment.

Another potential use of the WHIMCBS is for the modeling of prognosis for different Medicare subgroups. For example, Manton (1987; 1988; 1989) has studied the rate of movement between health states of persons with various chronic diseases. In his work, Manton has relied on an assortment of data sources from clinical studies, epidemiological studies, and expert opinion. From these data sources he has developed models for forecasting future distributions of disease within populations. Researchers working with MCBS and other Medicare data are in a position to develop similar forecasting mod-

els for the Medicare population. Applying the WHIMCBS to such forecasting work would offer both a health state structure for examining transition rates to other states and an analysis of changes in YHL for different groups over time.

There are limitations to our methodology that need to be mentioned. The use of a scaling method such as correspondence analysis may be less desirable than obtaining weights directly through elicitation of preferences of the survey participants for health states (Drummond et al., 1997). However, obtaining these preferences is generally seen as cost prohibitive. Therefore, these data are not available in large national survey studies like the MCBS. In addition, although we employed a sensitivity analysis varying the value of the worst health state, analysis varying the value used for excellent health and an ADL-limitation was not performed. We chose not to vary this value because the value we use has empirical support in studies of preferences for health states (Drummond et al., 1997). Further, the same value was used in the calculation of years of healthy life for monitoring the goals of healthy people 2000 (Erickson, 1998; Erickson et al., 1995). However, future studies examining the robustness of the model to this assumption may be warranted. Finally, we wish to point out that although the methods used in the development of the WHIMCBS index have been used to monitor national health goals, and have been successfully validated on a national sample, the approach is not the only or best way to index health status. Development of alternative approaches to complement the WHIMCBS may help us better understand the overall health of Medicare beneficiaries.

In the present study, health status weights for perceived health and activity limitation dimensions were empirically derived from beneficiary responses so as to maximize the association between responses on the two dimensions. An advantage of this method is that it incorporates each individual's perceived health in the weighted health status model. Because of this, the WHIMCBS scores of persons with disabilities, who may have low scores on activity limitation, can be adjusted for their perceived health. Such an approach may be important to beneficiaries with chronic disabilities (e.g., spi-

nal cord injuries), as well as to older persons experiencing the normal effects of aging. These persons may feel they are healthier than their physical limitations suggest from evaluation with a measure that is based entirely on activity limitation.

Currently, we see the WHMCBS as a useful population health indicator that will provide information to policy makers as to whether the general health of the Medicare population is improving or not. Increases or decreases as registered by the WHMCBS are indications of changing perceived health and activity limitation levels of beneficiaries. Comparisons of Life Years remaining to Years of Healthy Life remaining provides an indication of the amount of life lost to morbidity.

ACKNOWLEDGEMENTS

This research was supported by a grant from the National Institute of Child Health and Human Development (NICHHD) #HD07424-06. The authors would like to thank Peter Houck, M.D., Division of Clinical Standards & Quality, HCFA, Region 10, for his support of this project. The views expressed in this article are those of the authors and are not necessarily those of the Health Care Financing Administration, University of Washington, or the NICHHD.

REFERENCES

Adler, G. S. (1994). A profile of the Medicare Current Beneficiaries Survey. *Health Care Financing Review,* 15(4), 153-163.

Clausen, S.-E. (1998). *Applied Correspondence Analysis*. (Vol. 7). Thousands Oaks: Sage.

Drummond, M. F., O'Brien, B. J., Stoddart, G. L., & Torrance, G. W. (1997). *Methods for the economic evaluation of health care programmes*. (2nd ed.). Oxford ; New York: Oxford University Press.

Erickson, P. (1998). Evaluation of a population-based measure of quality of life: the Health and Activity Limitation Index (HALex). *Qual Life Res,* 7(2), 101-14.

Erickson, P., Wilson, R. W., & Shannon, I. (1995). *Years of Healthy Life: Statistical Note Number* 7. Hyattsville, MD: National Center for Health Statistics.

Field, M. J., & Gold, M. R. (1998). *Summarizing population health* Washington: National Academy of Sciences, Institute of Medicine.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Keeney, R. L., and Raiffa, H. (1993). *Decisions with Multiple Objectives*. (2nd ed.). New York: Cambridge University Press.

Manton, K. G. (1987). Forecasting health status changes in an aging U.S. population: assessment of the current status and some proposals. *Climatic Change*, 11, 179-210.

Manton, K. G. (1988). *Chronic Disease Modeling*. New York: Ox ford University Press.

Manton, K. G. (1989). Epidemiological, demographic, and social correlates of disability among the elderly. *The Milbank Quar terly*, 67 (suppl 2 pt 1), 13-58.

Patrick, D. L., & Erickson, P.(1993).*Health status and health policy: quality of life in health care evaluation and resource allocation*. New York: Oxford University Press.

SAS Institute, I. (1990). SAS/STAT User's Guide (Version 6). Cary: SAS Institute Inc.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability Theory. *American Psychologist*, 44(6), 992-932.

Shryock, H., Siegel, J., & al., e. (1971). *The methods and materials of demography*. (Vol. 2). Washington: U. S. Government Printing Office.

Social Security Administration. (1998). Period Life Table, 1996 (Table 4.C6). *Social Security Administration Bulletin, Supplement.*

Torrance, G. W., Erickson, P., Patrick, D. L., & Feldman, J. J. (1995). *Technical Notes. In Years of Healthy Life. Statis tical Note No. 7, by P. Erickson, R.W. Wilson, and I. Shannon*. Hyattsville, MD: National Center for Health Statistics.

Westat. (1997). A User's Guide to WesVarPC (Version 2.1). Rockville: Westat, Inc.

Wolfson, M. C. (1996). Health-Adjusted Life Expectancy. *Health Reports*, 8, 41-46.

JOURNAL OF OUTCOME MEASUREMENT®, 4(4), 740-762

ECRHS Screening Questionnaire Scoring: A Methodological Suggestion For Asthma Assessment

G. Biino
C. Rezzani
M. Grassi
A. Marinoni

Department of Applied Health Sciences
University of Pavia

Univocal definition and classification of Asthma have always been a matter of discussion, and that is reflected in the difficulty of constructing a measure of pathology severity. The European Community Health Survey is a multinational survey designed to compare the prevalence of asthma in subjects, aged 20 to 44 years, in several European areas. In each participating center a sample of 3000 adults filled a self-administered screening questionnaire composed by 9 dichotomous items. Aim of the present study is to investigate unidimensionality of the ECRHS screening questionnaire and to determine and validate a scoring of *asthma-like symptoms seriousness*. Dimensionality and scoring was determined through a Homogeneity Analysis by Alternating Least Square; while scoring validation was assessed by a cross validation technique. This study found the existence of a sole dimension underlying the screening questionnaire; furthermore a scoring of *asthma-like symptoms seriousness* was determined with the indication of a cut-off in order to distinguish between asthma symptomatic and non symptomatic subjects.

Requests for reprints should be sent to G. Biino, Dipartimento Di Scienze Sanitarie Applicate Universita' Degli Studi Pavia, Via Bassi 21, 27100 Pavia, Italy

INTRODUCTION

The main intent for conducting standardized international prevalence studies is the current concern that the prevalence of asthma may be increasing [1-7]. Although methodological differences in these studies make it difficult to compare the magnitude of the differences in asthma prevalence between countries, the trend of increasing prevalence in countries of widely differing lifestyles and ethnic groups is generally consistent.

The principal problem in assessing asthma prevalence is that the definition and classification of asthma has continued to be a subject of controversy. In fact the diagnosis of asthma involves an overall assessment of the patient's medical history, physical examination, and laboratory test results, and there are no standardised methods for combining the information from these various sources. In general two methods have been used in the past to measure asthma prevalence: questionnaires and physiological measurements. Standard written questionnaires have been the principal instrument for measuring asthma symptom prevalence in community surveys, and in homogeneous populations these have been standardised, validated, and shown to be reproducible [8]. While measurement of bronchial responsiveness in epidemiological studies of obstructive respiratory disease have been increasingly used in the past decade, and standard procedures are now well established [9]. Therefore a step towards a standardisation of the methods for combining various information could be assessing the capability of the European Community Respiratory Health Survey (ECRHS) *screening questionnaire* to discriminate between symptomatic and asymptomatic subjects, in order to select people to undergo instrumental measurements and clinical examinations.

Many studies have been conducted on the ECRHS *screening questionnaire* [8, 10,11] but almost all of them consider the prevalence or sensitivity and specificity of symptoms corresponding to individual items without considering the questionnaire as a whole and thus without computing a score.

Aim of our study is to investigate unidimensionality or multidimensionality of the ECRHS screening questionnaire. If there is

unidimensionality it means assuming the existence of one latent trait, for example *asthma-like symptoms seriousness*, being able to explain the inter-item correlation; if unidimensionality is present, we can determine and validate a scoring of *asthma-like symptoms seriousness* i.e. a sort of classification from an asymptomatic condition till a serious symptomatic stage based upon the questionnaire's item responses. The scoring of each subject, assuming values in the continuous numbering, would permit a more precise measuring of the extreme variability existing among individuals. On the contrary if there is multidimensionality it means the questionnaire measures many symptomatic traits, in which case it would be proper to ascertain them.

METHODS

Data and subjects

The ECRHS is a multinational survey designed to determinate the differences in the prevalence of asthma across Europe, to estimate variation in exposure for knowing risk factors for asthma and to asses the extent to which they may explain differences across Europe, and finally to estimate the variation in treatment practice for asthma in the European Community [12]. In each center a representative sample of 3000 adults, aged 20 to 44 years, should have filled a short screening questionnaire, containing seven main dichotomous items and two further items conditional on a positive answer to the first one (Appendix A). A random sub-sample of 600 subjects and an additional sample of up to 150 "symptomatic" individuals should have been then studied in more detail in phase II, with measurements of skin-prick test to common allergens, serum total and specific IgE, bronchial responsiveness to inhaled methacoline and urine electrolytes, as well as with an additional long questionnaire on asthma symptoms and medical history, occupation and social status, smoking, home environment, and use of medication and medical services. Both, the short and the long questionnaire, were derived from the International Union against Tuberculosis and Lung Disease (IUATLD) questionnaire on bronchial symptoms. The short questionnaire (*screen-*

ing questionnaire) was sent by mail and self-administered; while the long one was administered by interviewers.

In this paper we use data (n=6946) collected by the three Italian centers participating to ECRHS' phase I: Pavia, Verona and Torino. Data are divided into three samples: in the first one there are subjects from Verona and Torino who completed only the *screening questionnaire* sent by mail (n=4620); the second one contains the subjects of Pavia to whom the long questionnaire was administered door-to-door (n=1216); finally, the third one contains all the subjects, belonging to the three towns, for whom *screening questionnaire* and "post consensus" clinicians' diagnosis of asthma are available (n=1109).

Homogeneity Analysis

In order to investigate unidimensionality of the *screening questionnaire* items we run an Homogeneity Analysis by means of Alternating Least Squares (HOMALS procedure implemented in SPSS© Category Module) on the entire "mail" sample (the first sample: Torino plus Verona). Through this procedure [13] we obtained "category quantifications" for each item category. Thus, starting from dichotomous responses, we are allowed getting to numerical values called *optimal quantifications* of binary variables responses. These continuos responses are called optimal quantifications since they minimise the differences between observed and true scores, that is they minimise the measurement errors. Such optimal quantifications allow determining the *subjects' score* on latent dimensions. The score is the sum of the subject's item responses, recoded by optimal quantifications. As far as regards items 1.1 and 1.2, which are conditioned on the response to item1, when a subject answered "No" to item 1, we replaced missing values in items 1.1 and 1.2 with "No". Therefore, we have dealt with these items like all the others.

Besides, for each (latent) dimension of the HOMALS solution it is also possible to obtain the *discrimination coefficients* of the screening questionnaire items, whose average value is the *eigenvalue*, i.e. the total variability of the latent dimension underlying the items. The discrimination coefficients can be interpreted as the dimension

variability explained by each item, while the eigenvalue is an homogeneity index and it measures the fit of each dimension of the HOMALS solution; analogously to Principal Component Analysis of continuos variables, the eigenvalues are in decreasing order.

Each variable category receives as many quantifications as the number of latent dimensions; we are particularly interested in the first dimension where the category quantifications of each variable have maximum spread. We use some unidimensionality descriptive indices: first a graphical representation (screen plot: see Cattel [14]) of all eigenvalues extracted from the sample; secondly we computed Cronbach index (α) which is a reliability coefficient based on the internal consistency [15], and Greenacre (τ) which is the percentage of inter-item squared correlation explained by the relationship between items and score [16]. (A list of the descriptive indices is reported in Table 1). As threshold for unidimensionality goodness-of-fit we use (following Nunnally [17:p245]) $\alpha > 0.70$ and (following Morrison [18: p.228]) $\tau > 0.75$.

Where p is the number of items, y_j is item x_j (j=1,...,p) recoded with optimal quantifications, and $y = \sum_{j=1}^{p} y_j$ is the subject score.

We have also considered a "scaled" score version (min=0 & max=10). With more details.

$y^* = \sum_{j=1}^{p} y_j^*$, where:

$$y_j^* = \begin{cases} 0 & \text{if "NO"} \\ \dfrac{q_{NOj} - q_{YESj}}{\sum_{j=1}^{p} q_{NOj} - \sum_{j=1}^{p} q_{YESj}} \cdot 10 & \text{if "YES"} \end{cases}$$

the figure q_{NOj} and q_{YESj} are the optimal quantifications of the "No" and "Yes" binary item responses, respectively.

Table 1 Descriptive indices obtained by Homogeneity Ananlysis

Descriptive index	Formula
Discriminant measures	$\delta_j = \mathrm{cor}^2(y_j; y) = \dfrac{\mathrm{cov}^2(y_j; y)}{\mathrm{var}(y_j) \cdot \mathrm{var}(y)}$
First Eigenvalue	$\lambda_1 = \dfrac{\sum_{j=1}^{p} \delta_j}{p}$
Cronbach index	$\alpha = \dfrac{p}{p-1} \cdot \left(1 - \dfrac{1}{p \cdot \lambda_1}\right)$
Greenacre index	$\tau = \dfrac{\left(\dfrac{p}{p-1}\right)^2 \cdot \left(\lambda_1 - \dfrac{1}{p}\right)^2}{\overline{\Phi}^2}$ $\overline{\Phi}^2 = \dfrac{\sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \mathrm{cor}^2(x_j; x_k)}{p \cdot (p-1)/2}$

Finally, we computed the standard errors of the category quantifications. The eigenvalues of the HOMALS solution are equivalent to the Correspondence Analysis [16] solution which extracts the eigenvalues from the sample "Burt matrix", a block symmetric matrix of the all two-way contingency table of the binary variables analogous to the covariance matrix of continuos variables. Assuming that the

Burt matrix to be analysed is a frequency table and that the data are a random sample from an unknown population, the cell frequencies follow a multinomial distribution. From this, it is possible using "delta method" (provided by SPSS ANACOR procedure) to compute the standard deviations of the category quantifications, see Gifi [13: p 408-415]. "Small" standard deviations indicate that HOMALS would produce the same solutions for a slightly different sample from the same population, so we could state that HOMALS obtained an overall stable solution.

Cross-validation study

In order to validate the stability of HOMALS goodness-of-fit results (discriminant coefficients and eingenvalue) we performed a (grouped) jack-knife K-fold cross-validation procedure [18] applied to both the eigenvalue, λ and the discriminant measures, δ.

We selected at random K =10 groups, A_k $(k=1,\ldots,10)$ of n_k = $n/10$ (with n = 4620 and n_k = 462) subjects each (without replacement) from the "mail" sample subjects (Verona plus Torino). Let B_k (k=1,...,10) denote the groups of (n - n_k = 4158) individuals obtained by dropping A_k from the entire data set. We shall refer to B_k as *jack-knifed* sets and A_k as *cross-validation* sets. For each of the jack-knifed sets we performed a HOMALS analysis and computed an and δ's. The λ and δ's are then averaged over all B_k (k=1,...,10) in order to obtain the jack-knife eigenvalue, λ_{JN} and the jack-knife measures, δ_{JN}. We also computed the cross-validation eigenvalue, λ_{CV} and the cross-validation discriminant measures δ_{CV}: this is done by using the optimal quantification parameters obtained on B_k to calculate the λ and δ's on the corresponding cross-validation set A_k and by averaging over the A_k's.

The HOMALS results are stable if $\lambda_{JN}(\lambda_{CV})$ and δ_{JN} (δ_{CV}) are not too far from λ and δ's calculated on the entire data set. We also used the long questionnaire sample of Pavia (n=1137

subjects excluding 79 subjects who had at least one missing item response) as an additional cross-validation sample of the entire "mail" sample of Torino and Verona.

External Criterion Study

The subjects' *score* obtained by the HOMALS procedure assumes values in the continuous numbering, and thus it can be treated as a continuous or dichotomous variable.

Treating the *score* as a continuous variable, we tried to correlate it with a variable that measures bronchial hyperresponsiveness (BHR). For this propose we used the variable PD_{20}, obtained as the methacholine dose which causes a 20% fall in Forced Expiratory Volume in one second, FEV_1 [9] for ethical reasons the methacholine administration has been stopped at the dose to which a subject could not be considered as BHR symptomatic anymore. The dose values are 0.0312, 0.0625, 0.1250, 0.2500, 0.5000, 1.0000, 2.0000, 4.0000, 6.0000, 8.0000 mg/ml. in order to ascertain a correlation between the score and the BHR we excluded the subjects BHR asymptomatic (PD_{20}>8) and we perform a One Way Analysis of Variance, considering as a criterion the PD_{20} values and as response variable the *score.*

Treating the *score* as a dichotomous variable, for each observed score values sensitiity and specificity have been computed with reference to clinical diagnosis available from the "post consensus" sample (n=1011 excluding 98 subjects who had at least one missing item response).

Sensitivity is the probability that a subject with a score greater than the tested one is asthmatic in respect with the clinical diagnosis, whereas specificity is the probability that a subject with a score smaller than the tested one is non-asthmatic in respect with the clinical diagnosis. Therefore, in order to determine the best threshold of the *score*, the Receiver Operating Characteristic (ROC) curve, generated by varying the cut point on the questionnaire total score, has been plotted [20]. The ROC curve is constructed by plotting sensitivity on the y-axis and 1-Specificity on the x-axis, for all possible cut-off values of a diagnostic test.

The full protocol followed to classify asthma status of each subject is described elsewhere [21], Briefly, subjects were defined as "asthmatic" or "not asthmatic" according to the diagnoses that three experts formulated after having examined: responses to a clinical interview, respiratory function tests and finally allergy tests. Each expert physician formulated a diagnosis using his own experience and knowledge, without consulting the other two physicians. If there was agreement among the clinicians, a diagnosis label was immediately given. On the other hand in case of disagreement, diagnosis was formulated when the experts reached their greatest agreement, after discussing all the data.

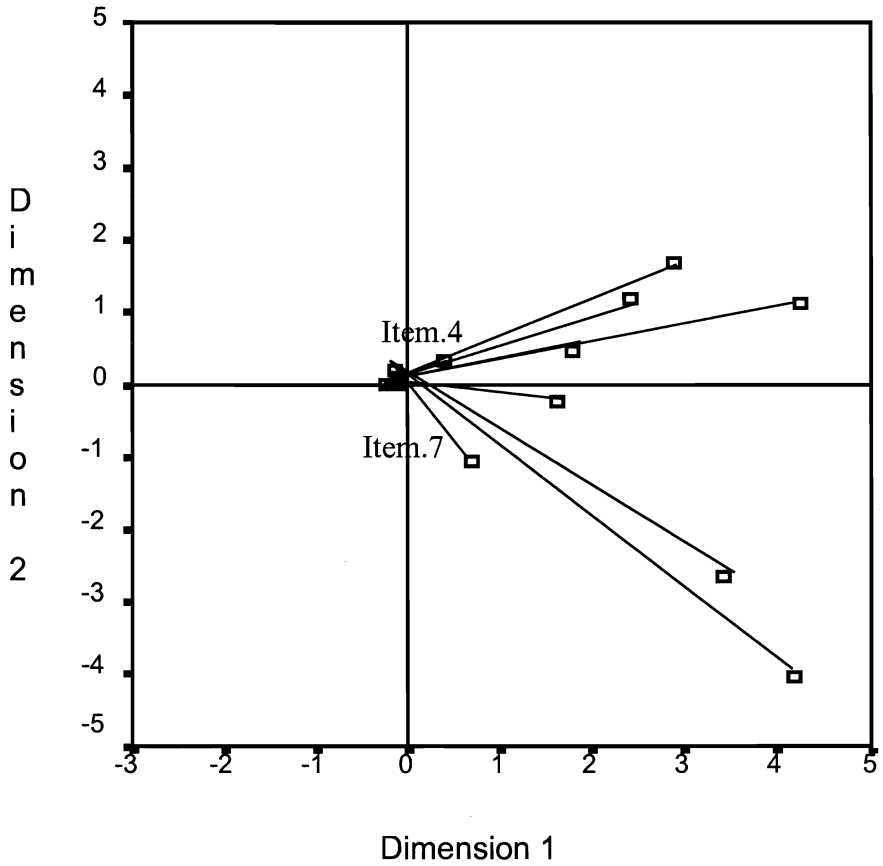
We also compare sensitivity and specificity with reference to clinical diagnosis available from the "post consensus" sample of the *score* cut-off, our classifier criterion, with another and more detailed criterion according to Toelle working definition of *current asthma* [22]. In 1992 Toelle and co-workers defined *current asthma* as the presence of symptomatic BHR, that is bronchial hyperresponsiveness to a challenge test, plus a positive reply to the items concerning recent wheeze or exercise wheeze present in the clinical interview.

RESULTS

Homogeneity Analysis

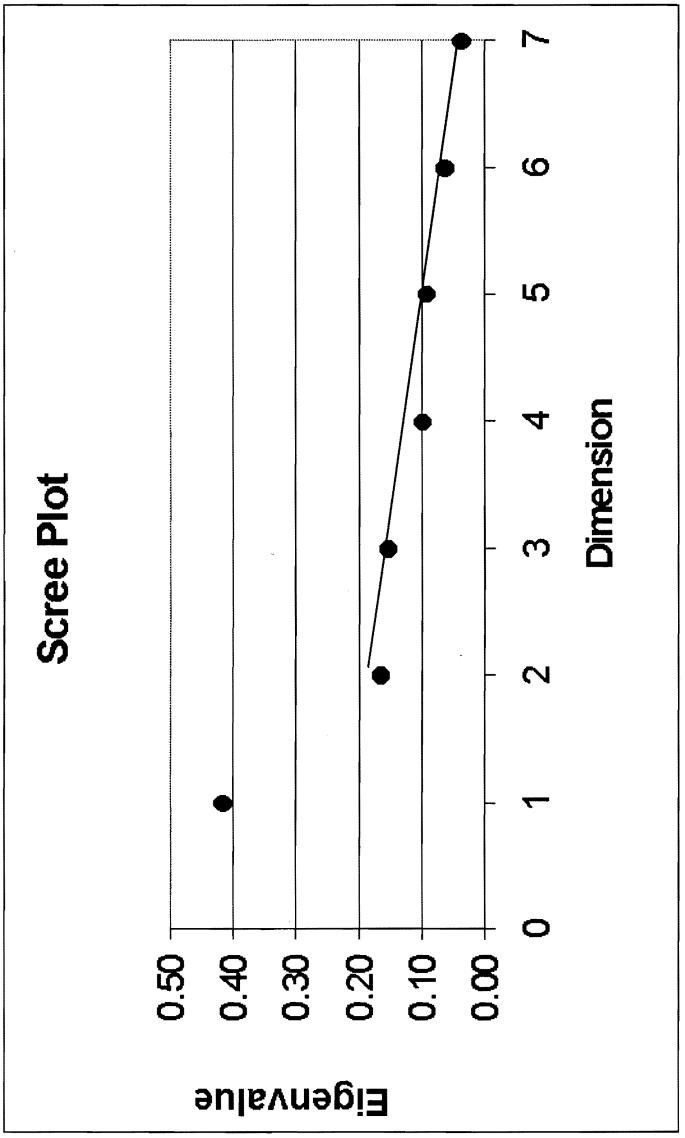
Investigating unidimensionality of the screening questionnaire in the "mail" sample, item 4: "Have you been woken by an attack of coughing at any time in the last 12 months?" and item 7: "Do you have any nasal allergies including hay fever?" have been excluded from the analysis since their discrimination coefficients were below 0.10 and this means that those items do not explain anything about latent trait variability. As a further proof in Figure 1 we have plotted items category quantifications of the screening questionnaire items. The weighted average squared distances of the category quantifications to the origin of the two-dimensional space are the discrimination coefficients. Hence, items that well discriminate between categories have their category points further apart, which in turn better separates the subjects.

Figure 1 Items category quantifications



The graphical representation of the eigenvalues extracted from the "mail" sample, excluding items 4 and item 7 (Figure 2), confirms the presence of a sole latent trait, as shown by the fact that just the first eigenvalue assumes an high value in respect to the others, whom

Figure 2 Eigenvalues extracted from the entire sample



are well fitted by a single straight line, indicating that one dimension should be retained. Therefore the latent trait underlying the ECRHS *screening questionnaire* items could be named *asthma-like symptoms seriousness* and thus we determined the subjects' *score* on the latent trait *asthma-like symptoms seriousness*. In Table 2 optimal quantifications of the screening questionnaire item responses obtained from the entire "mail" sample, with their standard errors, are reported; the last column reports optimal quantifications rearranged in a manner that the subjects' score, computed as a sum of item responses, has minimum equal to zero and maximum equal to ten.

Cross-validation study

The descriptive indices calculated in the cross-validation study are shown in Table 3. In the first part of the table the discrimination coefficients are represented, for example the item 1. "Have you had wheezing or whistling in your chest at any time in the last 12 months?" explains about the 63% of the score variability; in the central part of the table goodness of fit indices are illustrated (eigenvalue, Crombach index and Greenacre index). For example in the entire mail sample the average of the discriminant coefficients is about 0.42; Cronbach index is about 0.77, meaning that items are quite homogeneous; while the Greenacre index reveals that the relationship between items and score explains about 85% of the average inter-item squared correlations ($\overline{\Phi}^2$) reported in the last raw of the table, therefore both those indices confirm the unidimensionality found through the scree plot.

Descriptive indices computed in the jack-knifed, cross validation and long questionnaire samples are not far different from those obtained in the entire "mail" sample indicating that the scoring methodology have been cross-validated. The stability of those indices, computed starting from the optimal qualifications, shows that also the optimal quantifications are quite stable and thus could be applied to other data sets. The optimal quantification standard errors, visible in Table 2, are quite small confirming again the optimal quantifications stability.

Table 2 **Optimal quantifications and standard errors (SE) of screening questionnaire item responses for the "mail" sample, and rearranged optimal quanitifcations**

	"Yes" response (SE)		"No" response (SE)		Recoded "Yes"	Recoded "No"
Item1	2.42	(0.07)	-0.26	(0.33)	1.30	0
Item 1.1	4.34	(0.21)	-0.11	(0.01)	2.00	0
Item 1.2	2.96	(0.10)	-0.20	(0.01)	1.50	0
Item 2	1.74	(0.08)	-0.16	(0.01)	0.90	0
Item 3	1.56	(0.08)	-0.12	(0.01)	0.80	0
Item 4	0		0		0	0
Item 5	3.29	(0.15)	-0.13	(0.01)	1.60	0
Item 6	4.04	(0.24)	-0.08	(0.01)	1.90	0
Item 7	0		0		0	0

Table 3 **Items discrimination coefficients (δ) and goodness of fit indices: Eigenvalue (λ), Cronbach index (a), Greenacre index (τ) and average inter-item squared correlation ($\overline{\Phi}^2$).**

	Entire "mail" sample (n=4620)	Average of the jackknifed samples (n=4159)	Average of the cross-validation samples (n=462)	"Door-to-door" sample (n=1137)
$\delta_{\text{Item 1}}$	0,634	0,634	0,632	0,616
$\delta_{\text{Item 1.1}}$	0,469	0,469	0,475	0,499
$\delta_{\text{Item 1.2}}$	0,580	0,580	0,575	0,558
$\delta_{\text{Item 2}}$	0,272	0,272	0,268	0,252
$\delta_{\text{Item 3}}$	0,184	0,184	0,181	0,205
$\delta_{\text{Item 5}}$	0,425	0,425	0,422	0,481
δ_{Item6}	0,350	0,350	0,352	0,456
λ	0,416	0,416	0,415	0,438
α	0,766	0,766	0,762	0,786
τ	0,849	0,855	0,855	0,863
$\overline{\Phi}^2$	0,120	0,119	0,120	0,139

External Criterion Study

Treating the score as a continuos variable, the One Way Analysis of Variance considering the PD_{20} values as a criterion, gave the results visible in Figure 3. On X-axis there are the PD_{20} values, on Y-axis there are the score means of the subjects corresponding to each methacoline dose and bubbles' area is proportional to the subjects frequencies, which is reported near each bubble. For example there are 21 subjects with PD_{20} value equal to 21 and score equal to 0,46. An inverse relationship between the score and the methacoline dose is evident. In fact to increasing values of PD_{20} correspond decreasing values of the score, meaning that subjects with low scores, and thus potentially healthy, resist to high methacoline doses without having a fall in FEV_1 values.

Treating the score as a dichotomous variable for each of the 45 observed score values, among the 154 possible values, sensitivity and specificity have been computed. The best cut-off of point of the ROC curve (Figure 4) suggests the value 0.8, that represents a threshold underneath which a subject is assessed as non *asthma-like symptomatic* and above which is assessed as *asthma-like symptomatic*. With such cut-off the diagnostic test based on the screening questionnaire score has a sensitivity of 75.57 % and a specificity of 80.22%. Sensitivity is higher in respect with those determined by single items, ranging from 21.4% to 52.3%; by contrast, specificity is lower in respect with those determined by single items, ranging from 88.1% to 99.2%. Sensitivity and specificity of Toelle *current asthma* definition, with reference to clinical diagnosis available from "post consensus" sample, are 65.6% and 86.9%.

DISCUSSION

Most epidemiological studies of asthma symptoms prevalence employ written questionnaires as survey instrument. One of the most utilised is the ECRHS short screening questionnaire: it has been standardised, validated and translated into many languages.

This study shows that the ECRHS short screening questionnaire resulted to be unidimensional, that means that the items multi-

Figure 3 **Scatterplot of methacoline ordinal values vs. score means**

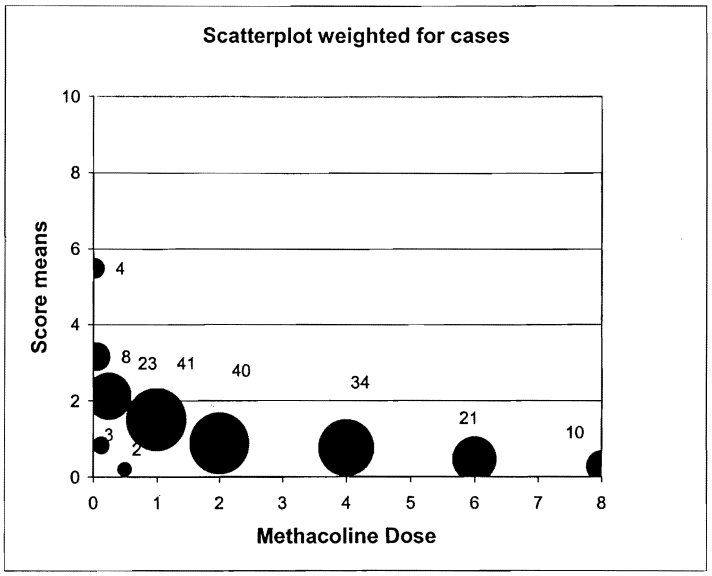
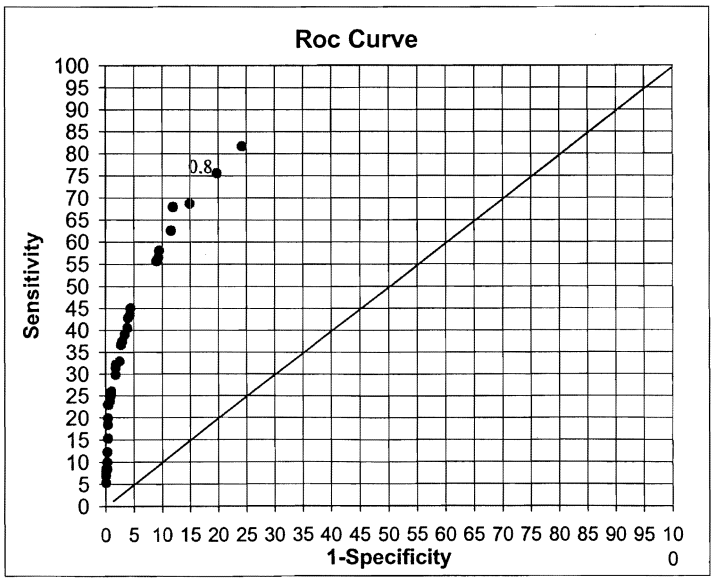


Figure 4 **Roc Curve of the screening questionnaire score**



variate relationships are explained by the relationship existing among items and a sole latent trait which we called "*asthma-like symptoms seriousness*". Unidimensionality investigation conducted to the exclusion of item 4."Have you been woken by an attack of coughing at any time in the last 12 months?" and item 7."Do you have any nasal allergies including hay fever?". From the analysis of data coming from the three Italian ECRHS centres, Pavia Verona and Torino, those items do not seem to be correlated with "*asthma-like symptoms seriousness*", which the questionnaire intended to measure, furthermore they could have been excluded even on the basis of a contents analysis since they seem to be aspecific or attributable also to other pathologies. Despite the importance of unidimensionality, there is not an accepted and effective index of the unidimensionality of a set of items, but there are many. In our study we only used descriptive indices of unidimensionality. Firstly the scree plot and Greenacre index, graphical and a numerical descriptive indices respectively, which just gives a general idea of dimensionality; then Cronbach index which is a measure of homogeneity or internal consistency and thus related to dimensionality, but its chief defect is its tendency to increase as the number of items increases [23]. Generally when Cronbach alpha of a item set is over 0.80, commonly accepted criteria lead to the conclusion that the items set is unidimensional; but in the early stages of a research a value of 0.70 will suffice [17: p 245]. In our study Cronbach index of ECRHS short screening questionnaire items is about 0.77 but considering that it has been calculated on the base of only seven items it is an appreciable value, in fact, if we add some items which are strictly correlated with the ECRHS screening questionnaire items, Cronbach index easily overcomes the value of 0.80. Once verified the unidimensionality a scoring of *asthma-like symptoms seriousness* was determined giving the possibility of graduating the pathology seriousness from an asymptomatic condition till its most serious symptomatic stage. The optimal quantifications, found through the HOMALS procedure, give now the categorical values of items' responses, but they are strictly related to our data set. The stability of descriptive indices, calculated by cross-validation study, and the optimal quantifications standard errors width

show that the optimal quantification are quite stable. Therefore a further analysis could regard the application of the optimal quantifications we found to the data of the second cycle of the ECRHS phase I (screening questionnaire), in order to further validate our optimal quantifications on a different data set. Besides it could be possible to find some fixed valued of the optimal quantifications replicable in any future survey. Finally the new scoring could be used or as a continuous variable, together with other measures of asthma symptoms and thus analysed with continuous variables techniques, or as a dichotomous variable in order to do a first selection of a sample of subjects to submit to further instrumental measurements and clinical examinations.

Treating the score as a dichotomous variable the cut-off point low value (0.8), is justified by the fact that the questionnaire is unidimensional and that the items are all quite homogeneous, and thus they all measure *asthma-like symptoms seriousness*. An interesting result is that sensitivity is almost equal to specificity (both about 80%), which constitutes a criterion in evaluating two-class (asthma-like symptomatic and asymptomatic) classification rules. Specifically, it is better to have a classification rule such that the false negative rate is equal (or almost equal) to the false positive rate (minimax error *rate* criteria: see Hand [24]).

From our classification rule a subject results non *asthma-like symptomatic* answering affirmatively to item 3."Have you been woken by an attack of asthma in the last 12 months?" or to item 6. "Are you currently taking any medicine for asthma?" were defined as symptomatic (ECRHS criterion). Sensitivity and specificity of the ECRHS criterion computed in respect with post consensus clinical diagnosis were 55% and 89% respectively. These sensitivity and specificity values are decidedly inferior to those obtained by means of our criterion.

In order to compute sensitivity of our diagnostic test we made reference to clinical diagnosis. In fact since the "gold standard" for the diagnosis of asthma seems to be clinical diagnosis, and it seems reasonable to use a respiratory group of physicians as the arbiters of what is and what is not asthma [1] we followed this procedure. For

screening purposes, sensitivity is generally preferable to specificity. If sensitivity tends to 9 (for any prevalence value). This means that the subjects identified as *"non asthma-like symptomatic"* by the decision rule are certainly the healthy subjects, while those identified *"asthma-like symptomatic"* are all the sick ones plus a quota of healthy subjects (false positives). Therefore for a two-phase diagnosis only the group of positive subjects (generally much smaller than that of the negatives) will be submitted, in a second occasion, to a more careful survey to eliminate the false positive subjects.

Sensitivity we found is higher i respect with that determined by single items of the short screening questionnaire. Sensitivity we found is also higher than that obtained for the single items in respect with the bronchial hyperresponsiveness, ranging from 27.1% to y0.8% [11]. Our sensitivity is also higher while specificity is a little lower than those of Toelle *current asthma* definition. Toelle *current asthma* defintition, based on questionnaire responses and on instrumental examinations, is quite complex to implement. On the contrary the method we suggest, based only on the ECRHS short screening questionnaire, offers several advantages: it is cheap, convenient, requiring no special equipment, and conducts to appreciable results. There fore we suggest to simply using the ECRHS short questionnaire in order to discriminate between symptomatic and asymptomatic subjects.

Treating the score as a continuous variable, since it is inversely correlated with the methacoline and thus it measures the *asthma-like symptoms seriousness*, it can be used as any other continuous variable in asthma surveys.

Therefore these findings are of special interest if one considers the cost of epidemiological studies that involve medical examinations and numerous instrumental tests.

ACKNOWLEDGEMENTS

We wish also to thank the Italian group of ECHRS to have made available the data of the research. The members of ECRHS group are the following:

ECRHS- Italy: University of Verona: de Marco R, Lo Cascio V, Campello C, Rossi F, Biasin C, Cannistrà A, Cenci B, Destefani E, Ferrari M, Girotti M, Lamprotti G, Martini C, Olivieri M, Poli A, Tardivo S, Verlato G, Villani A, Zanolin ME; **University of Pavia:** Marinoni A, Cerveri I, Alesina R, Basso O, Berrayah L, Brusotti R, Fanfulla F, Moi P, Zoia MC; **University of Turin:** Bucca C, Romano C, Aime M, Cerutti A, Chiampo F, Gallo G, Rola, Sulotto F; **ASL 42 Pavia:** Casali L, Fratti C, Karytinos P; **ASL 7 Torino:** Bugiani M, Arossa W, Caria E, Carosso A, Castiglioni G, Piccioni P.

REFERENCES

1. Burr ML. Is Asthma increasing? *J. Epidemiol Community Health* 1987; 41: 185-189.
2. Anderson HR. Increase in hospital admissions for childhood asthma: trends in referral, severity, and readmissions from 1970 to 1985 in a health region of United Kingdom. *Thorax* 1989; 44: 614-619.
3. Weitzman M, Gortmaker SL, Sobol AM, et al. Recent trends in the prevalence and severity of childhood asthma. *JAMA* 1992; 268: 2673-2677.
4. Balfe D, Crane J, Beasley R, Pearce N. The worldwide increase in the prevalence of asthma in children and young adults. *Continuing Med Education J* 1996; 14: 433-442.
5. Anderson HR, et al Trends in prevalence and severity of childhood asthma. *BMJ.* 1994; 308 :1600-1604.
6. Woolcock A.J., Peat J.K. Evidence for the increase in asthma

worldwide. *Ciba Found Symp* 1997 ; 206: 122 - 134; discussion 134-139, 157-159. Review.

7. Crater SE, Platts-Mills TA. Searching for the cause of the increase in asthma. *Curr Opin Pediatr* 1998; 10(6): 594-599
8. Burney PGJ et al. Validity and repeatability of the IUATLD (1984) Bronchial Symptoms Questionnaire: an international comparison. *Eur Respir J* 1989; 2: 940-945.
9. Rijcken B., Schouten J.P. Measuring bronchial responsiveness in epidemiology. *Eur Respir J* 1993; 6: 617-618. Editorial.
10. Burney PGJ, Chinn S. Developing a New Questionnaire for Mea suring the Prevalence and Distribution o f Asthma. *Chest* 1987; 91(6): 79-83. Supplement.
11. Galobardes et al. Effect of the method of Administration, Mail or Telephone, on the Validity and Reliability of a Respiratory Health Questionnaire. The Spanish Centers of the European Asthma Study. *J Clin Epidemiol* 1998; 51(10): 875-881.
12. Burney, PGJ; Luczynska. C; Chinn, S; Jarvis, D. The European Community Respiratory Health Survey. *European Respiratory Journal* 19 94; 7: 954-960.
13. Gifi, A. *Nonlinear Multivariate Analysis*. Wiley, New York, 1990.
14. Cattell, R.B. The screen test for the number of factors. *Mult. Behav. Res*. 1966; 1: 245-276
15. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951; 16: 297-334.
16. Greenacre, MJ. *Correspondence Analysis in Practice*. Academic Press, London, 1993.
17. Nunnally, J.C. *Psychometric Theory*. McGraw-Hill, New York, 1978.
18 Morrison, D. F. *Multivariate statistical methods*. New York: McGraw-Hill, 1967.
19. Sprent, P. *Data driven statistical methods*. London: Chapman & Hall, 1998
20. Tanner and Swets,, *Psycologiacal Review* 1954.
21. Cerveri I et al. Can clinical judgement be a gold standard for identifying bronchial asthma in epidemiological studies?

Interobserver agreement and contribution of specific diagnosis procedures. *Monaldi Arch Chest Dis* 1998 (in press).

22. Brett G et al. Toward a Definition of Asthma for Epidemiology. *Am Rev Respir Dis* 1992; 146: 633-637.
23. Hattie J., Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement* 1985, 9: 139-164.
24. Hand DJ. *Construction and assessment of classsification rules*. Wiley, New York, 1997.

APPENDIX A. Short Screening Questionnaire Administered in the First Phase of the ECRHS

---

1. Have you had wheezing or whistling in your chest at any time in the last 12 months?
   IF NO GO TO QUESTIONS 2, IF YES:
   - 1.1 Have you been at all breathless when wheezing noise was present?
   - 1.2 Have you had this wheezing or whistling when you did not have a cold?
2. Have you been awakened with a feeling of tightness in your chest at any time in the last 12 months?
3. Have you been awakened by an attack of shortness of breath at any time in the last 12 months?
4. Have you been awakened by an attack of coughing at any time in the last 12 months?
5. Have you had an attack of asthma in t he last 12 months.
6. Are you currently taking any medicine (including inhalers, aerosols, or tablets) for asthma?
7. Do you have any nasal allergies including hay fever.

JOURNAL OF OUTCOME MEASUREMENT®, 4(4), 763-793

Measurement Properties of the Symptom Impact Inventory

Arlene Michaels Miller, PhD, RN

JoEllen Wilbur, PhD, RN, FAAN,

Andrew Montgomery, PhD

Peggy J. Chandler, PhD

Nikolaus Bezruczko, PhD [1]

University of Illinois at Chicago

The purpose of this study is to evaluate the measurement properties of the Symptom Impact Inventory using both psychometric and Rasch analyses. This inventory is designed for generally healthy midlife women. The sample included 340 midlife women aged 45-65 representing two studies. The first study involved Black and White employed sedentary women (n = 161) who volunteered for a walking intervention. The second study of migration and health included women who were recent immigrants from the former Soviet Union (n = 179). The women reported experiencing an average of 13.44 symptoms (S.D.=7.88) with a range of 1 to 32. Principal components analysis identified 5 components in this sample. Rasch measurement analysis found excellent model fit for the Symptom Impact Inventory with only 2 symptoms, *Decreased appetite* and *Decreased sexual desire or interest,* unstable in scale dimensionality analyses. Person and item parameters were reliable, and comparisons with groups known to differ on symptom reporting provided substantial validity. Although the two sample groups differed significantly on most demographic characteristics, a cross-cultural comparison found the scale structure remarkably robust.

Requests for reprints should be sent to Nikolaus Bezruczko, Ph.D., 1524 E. 59th St., Chicago, IL 60637

INTRODUCTION

Symptoms are subjective phenomena often regarded as indicators of conditions that depart from normal function, sensation, or appearance. Symptoms may be nonspecific or distinctly pathognomic and their character may or may not be related to underlying physiologic processes. Researchers generally agree that complex interactions among biological, psychological, and social factors influence symptom experience, as well as their interpretation. Contemporary researchers use symptom scales and inventories throughout the health sciences.

Literature review

One early instrument to focus on symptom measurement was the Symptom Distress Scale (McCorkle, 1987), later modified for cancer and heart disease patients (Samarel, et al., 1996; Sarna, 1998). This instrument collects symptom frequency, intensity, and distress ratings that are summed to represent total symptom experience. Researchers have developed other general-purpose health-related symptom instruments such as the Sickness Impact Profile and the Nottingham Health Profile (see Hunskaar & Vinsnes, 1991; Grimby et al., 1993). The emphasis in the present analyses involves symptoms surrounding menopause and normative aging.

Inventories and questionnaires focused on menstrual symptoms often collect specific information concerning timing in the cycle, severity, comparison with peers, as well as alleviating factors (Hightower, 1997; Mitchell & Woods, 1996; Taylor, 1994; Woods, et al., 1997). Scale precision of these menstrual symptom inventories, as well as a measure of the impact of scores on women's normal activities, however, is unknown. While these inventories commonly assess a broad range of symptoms, they are primarily validated with small, homogenous samples and/or include women recruited primarily from clinical populations (Holte & Mikkelson, 1991; Hunter, Battersby & Whitehead, 1986), (Perz, 1997).

Substantial research has shown menstrual symptom surveys

are important for investigating symptom prevalence, relationships among menopausal symptoms, as well as influence on midlife women's health (Bungay, Vessey & McPherson, 1980; Neugarten & Kraines, 1965; Jaszmann, Van Lith & Zaat, 1969, McKinlay & Jeffreys, 1974). Recent symptom research has examined presence or absence of direct hormonal relationships, while other symptom studies have identified specific complexes or patterns in peri- and post-menopausal women. Woods, et al. (1997) describe persistence of stress-related premenstrual syndrome (PMS) symptom patterns in women over age 40, while Mitchell and Woods (1996) describe symptom clusters that may differ in the early, middle and later menopausal transition. Studies have established an unequivocal relationship between fluctuation of estrogen levels and vasomotor instability (hot flashes and night sweats), and between estrogen decline and vaginal changes (Holte & Mikkelson, 1991; Hunter, Battersby, & Whitehead, 1986). Evidence for direct relationships between hormonal changes at menopause and other commonly reported symptoms, however, remains inconclusive.

Using a 22-item symptom inventory, Wilbur, Miller, Montgomery, and Chandler (1998) found only hot flashes were significantly related to serum levels of estradiol and FSH in 35-69 year old women. They found no relationships among hormone level or age and vaginal dryness, trouble sleeping, or cognitive symptoms. Similarly, Matthews, et al. (1990) concluded that changes in psychological characteristics were minimal across a 2½-year longitudinal study of midlife women, and no relationships were found between menopausal status and depression, nervousness, sleeplessness, or neck and skull aches. Avis, et al. (1994) found psychological symptoms appear to increase when the peri-menopausal period is lengthy, but menopause onset *per se* is not associated with increased depression. On the contrary, Busch, Zonderman, and Costa (1994) identified depression *decline* over time, and no relationship between distress and menopausal transition. In fact, many studies suggest anxiety and depressed mood symptoms are more strongly related to stressful life events, role change, and loss of significant social relationships than to hormonal changes at menopause (Miller, Wilbur, Montgomery, &

Chandler, 1998; Mitchell & Woods, 1996). Although changes in hormonal patterns may contribute to sleep disorders (Shaver, et al., 1988), no direct relationship has been established between hormonal changes and joint pains, backaches, or fatigue. Cross-sectional studies have demonstrated, however, that these symptoms are frequently experienced by midlife women (Wilbur, Miller, & Montgomery, 1995; Mitchell & Woods, 1996).

This rapidly growing knowledge of midlife women's health would benefit substantially from inventories that not only survey or account for symptoms but operationally define a variable that measures the impact of these symptoms on women. While inventories may show women reliably vary in their ratings of symptom frequency, severity, and bothersomeness, they do not represent the cumulative impact of the symptom experience nor express their combination as unidimensional impact measures. These shortcomings are important, because objective, equal interval scales would facilitate group comparison, as well as measure the magnitude of pre- and post-treatment effects (see Wright & Linacre, 1989). All symptom instruments in current use yield only ordinal raw scores. A linear measure of symptom impact would provide a major advance in midlife women's health research.

Another issue concerns the sensitivity of inventory items to changes in the symptom experience. Research, for example, has demonstrated conflicting results between physical activity and vasomotor, somatic and psychological symptoms in midlife women (Guthrie, Smith, Dennerstein, & Morse, 1995; Li, Holm, Gulanick, Lanuza & Penckofer, 1999, Sternfeld, Quesenberry, & Husson, 1999; Wilbur, Holm & Dan, 1992). We speculate this discrepancy may be due in part to inventories and surveys that are unable to detect subtle changes in symptom experience. Equal interval measures on absolute scales should provide greater symptom sensitivity than traditional psychometric raw score analyses.

A final issue concerns the weak validity of symptom inventories in general, and menopausal symptoms in particular. In our review, we found only correlational studies asserting significant relationships between scores/ratings and criterion variables. These stud-

Table 1 **Demographic variables for the overall sample and by sample group**

	Overall Sample (n=340)		Women's Walking Program (n=161)		Migration & Health Project (n=179)	
	n	%	n	%	n	%
Race*			Chi-square = 84.30, df=1, p<.001			
White	278	82	99	62	179	100
Black	62	18	62	38		
Age group*			Chi-square = 133.26, df=3, p<.001			
45-49	126	37	101	63	25	14
50-54	54	16	36	22	18	10
55-59	80	23.5	18	11	62	35
60-65	80	23.5	6	4	74	41
Marital status*			Chi-square = 15.24, df=1, p<.001			
Married	201	59	73	45	128	72
Not married	128	36	72	45	51	28
missing	16	5	16	10		
Number of children*			Chi-square = 45.63, df=3, p<.001			
none	39	11	25	16	14	8
one	101	30	32	20	69	38
two	147	43	55	34	92	52
more than two	37	11	33	20	4	2
missing	16	5	16	10		
Education*			Chi-square = 144.45, df=4, p<.001			
HS grad or less	11	3	7	4	4	2
Some college	51	15	40	25	11	6
College grad	76	22	23	14	53	30
Grad/Prof school	122	36	15	9	107	60
Grad/Prof degree	63	19	59	37	4	2
missing	17	5	17	11		
Income*			Chi-square = 161.21, df=2, p<.001			
<$20,000	118	35	5	3	113	63
>$20,000, <$50,000	127	37	64	40	63	35
>$50,000	76	22	73	45	3	2
missing	19	6	19	12		
Menopausal status*			Chi-square = 51.01, df=3, p<.001			
Premenopausal	89	26	69	42	20	11
Perimenopausal	48	4	25	16	23	13
Postmenopausal	163	48	58	36	105	59
Hysterectomy	40	12	9	6	31	17

* significantly different by group

ies are largely ad hoc empirical attempts to support symptom inventories or checklists unguided by any strong theoretical perspective (see Alder, 1998 for critique of menopausal inventories). None demonstrated that symptom items define a quantitative variable with plausible conceptual relationships to midlife women's menopausal experience.

Purpose

The purpose of this study is to evaluate the measurement properties of the Symptom Impact Inventory using both psychometric and Rasch analyses. The Rasch model, in particular, will be used to assess 1) delineation of the symptom impact construct, 2) item cohesion and scale invariance, and 3) impact measure precision. Validity of the obtained symptom impact measure is then examined by sample group, age, menopausal status, and depression scores.

METHOD

Sample

The overall sample is 340 women, aged 45-65 (Mean = 53.61, SD = 6.08). This sample contrasts two groups. The first group involves 161 women (47%) participating in the Women's Walking Program, an exercise intervention for healthy mid-life Black (38.5%) and White (61.5%) women (Wilbur, Miller, Chandler, & Montgomery, 1998). The second group includes 179 women (53%) participating in the Migration and Health Project, a cross-sectional study of the effects of immigration and resettlement on the health and psychological well-being of midlife women from the former Soviet Union (Miller, Wilbur & Chandler, 1999). Table 1 summarizes demographic data for the overall sample and these two groups.

These groups were selected to contrast profoundly different symptom experiences. Women immigrating from the former Soviet Union are known to experience significantly more symptoms with greater impact than generally healthy U.S. women (Aroian, Spitzer & Bell, 1996). For example, midlife Russian women show higher

rates of obesity, hypertension, hyperlipidemia, as well as anemia, gastrointestinal, and gynecological disorders than comparable U.S. women (Smith, 1995; Kohlmeier, et al., 1998). Prevalence of depression, anxiety, and other stress-related disorders are also high (Cwikel, et al., 1997; Aroian & Norris, 1999). Table 1 illustrates the significant differences between the two sample groups on race, age, marital status, number of children, education, income, and menopausal status.

The Women's Walking Program. The purpose of Women's Walking Program was to determine the effectiveness of a 24 week home-based, moderate intensity walking program to improve physical health and alleviate common physical and psychological symptoms. Women were recruited through newspaper advertisements, university e-mail notices, posters, and local television news coverage. Enrollment criteria included only healthy Black or White, employed, sedentary women not on hormone replacement therapy who volunteered to participate in an exercise intervention. Mean age of the women in the Women's Walking Program was 49.64 (SD = 4.25), approximately 50% were currently married, 83% had at least one child, 83% of the women reported household incomes of greater than $30,000, and all women worked outside home. The Symptom Impact Inventory was administered among other baseline measures.

Migration and Health Project. The purpose of this cross-sectional study was to examine the health effects of immigration at midlife on women from the former Soviet Union. They were recruited through Russian-language newspaper advertisements, English as a second language classes, posters in neighborhood businesses, and recommended by other women in study. Eligibility criteria included having immigrated after age 40 from the former Soviet Union to the U.S. within past five years. Approximately 50% had lived in the U.S. for less than 2 years (Mean years in U.S. = 1.95; SD = 1.69). Mean age of the women was 57.18 years (SD = 5.24). Seventy two percent were married, 92 percent had at least one child. Household income of less than $30,000 was reported by 88% of the women, and only fifteen percent worked outside home.

Instruments

Symptoms. The Symptom Impact Inventory surveys 34 physical and psychological symptoms. Consistent with Larson's (1994) symptom management model in which perception, evaluation, and responses dynamically interact, the Symptom Impact Inventory collects ratings of symptom frequency, severity, and bothersomeness. Women rated the frequency (0=never, 1=occasionally, 2=sometimes, 3=often), severity (1=mild, 2=moderate, 3=severe), and bothersomeness (0=none, 1=not much, 2=somewhat, 3=a lot) of each symptom that occurred during the previous two weeks. A linear impact score was obtained by summing these three items for each symptom. The symptoms were adapted from a 28-item list developed by Kaufert and colleagues (Kaufert & Syratuik, 1981; Kaufert, Gilbert & Hassard, 1988) and used by the investigators in a prior study of 200 mid-life women in four occupations (Wilbur, Miller & Montgomery, 1995). Five items were added based on a literature review of instruments measuring menstrual, menopausal and other common mid-life symptoms (Neugarten & Kraines, 1965; Woods, Most & Dery, 1982). The sixth item, *Difficulty thinking of the right word,* representing word retrieval, was added to assess mid-life cognitive changes (Hoyer & Rybash, 1994).

For women from the former Soviet Union, the Symptom Impact Inventory was translated into Russian using the translation/back translation method (Brislin, 1990). Focus groups were held to validate conceptual equivalence and appropriateness of the translated instrument. A pilot study examined administrative feasibility of this instrument. It was then embedded in a larger questionnaire, and self-administered under supervision of bilingual research assistants in participants' homes or convenient meeting places.

Depression. To establish validity of the Symptom Impact Inventory, a measure of depression was included. The Center for Epidemiological Studies Depression Scale (CES-D) is a 20-item questionnaire developed to assess depressive symptomatology with emphasis on mood in the general population (Radloff, 1977). Partici-

pants respond to questions about how they felt or behaved in the past few weeks on 4-point Likert scales (0-3). Items are coded to reflect increasing severity and are summed for a total score. Higher scores indicate more depressed mood. Validity is supported by a significantly higher percentage of psychiatric patients (70%) than the general population (21%) who score at or above an arbitrary cutoff score of 16. For purposes of this study, women with a score of less than 16 were identified as not depressed and with a score of 16 or more as depressed. CES-D internal consistency is reported to be 0.85 for the general population and 0.90 for the psychiatric inpatient sample (Radloff, 1977). Other research has found the CES-D valid and reliable across age, sex, and racial groups, and it has been used successfully with midlife women (Kaufert, 1994; McDowell & Newell, 1996; Wilbur, Miller & Montgomery, 1995).

Menopausal status. To establish known groups validity, a measure of menopausal status was included. Menopausal status was established by questions derived from the Massachusetts Women's Health Study, which survey menstrual regularity over the prior 12 months (McKinlay, McKinlay & Avis, 1989). Women are classified premenopausal if they report regular menses during the preceding 12 months; perimenopausal if they report menses during the preceding 12 months but with intermittent amenorrhea, menstrual irregularity or changes in flow; and postmenopausal if they report no menses during the preceding 12 months and had no surgical menopause. Women with surgical menopause were classified in the fourth group, hysterectomy. In a previous study, these self-report classifications were shown to cross-validate with serum levels of follicle stimulating hormone (FSH) and estradiol (Wilbur, Miller & Montgomery, 1995).

Statistical Procedures

Psychometric analyses include descriptive statistics, chi-square, inter-item correlations, Cronbach's alpha, and principal components analysis. Descriptive statistics were calculated and compared across the two groups of women (Women's Walking Program; Migration

and Health Project). Positive inter-item correlations support internal consistency, while coefficient alpha provides some evidence of ordinal scale reproducibility. Principal components analysis was performed to identify relationships among symptoms.

Our goal, however, is interval level symptom measurement or scale stability. Thus, primary emphasis is on the conformity of ratings to quantitative requirements for linear measurement when both the person and the item parameters are estimated by the Rasch model (Wright & Masters, 1982, see also Wright & Stone, 1979). In the Rasch model, parameters are estimated for each person (B_n) and item (D_i) and their differences (B_n - D_i) are compared against unidimensional measurement model predictions. In this application, the parameters were estimated with Winsteps software (Linacre, 2000) which implements an unconditional maximum likelihood procedure (Wright & Panchapakesan, 1969). The estimated parameters are conditioned on total ratings, and an iterative construction of conjointly additive scale units. These fundamental properties are currently only available in Rasch models. Model fit is evaluated for both women and items, comparing observations with modeled expectations, and the obtained residual is standardized by its variance, squared, then averaged separately over women and items, respectively.

This approach represents an important shift from specific sample characteristics common in traditional "true score" analyses to differences between person and item parameters and their objective magnitude. When quantitative properties of these differences conform to the one-parameter logistic, person and item estimates are measures on an absolute scale. A related property of this difference (person minus item) is a systematic accumulation of scale units across the interval scale further supporting scale linearity. When differences deviate from model expectations, they are examined by fit statistics and principal components residual factor analysis (Linacre, 1996; 1999). Standard errors for each person and item along the entire variable provide maximum scale precision. These properties are extremely important when developing instruments with the sensitivity to measure change. (More technical explanations can be found in

Figure 1 **Map of women and items on symptom impact variable**

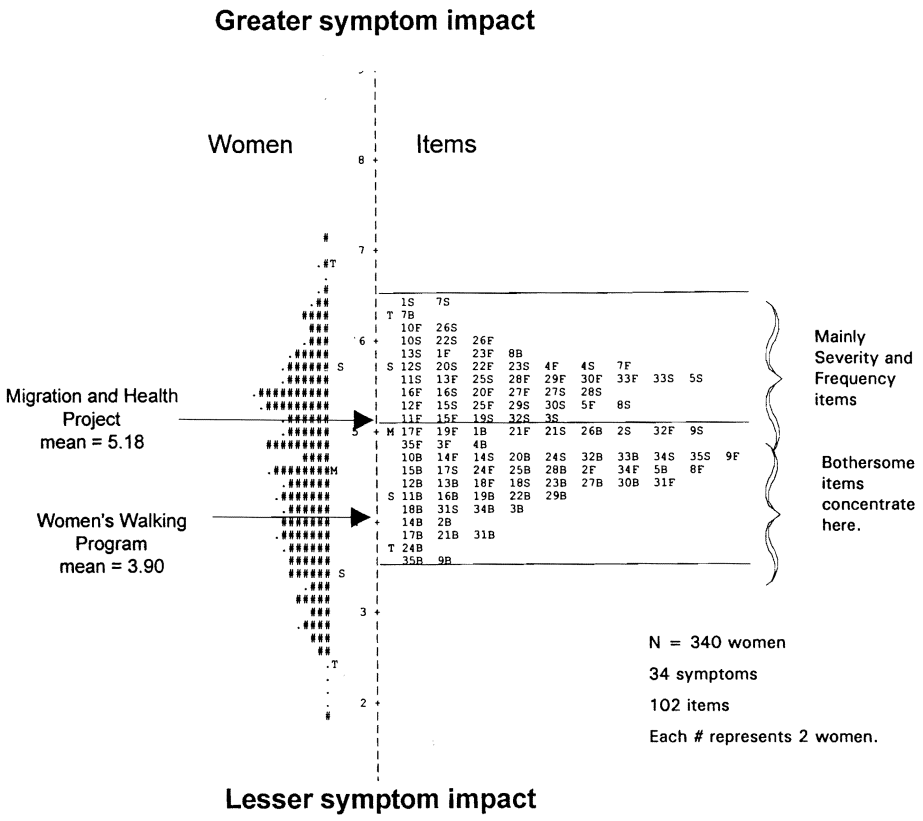


Figure 1. Map of women and items. Numbered vertical line is an equal interval logit scale with scale origin set at lowest obtained person measure. Item abbreviations are along the right side and women are distributed along the left. Women reporting few symptoms appear at the bottom, while those with many symptoms at top. Going up the scale, symptoms become progressively less common but their occurrence is cumulative, i.e., women high on scale tend to report the symptoms below. Item abbreviations refer to number on the questionnaire, and F, S, B indicate Frequency, Severity, and Bothersomeness, respectively. Bothersomeness items appear along the lower section, while Severity and Frequency items are intermingled in the top. Effect size = 1.33 SD units for Women's Walking Program versus Migration and Health Project comparison.

Wright & Masters, 1982; and Wright & Stone, 1979).

Finally, validity was assessed by examining differences in the Rasch model symptom impact measure by sample group, age, menopausal status and depression score using analyses of variance and effect sizes. A Pearson correlation coefficient between Rasch symptom impact measures and total depression scores was calculated.

RESULTS

Psychometric analyses

The mean number of symptoms reported by the women was 13.44 (S.D. 7.88) with a range from 1 to 32 (women who reported no symptoms were not included in these analyses). The number of symptoms varied significantly by sample group ($F = 177.37$; df 1, 338; $p<.001$); the Migration and Health Project women (mean 17.82, S.D. 6.72) reported more symptoms than those from the Women's Walking Program (mean 8.57, S.D. 6.01). Each symptom was reported as experienced during the past two weeks by at least 35 (10%) women. Most commonly reported symptoms were *Aches and stiffness in muscles and joints* (73%), *Fatigue or lack of energy* (72%), *and Irritability* (63%). The proportion of women reporting each of these was significantly different by sample group (chi-squares = 14.82, 77.55, and 65.72 respectively) with more women from the Migration and Health Project reporting that they experienced each symptom than women from the Women's Walking Program. The least commonly reported symptoms were *Heavy menstrual periods* (10%), *Pain during sexual intercourse* (11%), and *Vaginal irritation* (12%). There was no significant difference in the proportion of women reporting these symptoms by sample group.

Cronbach's alpha for frequency of symptoms was 0.932 (n of cases = 340, n of items = 34) and for the linear impact score it was 0.938. Item to total correlations were all significant and strong (>.30) except for *Heavy menstrual periods,* which was .023. For this rea-

Figure 2 **Content structure of a symptom impact variable**

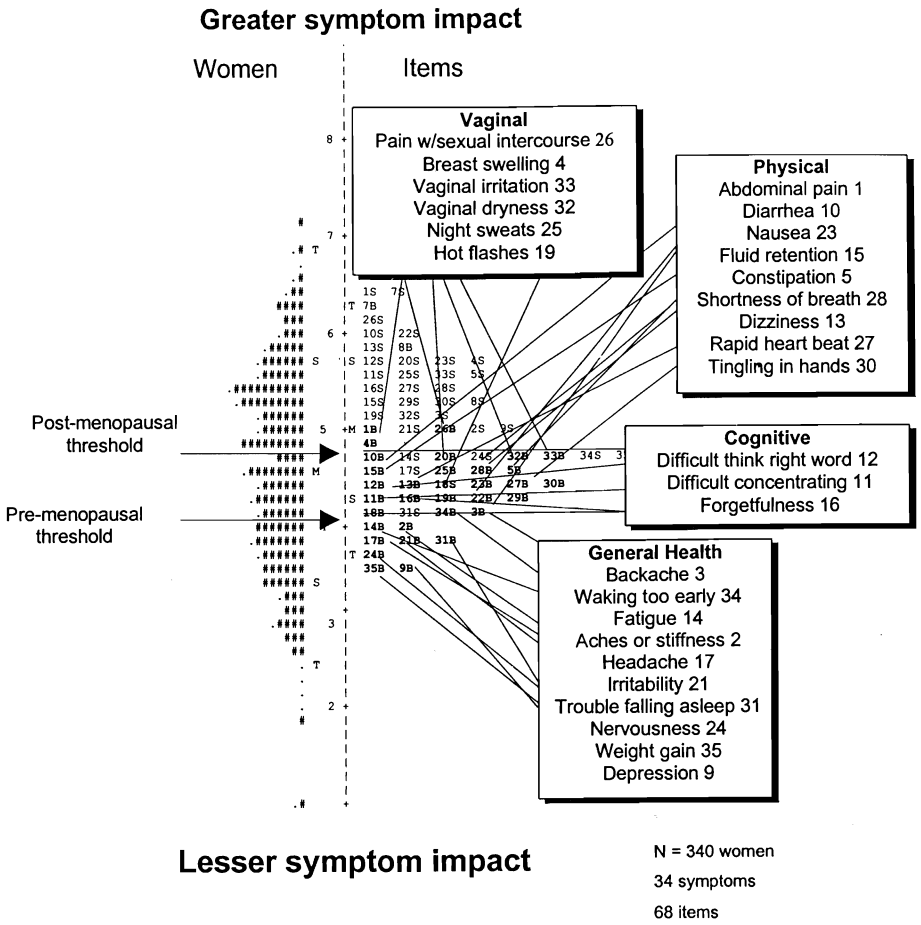


Figure 2. Content structure of Bothersomeness items. Map shows only Severity and Bothersomeness items. Call outs contain the symptoms with their item number and refer only to Bothersomeness items. Scale origin is located at the lowest obtained person measure. The Impact construct is defined by General Health symptoms in the lower region of the scale, and three separate clusters higher on the structure, Cognitive, Vaginal, and Physical symptoms. Effect size = 0.56 SD units between Pre- and Postmenopausal thresholds.

son, that symptom was dropped before performing a principal components analysis on the remaining symptoms (Kerlinger, 1973).

The best and most understandable fit using principal components analysis identified 5 components that explained 53.89% of the variance. Factor 1 included 18 symptoms that could be described as psychological and general health since they ranged from *Depression* and *Difficulty concentrating* to *Shortness of breath* and *Aches or stiffness*. Factor 2 (3 symptoms) could be described as gastrointestinal since it included *Diarrhea, Nausea*, and *Abdominal pain*. Factor 3 (4 symptoms) was made up of *Hot flashes, Night sweats, Constipation*, and *Loss of urine when coughing*. Factor 4 (4 symptoms) contained *Vaginal dryness, Pain during sexual intercourse, Vaginal irritation,* and *Decreased sexual desire or interest*. Factor 5 (4 items) contained *Increased appetite, Weight gain, Decreased appetite* (negative loading) and *Skin problems*. Scores for each component were calculated as the sum of the linear impact score and compared across sample groups. For each component, women from the Migration and Health Project had significantly higher scores than the Women's Walking Program (Component 1 $F = 287.69$; Component 2 $F = 40.22$; Component 3 $F = 74.36$; Component 4 $F = 18.74$; Component 5 $F = 34.78$; df 1, 338; $p < .001$). In spite of the known differences between the two sample groups of women, the principal components analysis was not run by sample group since neither was large enough to provide the recommended 10 subjects per symptom. Since principal components analysis is population based, the sample group differences may affect the replicability of the components.

Rasch measurement analyses

The Rasch model transforms nonlinear raw data such as Likert ratings to equal interval log odds units or logits. Figure 1 graphically presents results from operationally defining the symptom impact variable with 102 items (34 symptoms with 3 responses - frequency, severity, and bothersomeness for each symptom). For this analysis, scale center, 0, was set at the lowest woman impact measure to eliminate negative measures. More commonly reported items are in the

Figure 3 **Item plots for frequency, severity , and bothersomeness**

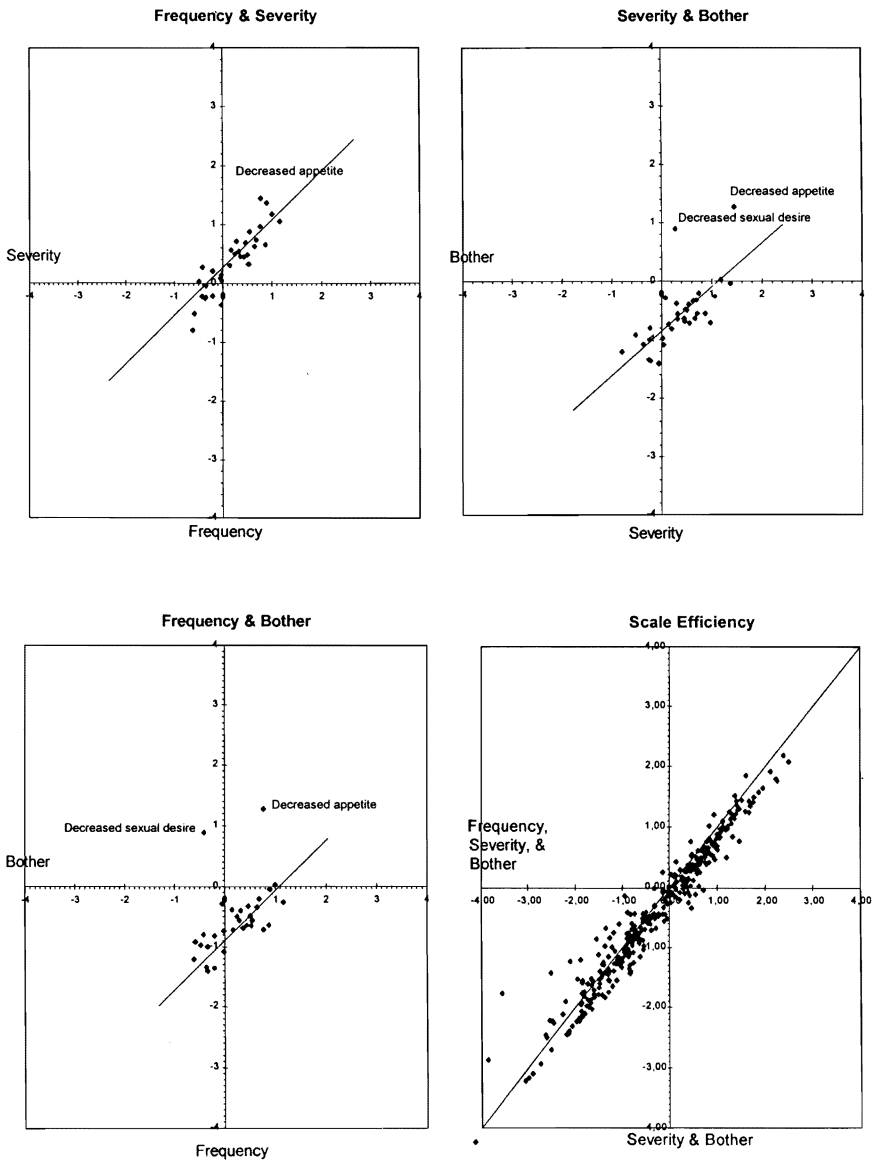


Figure 3. Item plots for Frequency, Severity, and Bothersomeness. All values are in logits.

Figure 4 Sample Group Comparison

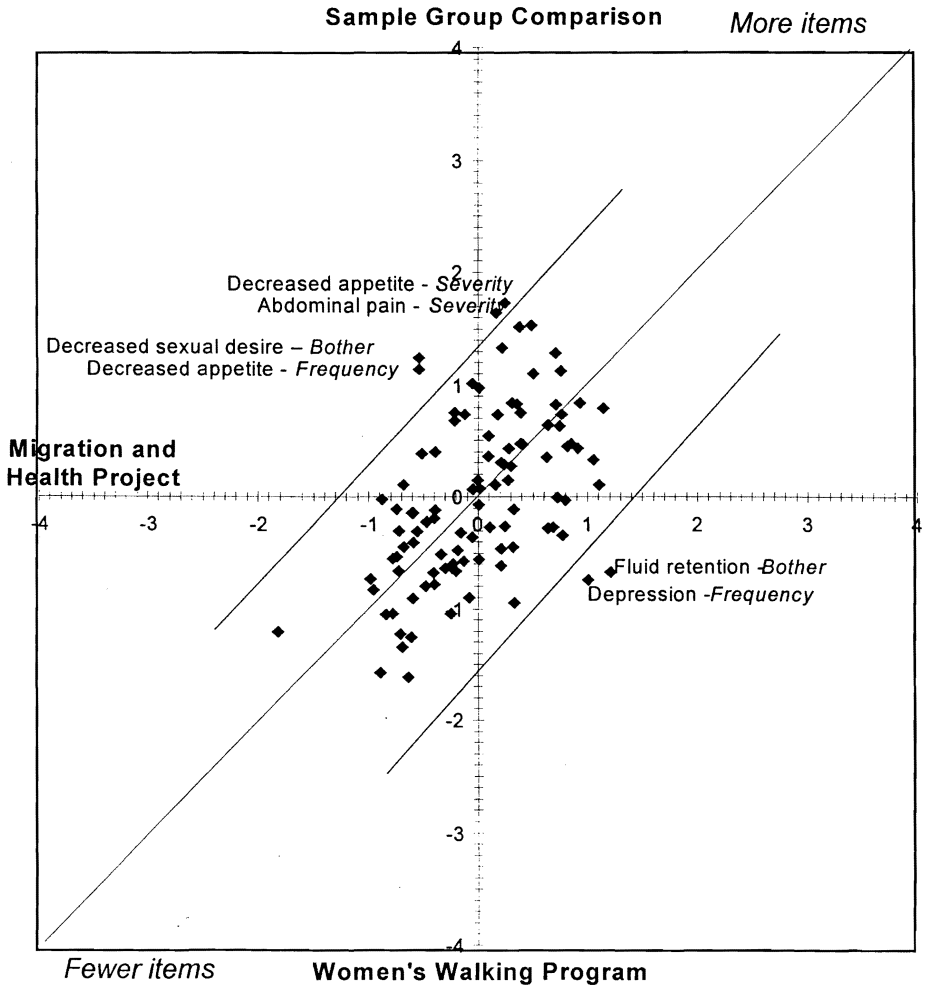


Figure 4. Item plot for two sample groups, Women's Walking Program and Migration and Health Project. One hundred and two items (one each of Frequency, Severity, and Bothersomeness for 34 symptoms) plotted for 340 women from two sample groups. Scale units are in logits. Bands indicate 99 percent quality control lines to test invariance of the symptom structure.

lower region, and women indicated their lesser impact, i.e., assigned them lower ratings. Bothersome ratings always appear lower than the corresponding frequency or severity ratings. The upper region shows greater symptom impact, but women tend to report them simultaneously with lesser impact symptoms. *Abdominal pain* and *Diarrhea*, for example, are relatively rare (only 35 and 13 percent in this sample); their frequency, severity, and bothersome ratings are not high; but women tend to report them simultaneously with lesser impact symptoms. Overall item reliability is 0.94.

Rasch analysis also reports impact of these symptoms on women in common logit units. Women with fewer symptoms appear at the bottom of the scale and those with many at the top. For this sample, some women are below scale floor indicating very low symptom impact. As expected, the mean for the Migration and Health Project women is significantly higher than that of the Women's Walking Program women confirming a greater symptom impact. Overall sample reliability is 0.89.

Fit analysis. Both women and items fit the Rasch model remarkably well. Infit and outfit mean squares for women were both 0.96 with standard deviations of 0.40 and 0.47, respectively. Likewise, mean squares for items were 1.03 and 1.04 with standard deviations of 0.23 and 0.27, respectively. All standardized values were near expectations (Mean = 0, SD = 1). The only symptoms showing unusual fit values are *Decreased appetite* (mean square fit values: Bothersomeness infit = 1.52, outfit = 1.72; Frequency infit = 1.26, outfit = 1.56) and *Decreased sexual desire or interests* (Bothersomeness infit = 2.32, outfit = 2.44; Frequency infit = 1.37, outfit 1.59). In addition, a few isolated items also showed large fit values: *Diarrhea* (Bothersomeness infit and outfit both = 1.70), *Hot flashes* (Frequency infit = 1.39, outfit = 1.54), *Abdominal pain* (Frequency outfit = 1.53) and *Increased appetite* (Frequency infit = 1.39, outfit = 1.44).

Principle components factor analysis of model residuals showed less than 10% systematic variation, suggesting that symptom impact uniquely defines a unidimensional scale. Analysis of disaggregated Migration and Health Project and Walking Program

Table 2 **Rasch model impact measures in mean logit and standard deviation by sample group, age, menopausal status, depression scores coded, and the significance of the differences between the groups as determined by analysis of variance and effect sizes.**

	n	Symptom impact measure Mean (S.D.)	F	p	Effect size
Sample group			150.07	<.001	1.33
Women's Walking	161	3.90 (0.92)			
Program	179	5.18 (1.00)			
Migration and Health Project					
Age group			26.87	<.001	1.23
45-49	126	4.04 (1.04)			
50-54	54	4.28 (0.92)			
55-59	80	4.91 (1.20)			
60-65	80	5.26 (0.94)			
Menopausal status			7.89	<.001	0.87
Premenopausal	89	4.09 (1.05)			
Perimenopausal	48	4.57 (0.98)			
Postmenopausal	163	4.77 (1.19)			
Hysterectomy	40	4.81 (1.14)			
Depression scores			193.97	<.001	1.56
Not depressed	158	3.85 (0.96)			
Depressed	164	5.30 (0.90)			

Note. Effect size is calculated as the difference between largest and smallest mean impact measures divided by their pooled standard deviations so that $(M_x\text{-}M_n)/((SD_x + SD_n)/N)$ where N is number of groups.

women, however, confirm that scale effectiveness is optimal when the mean symptom impact is near the scale center.

Scale content. Figure 2 illustrates the arrangement of the bothersomeness item of symptom groups along the impact variable. General Health symptoms appear at the bottom of the scale. Another cluster, slightly higher up, could be labeled Cognitive. A third cluster, Physical, is intermingled with Vaginal symptoms. Average logit scores by menopausal status can also be identified on this scale. As expected, pre-menopausal women report significantly lower symptom impact than post-menopausal women.

This scale is highly reliable. Finding an invariant content structure such as this is an important source of support for validity, evidence commonly neglected in traditional scale development.

Dimensionality. With only two exceptions, Figure 3 illustrates linear relationships among frequency, severity, and bothersomeness items. The exceptions are *Decreased appetite* and *Decreased sexual desire or interest*, symptoms with scale values that have differed significantly in all comparisons. As expected, items rated for bothersomeness show significantly less impact than corresponding items rated for frequency or severity. Frequency and severity plots, however, show items for these dimensions are equivalent. This is consistent with the plot of the overall instrument and with only items of the bothersomeness and severity dimensions.

Item invariance. Scale integrity in the Rasch model is chiefly determined by item stability. Item scale values that remain statistically invariant across culturally diverse samples offer researchers measurement on an absolute scale. The item plot in Figure 4 shows remarkable stability when these items are estimated separately for the Women's Walking Program and the Migration and Health Project women. Among 102 items, only six are statistically unstable. Women in the Migration and Health Project tended to rate *Fluid retention* (bothersomeness) and *Depression* (frequency) higher than those from the Women's Walking Program. On the other hand, responses from the Women's Walking Program rated *Decreased appetite* (severity), *Abdominal pain* (severity), *Decreased sexual desire or interest* (bothersomeness), and *Decreased appetite* (frequency) higher than

those from the Migration and Health Project. As can be seen, all remaining item differences fall within the 99 percent confidence intervals.

Validity. Table 2 provides descriptive statistics of the Rasch model impact measures and the results of comparing women by sample group, age, menopausal status and depression scores. Both analysis of variance and effect sizes indicate the significant differences. As expected, women from the Migration and Health Project reported significantly higher symptom impact than women from the Women's Walking Program and the effect size (1.33 SD units) was substantial. Older women experience more symptoms than younger women. The corresponding difference by menopausal status is expected, due to the high correlation between menopausal status and age. These predictable differences provide empirical support for validity of the symptom impact construct. When the depression scores were recoded to identify those women who scored at the arbitrary cut point of 16 or more, suggesting clinical depression, their Rasch model impact measures were significantly different. This is further supported by the positive correlation between symptom impact measures and the total depression scores (0.67).

DISCUSSION

Systematic ways to characterize symptom experience are important for developing interventions and increasing knowledge (Giardino & Wolf, 1993; Vessey & Richardson, 1993). Some recent efforts in symptom management address interventions to modify symptoms that may be independent of treating underlying disease or syndrome. These innovative approaches require precise evaluation that would be facilitated by symptom representations on graduated scales rather than the ordinal scales in common use. With advances in symptom measurement, researchers could assess the responsiveness of symptoms to interventions or treatments, measure the impact of symptoms on health and well-being, and document changes in symptom experience (Kirschner & Guyatt, 1985).

Inventories that consider multiple dimensions provide more

information than checklists that only indicate the presence or absence of symptoms (Samarel, 1996; Sarna, 1998; Hightower, 1997; Taylor, 1994). The present analyses examined not only the linear measurement properties for three dimensions of symptom experience (frequency, severity, and bothersomeness) but also examined the plausibility of combining these into a conceptual construct of symptom impact.

While the explicit target of these analyses is symptom experience related to women's hormonal changes at midlife, the context of the inventory is much broader and includes overall physical and psychological symptoms. Although not directly related to menopause, these other symptoms are experienced with regularity by many midlife women and thus need to be included in any comprehensive model of symptoms. For example, we found *Aches and stiffness in muscles and joints*, *Fatigue or lack of energy,* and *Irritability,* symptoms not typically associated with menopause, to be among the most commonly reported symptoms of midlife women. An inventory expanded by these symptoms improves the understanding of women's typical symptom experience, and helps to identify important "marker" clusters of symptoms.

To advance from raw data to equal interval measures, the Symptom Impact Inventory ratings were submitted to Rasch measurement analysis and evaluated according to several criteria. First a qualitative appraisal was conducted to determine if this item ordering shows the expected gradation from low to high impact. In the observed hierarchy, lesser symptom impact represents women reporting fewer, milder symptoms, while greater impact represents women reporting many, severe symptoms. Women in between reflect graduations of impact.

A second appraisal examined conformity of both women and items to an impact hierarchy assessed through fit statistics. Appropriate fit indicates that symptom experiences quantitatively accumulate across the scale empirically verifying the conjoint additivity necessary for linear measurement (Perline, Wright, & Wainer, 1979). This property is completely obscured in ordinary ratings analyses but meaningful arithmetic operations on scale values require it (Stucki,

Daltroy, Katz, Johannesson, & Liang, 1996; see also Michell, 1999 for additivity discussion in psychology).

Results show the Symptom Impact Inventory defines an equal interval scale with generally excellent Rasch model fit. Even when contrasting the two sample groups, scale structure, with few exceptions, was remarkably robust. The only issue of substantial concern is the lack of very high and very low impact items and their implications for longitudinal studies. While the Symptom Impact Inventory was generally effective with this sample of women, certain future conditions may challenge the sensitivity of the scale. For example, change for groups reporting extremely low impact would tend to be underestimated by the Symptom Impact Inventory because too few items define this portion of scale. Fortunately, routine scale development could address this issue.

The three dimensions for measuring impact enrich symptom understanding. For example, results confirm that symptom experience is more than a simple occurrence; the scale definition is significantly extended by the severity and frequency ratings. An unexpected finding, however, is that bothersome items, without exception, have less impact than either severity or frequency. On this foundation of bothersomeness, however, higher severity ratings distinguish women experiencing greater impact. Content structure of this variable also contributes to a better understanding of the symptom experience. Lesser impact is represented by bothersomeness items of symptoms forming a General Health cluster. As women experience the bothersomeness of these symptoms, impact increases. Somewhat higher on the scale the bothersomeness of Vaginal symptoms intermingled with Physical symptoms. If women experience these symptoms, impact increases even further. Finally, symptom severity and frequency further extend the impact scale range. Because this content organization will reappear whenever the Symptom Impact Inventory is administered, future research should consider the theoretical implications of this ordering.

Statistical conformity of symptoms to a hierarchical order establishes that women tend to report not only symptoms at their own scale position but all symptoms below. For example, women report-

ing greater impact items such as severity of *Nausea, Fluid retention, Constipation*, *Shortness of breath*, and *Dizziness* also tended, statistically, to report lesser impact General Health symptoms such as F*atigue, Aches and stiffness in muscles and joints, Irritability*, and *Depression*. The Rasch model fit analysis examines scale continuity to identify specific "women by symptom interactions" that do not conform to this linear measurement ideal. This comparison isolates unusual interactions that may distort overall scale definition. In this analysis, the excellent model fit verifies that midlife women's symptom experience indeed conforms to this cumulative structure and can be measured by a linear scale.

Unlike traditional raw data approaches, the Rasch model estimates reliability for both measured women and scaled items. The goal is to establish whether item structure would reoccur with other samples, as well as the stability of the obtained women measures, and in these analyses high reliability was established for both. Reliability, however, does not diminish the importance of scale invariance. Meaningfulness ultimately depends on items preserving not only their order, usually assessed by reliability, but stable scale values. Unstable scale structures that are psychometrically reliable essentially perpetuate confusion. This study examined invariance by comparing scales estimated separately for two sample groups, women from the Migration and Health Project who are immigrants from the former Soviet Union and generally healthy volunteers for the Women's Walking Program. Results revealed only two unstable symptoms.

One of these symptoms, *Decreased sexual desire or interest*, was experienced by 117 women (34%). While Frequency was rated high for 52 women (44% of women who experienced it) and Severity was high for 29 (25%), only 7 (6%) rated Bothersomeness high. This could reflect the fact that 123 women or 36% of the total sample were not married. For some unmarried women, there may be confusion about how to respond to this question. On occasion, women have asked what was meant by *Decreased sexual desire or interest*; decreased from what? It is possible that the meaning of this symptom is different for different women.

The other unstable symptom, *Decreased appetite*, was experienced by only 56 women (16% of the total sample). Of those women who experienced the symptom, 7 rated Frequency high, only 2 rated Severity high, and only 1 rated Bothersomeness high. The lack of stability of this symptom could reflect the way midlife women feel about a decrease in appetite. Some women in midlife may believe they need to lose weight, consequently, a decrease in appetite is not regarded as a symptom of a health problem but rather as a solution to weight gain. For this reason, it is not surprising that this symptom might not maintain a stable fit in a linear scale of symptom impact.

Validity of the Symptom Impact Inventory was supported by the known groups technique. First, the Rasch model symptom impact measures were lower for the Migration and Health Project group than for the Women's Walking Program group. These groups were from profoundly different cultural backgrounds. Cross-cultural differences previously reported suggest significant population differences are fundamental to valid symptom measurement (Bell, 1995; Berg & Taylor, 1999; Dennerstein, et al., 1993). As expected, impact scores were lowest in the youngest women who would be the least likely to experience multiple symptoms related to aging. Consistent with prior studies (Wilbur et al., 1998), differences were also found between menopausal status groups with the premenopausal women having the lowest symptom impact scores. A positive relationship between the impact scores and depression scores is consistent with the expectation that higher depression should occur with higher impact. This is further reinforced by significant differences between the depressed and nondepressed women.

It is not clear, however, if these are several differences, or only one major difference that confounds these findings. In these analyses, older women, post menopausal women, and women from the Migration and Health Project all scored higher on the scale. However, the women from the Migration and Health Project were older and more likely to be postmenopausal. Further research needs to examine the cross cultural differences in a sample that more similarly reflects the age and menopausal status of the comparison group.

Several questions remain for future research. First, although

developed for healthy midlife women, it is important to determine that the Symptom Impact Inventory is a general purpose symptom measure that could be useful to women of other age groups or women who may not be healthy. Second, while symptom positions on this scale are fixed by statistical estimation, women positions reflect immediate symptom experience and are expected to change over time. Analyses that examine change in the symptom experience over time would demonstrate the utility of the Symptom Impact Inventory in assessing change. Finally, future work needs to examine the ability of the Symptom Impact Inventory to identify debilitation or functional limitations.

In summary, both traditional psychometric and Rasch analyses support the Symptom Impact Inventory as a reliable and valid tool for measuring the symptom experience in midlife women. The Rasch model identified an equal interval distribution which will be useful in the assessment of change. In spite of the strikinge differences between the two sample groups, the item structure of the Symptom Impact Inventory remains stable. In the rapidly advancing field of midlife women's health, this scale should make important contri-

---

This study was funded by grants from NIH, NINR, #NR04134; NIH, NICHD, #R15 HD534225; and the University of Illinois at Chicago, College of Nursing, Dean's Fund.

Arlene Michaels Miller, PhD, RN, Associate Professor; JoEllen Wilbur, PhD, RN. FAAN, Associate Professor; Andrew Montgomery, PhD, Research Specialist; Peggy Chandler, PhD, Research Assistant Professor; and Nikolaus Bezruczko, PhD, Visiting Professor. Department of Public Health, Mental Health, and Administrative Nursing, University of Illinois at Chicago. We are deeply grateful to Ben Wright, PhD, for his comments and suggestions.

REFERENCES

Alder, E. (1998). Blatt-Kupperman menopausal index: a critique. *Maturitas, 29,* 19-24.

Aroian, K.J., Spitzer, A., & Bell, M. (1996). Family stress and support among former Soviet immigrants. *Western Journal of Nursing Research, 18*, 655-674.

Aroian, K.J. & norris, A.E. (1999). Somatization and depression among fomer Soviet immigrants. *Journal of Cultural Diversity*, 6 (3), 93-101.

Avis, N.E., Brambilla, D., McKinlay, S., & Vass, K. (1994). A longitudinal analysis of the association between menopause and depression: Results from the Massachusetts Women's Health Study. *Annals of Epidemiology, 4,* 214-220.

Bell, M.L. (1995). Attitudes toward menopause among Mexican-American women. *Health Care for Women International, 16,* 425-435.

Berg, J., & Taylor, D. (1999). Symptom experience of Filipino American midlife women. *Menopause, 6,* 105-114.

Brislin, R.W. (1990). Applied Cross-Cultural Psychology: An introduction. In R. W. Brislin (Ed.), *Applied Cross-Cultural Psychology.* Newbury, PA: Sage Publications.

Bungay, G.T., Vessey, M.P., & McPherson, C.K. (1980). Study of symptoms in middle life with special reference to menopause. British Medical Journal, 181-183.

Busch, C. M., Zonderman, A.B., & Costa, P.T. (1994). Menopausal transition and psychological distress in a nationally representative sample: Is menopause associated with psychological distress? *Journal of Aging and Health, 6*, 209-228.

Cwikel, J., Abdelgani, A., Goldsmith, J.R., Quastel, M., & Yevelson, I.I. (1997). Two-year follow-up study of stress-related disorders among immigrants to Israel from the Chernobyl area. *Environmental Health Perspectives, 5*, (supplement 6), 1545-1550.

Dennerstein, L., Smith, A.M., Morse, C., Burger, H., Green, A., Hopper, J., & Ryan, M. (1993). Menopausal symptoms in

Australian women. *Medical Journal of Australia, 159,* 232-236.

Engstrom, J. (1988). Assessment of the reliability of physical measures. *Research in Nursing & Health, 11,* 383-389.

Giardino, E.R., & Wolf, Z. (1993). Symptoms: Evidence and experience. *Holistic Nurse Practitioner, 7,* 1-12.

Grimby, A., Milsom, I., Molander, U., Wiklund, I., Ekelund, P. (1993). The influence of urinary incontinence on quality of life of elderly women. *Age Aging, 22,* 82-89.

Guthrie, J.R., Smith, A.A.A., Dennerstein, L., & Morse, C. (1995). Physical activity and the menopause experience: A cross-sectional study. Maturitas, 20, 71-80.

Hightower, M. (1997). Effects of exercise participation on menstrual pain and symptoms. *Women and Health, 26*, 15-27.

Holte, A., & Mikkelson, A. (1991). Psychosocial determinants of climacteric complaints. *Maturitas, 13,* 205-215.

Hoyer, W.J., & Rybasch, J.M. (1994). Characterizing adult cognitive development. *Journal of Adult Development, 1,* 7-12.

Hunter, M., Battersby, R., & Whitehead, M. (1986). Relationships between psychological symptoms, somatic complaints, and menopausal status. *Maturitas, 8*, 217-228.

Hunskaar, S., & Vinsnes, A. (1991). Quality of life in women with urinary incontinence as measured by sickness impact profile. *Journal of American Geriatric Society, 39,* 378-382.

Jaszmann, L., van Lith, N.D., & Zaat, J.C.A. (1969). The perimenopausal symptoms: the statistical analysis of a survey. *Medical Gynecology and Sociology,* 4, 268-27.

Kaufert, P.A. (1994). A health and social profile of menopausal women. *Experimental Gerontology, 29,* 343-350.

Kaufert, P.A., Gilbert, P., & Hassard, T. (1988). Researching the symptoms of menopause: An exercise in methodology. *Maturitas, 10,* 117-131.

Kaufert, P.A., & Syratuik, J. (1981). Symptom reporting at menopause. *Social Science and Medicine, 151*, 173-184.

Kerlinger, F.N. (1973). Foundations of Behavioral Research. New York: Holt Reinhart & Winston.

Kirschner, B., & Guyatt, G. (1985). A methodological framework for assessing health indices. *Journal of Chronic Diseases, 38,* 27-36.

Kohlmeier, L., Mendex, M., Shalnova, S, Martinchik, A., Chakraborty, H., & Kohlmeier, M. (1998). Deficient dietary iron intakes among women and children in Russia: Evidence from the Russian Longitudinal Monitoring Survey. *American Journal of Public Health, 88,* 576-580.

Kotulak, R. (2000). Study finds lonely hearts often have sick hearts. *Chicago Tribune, August 8.*

Larson, P.J., Carrieri-Kohlman, V., Dodd, M., Douglas, M., Faucett, J., Froelicher, E., Gortner, S., Lee, K., Miaskowski, C., Savedra, M., Stotts, N., Taylor, D., & Underwood, P. (1994). A model for symptom management. *Image, 26,* 272-276.

Li, S., Holm, K., Gulanick, M., Lanuza, D., & Penckofer, S. (1999). The relationship between physical activity and perimenopause. *Health Care of Women International, 20,* 163-178.

Linacre, J. M., & Wright, B. D. (2000). *Winsteps: Rasch Analysis For All Two-Facet Models, Version 3.04.* Chicago: MESA Press.

Linacre, J. M. (1999). *Learning from principal components analysis of residuals.* Presented at Chicago Objective Measurement Table, April 22. See also MESA Research Note No. 7.

Linacre, J. M. (1996). Structure in Rasch residuals: Why principal components analysis? *Rasch Measurement Transactions, 12,* 636.

Matthews, K.A., Wing, R.R., Kuller, L.H., Meilahn, E.N., Kelsey, J.L., Costello, E.J., & Caggiula, A.W. (1990). Influences of natural menopause on psychological characteristics and symptoms of middle-aged healthy women. *Journal of Consulting and Clinical Psychology, 58,* 345-353.

McCorkle, R. (1987). The measurement of symptom distress. *Seminars in Oncology Nursing, 3,* 248-255.

McDowell, I., & Newell, C. (1996). *Measuring health: A guide to rating scales and questionnaires.* New York: Oxford Uni-

versity Press.

McKinlay, S.M., & Jeffreys, M. (1974). The menopausal syndrome. British Journal of Preventive & Social Medicine, 28, 108-115.

McKinlay, S.M., McKinlay, J.B., & Avis, N.e. (1989). The Massachusetts women's health study: A longitudinal study of the health of mid-aged women and the epidemiology of the menopause. *Psychology of Women, 16*, 2-6.

Merbitz, M., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation, 70,* 308-312.

Michell, J. (1999). *Measurement in Psychology*. Cambridge: Cambridge University Press.

Miller, A.M., Wilbur, J., Montgomery, A., & Chandler, P. (1998). Social role quality and psychological well-being. *AAOHN Journal, 46*, 371-378.

Miller, A.M., Wilbur, J., & Chandler, P. (1999). Menopausal attitudes, symptoms and acculturation in midlife women from the former Soviet Union. Looking forward, looking back. The place of women's everyday live in health research. *Proceedings of the 12th Conference of the Society for Menstrual Cycle Research*, June 5-7, 1997, Chicago, IL, 264-268.

Mitchell, E., & Woods, N, (1996). Symptom experience of midlife women: Observations from the Seattle midlife women's health study. *Maturitas, 25*, 1-10.

Neugarten, B.L., & Kraines, R.J. (1965). Menopausal symptoms in women of various ages. *Psychosomatic Medicine, 27*, 266-273.

Perline, R., Wright, B. D., & Wainer, H. (1979). Rasch Model as Additive Conjoint Measurement. *Applied Psychological Measurement, 3,* 237-255.

Perz, J. (1997). Development of the menopause symptom lists: A factor analytic study of menopause associated symptoms. *Women and Health, 25,* 53-69.

Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psy-*

chological Measurement, 1, 385-401.

Samarel, N., Leddy, S., Greco, K., Cooley, M.E., Torres, S., Tulman., & Fawcett. J. (1996). Development and testing of the Symptom Experience Scale. *Journal of Pain and Symptom Management, 12,* 221-228.

Sarna, L., (1998). Effectiveness of structured nursing assessment of symptom distress in advanced lung cancer. *Oncology Nursing Forum, 25,* 1041-1048.

Shaver, J., Giblin, E., Lentz, M., & Lee, K. (1988). Sleep patterns and stability in perimenopausal women. *Sleep, 11*, 556-561.

Smith, L.S., (1995). Russian Americans. In J. N. Giger & R. E. Davidhizar (Eds.). *Transcultural Nursing*. St. Louis, MO: Mosby.

Sternfeld, B., Quesenberry, C.P., & Husson, G. (1999). Habitual physical activity and menopausal symptoms: A case-control study. *Journal of Women's Health, 8,* 115-123.

Stucki, G., Daltroy, L., Katz, J. N., Johannesson, M., & Liang, M. H. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *Journal of Clinical Epidemiology, 49*, 711-717.

Taylor, D. (1994). Evaluating therapeutic change in symptom severity at the level of the individual woman experiencing severe PMS. *Image, 26,* 25-33.

Vessey, J.A., & Richardson, B.L. (1993). A holistic approach to symptom assessment and intervention. *Holistic Nursing Practice, 7,* 13-21.

Wilbur, J., Holm, K., & Dan, A.J. (1992). The relationship of energy expenditure to physical and psychological symptoms in women at midlife. *Nursing Outlook, 35*, 269-276.

Wilbur, J., Miller, A.M., & Montgomery, A. (1995). The influence of demographic characteristics, menopausal status, and symptoms on women's attitudes toward menopause. *Women and Health, 23,* 19-39.

Wilbur, J., Miller, A.M., Montgomery, A., & Chandler, P. (1998). Women's physical activity patterns: Nursing implications.

JOGNN, 27, 383-392.

Woods, N.F., Most, A., & Dery, G.K. (1982). Estimating perimenstrual distress: A comparison of two methods. *Research in Nursing and Health, 5,* 81-91.

Woods, N.F., Lentz, M., Mitchell, E.S., Heitkemper, M., & Shaver, J. (1997). PMS over 40: Persistence of a stress-related symptom pattern. *Research in Nursing and Health, 20,* 329-340.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70,* 857-860.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

Wright, B. D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

Wright, B. D., & Stone, M. (1979). *Best Test Design.* Chicago: MESA Press.

Journal of Outcome Measurement

Volume 4

Author and Title Index