Volume 4, Number 3, 2000/2001 ISSN 1090-655X

Journal of Outcome Measurement®

Dedicated to Health, Education, and Social Science

REHABILITATION FOUNDATION INC.
EST. 1993
Research & Education

JOURNAL OF OUTCOME MEASUREMENT®

Volume 4, Number 3 2000/2001

JOURNAL OF OUTCOME MEASUREMENT®, 4(3), 635- 666

COMMENSURATE RATINGS OF HEALTH CARE

Gordon G. Bechtel
University of Florida
and
Florida Research Institute

A new descriptive item statistic, termed the mean cumulative logit (MCL), is advocated for scoring ratings of health care at the population level. The advantages of the MCL are demonstrated on data from the Consumer Assessment of Health Plans Study (CAHPS). The CAHPS data require (1) the comparison of binary and ordinal ratings in a common metric, (2) a treatment of unit and item nonresponse, and (3) the control of ordinal item correlations. These requirements are handled by a cumulative logit model that is applicable to unweighted (incomplete) and weighted (complete) data. The former case gives item satisfactions from a patient perspective. The latter case generates these satisfactions as social utilities from a provider perspective. From both of these viewpoints the perceived quality of health care is greater for fee-for-service plans than managed care plans in the field test population studied here.

KEYWORDS: Mean Cumulative Logit (MCL); Binary and Ordinal Items; Nonresponse; Patient and Provider Perspectives; Social Utility.

Requests for reprints should be sent to Gordon G. Bechtel, University of Florida, 212 Bryan Hall, P.O. Box 117155, Gainesville, FL 32611-7155

Introduction

Background and Aims

The measurement of national satisfaction, or life quality, was stimulated in the 1970's by Levy and Guttman (1975), Andrews and Withey (1976), and Clogg's (1979) latent class analysis of data from the 1975 General Social Survey. It was also spurred in that decade by the U.S. National Science Foundation's program of Research Applied to National Needs (RANN). RANN gave impetus to the measurement of consumer satisfaction (Bechtel, 1977), which was combined with more general life-quality research in several marketing conferences (Bechtel, 1978; 1983). When RANN faltered it was replaced by the quality-control revolution stimulated from earlier work of W. Edwards Deming (Mann, 1994). This revolution has led to worldwide preoccupation with consumer satisfaction, and in the public sector it surfaced as the "National Performance Review" spearheaded by the Vice President (Gore, 1993). The correlated effort in the private sector is known as "Outcome Evaluation," and consumer satisfaction in particular service categories such as health care straddles both sectors.

The aim of this paper is to develop markers of consumer satisfaction with health care delivery that are comparable over different questionnaire items, different population aggregates, and different delivery modes in the United States. Such markers were called for by William Scanlon of the General Accounting Office in his closing testimony on January 18, 1996 to the House of Representative's Government Reform and Oversight Subcommittee. These markers will assist health care providers in evaluating how they "measure up", and they will assist consumers in their "comparative shopping" among health plans.

Consumer Assessment of Health Plans Study (CAHPS)

In 1995 the Agency for Healthcare Research and Quality initiated the Consumer Assessment of Health Plans Study (CAHPS) to help consumers select health care plans and services. This agency in col-

laboration with the Research Triangle Institute, RAND, and the Harvard Medical School, designed survey items tapping consumer evaluations of health care.

In positioning this effort from both patient and provider perspectives the Agency for Healthcare Research and Quality (1996, p. 1) states:

> Overall, most surveys of health plan members to date have tended to focus mainly on meeting the needs of institutional purchasers and plans, a single type of care delivery system (managed care), and the privately insured. To expand on this work, the Agency .... has funded the Consumer Assessment of Health Plans Study (CAHPS). The overall goal of CAHPS is to provide an integrated set of carefully tested and standardized survey questionnaires and accompanying final report formats that can be used to collect and report meaningful and reliable information from health plan enrollees about their experiences.

Summarizing the results of this questionnaire development the Agency for Healthcare Research and Quality (1996, p. 14) continues:

> The CAHPS team used cognitive testing to explore the strengths and weaknesses of objective reports versus ratings as measures of consumer experience with health care and health plans. "Reports" require consumers to use a Yes/No response format to report health care or health plan experiences ("Did something happen or not?" is the key question). In contrast, "ratings" measure consumers' *reactions* to their experiences, using such scales

as "poor to excellent" or "very dissatisfied to very satisfied." In determining the best measure of consumer experience, we needed to learn which health care events or interactions are most accurately reported with a Yes/No response format and which health care events or interactions are most accurately reported with ratings. The results of the CAHPS cognitive testing showed that

- *Reports of problems* with health care providers or the health plan are most effectively asked in a yes/no response format,

- *Interactions* with health care providers or the health plan are most effectively measured using a graded rating scale.

The Present Analysis

The preceding distinction in accessing consumer experiences requires an item model that scales yes/no and graded responses in a common metric. By placing binary and ordinal items on the *same* scale, the present analysis gives health-care providers and consumers a meaningful report that profiles modes of health delivery item by item. This scale output also provides separate yardsticks on which to compare different health delivery modes item by item. This comparative profiling gives providers a competitive benchmark for quality improvement, and it gives consumers a context in which to pursue idiosyncratic preferences in their choice among delivery modes.

The CAHPS Dataset

Population and Sample

The first field test data for the CAHPS core instrument was carried out in October and November of 1996 on samples of adults covered by private health plans. The data was collected from computer-assisted telephone interviews (CATI) reported by Williams et al. (1997). A total of 539 eligible protocols were recovered from respondents distributed across three sites: Health Insurance Plan of California, Michigan State University, and Connecticut Business and Industry Association. This sample was evenly split between health plan enrollees covered by gatekeeper (HMO) and non-gatekeeper (fee-for-service) plans.

The analysis below uses a subsample of 379 respondents who reported making one or more visits *for themselves* "to a doctor's office or clinic, or a hospital emergency room" *in the last 6 months*. The average length of the completed telephone interview for this subsample was 16.5 minutes. In the analysis below this subsample is divided into 192 fee-for-service enrollees and 187 managed care enrollees.

Core Items and Responses

Table 1 exhibits the 19 CAHPS field-test items with *yes/no* and *never,sometimes / usually /always* responses. Each item in this table has been paraphrased from the actual CATI questionnaire. Thus, *no* represents a favorable consumer response. Also, the comma between *never* and *sometimes* indicates that these two response options have been collapsed into a single category. This grouping is advocated by Williams et al. (1997, p. 61) due to the sparse frequencies in these two lower response options. Hence, this quaternary format in Table 1 becomes a ternary scale with the middle category *usually*.

Table 1
CAHPS Core Items and Responses

Response Categories	Item
Yes / No (reports of problems)	Less involved than you wanted
	Preventative care not encouraged
	Did not know history
	Could not get same day appointment
	Could not get appointment as soon as wanted
	Waited more than 15 minutes
	Did not get off-hours help
	Did not get daytime help
	Not able to see specialist
	Specialist care did not meet needs
	Doctor did not provide tests or treatment
	Plan did not pay for tests or treatment
	More paper work than reasonable
	Got no information from customer service
Never, Sometimes / Usually / Always (interactions)	Listened carefully
	Explained things
	Respected what you said
	Spent enough time
	Treated with courtesy and respect by office staff

Commensuration of Binary and Ordinal Ratings

A Common Satisfaction Metric

Let i represent any of the 14 binary items in Table 1. Then for item i the (below the line) cutpoint in Figure 1 partitions a satisfaction scale into two successive intervals. The observed *yes* and *no* response categories are represented by these two unobserved line segments. Thus, an ordered binomial is modeled as two successive intervals on a satisfaction continuum. The intervals on this scale are separated by cutpoint τ_i .

Next, let j represent any of the *never,sometimes/usually/always* items in Table 1. Now, selectively for item j, the two (above the line) cutpoints in Figure 1 partition the *same* scale into three successive intervals that are separated by the j-specific cutpoints τ_{j1} and τ_{j2}. This cutpoint specificity is necessary in fitting the CAHPS satisfaction data.

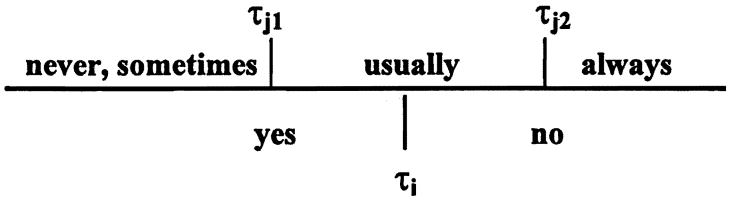

Figure 1. *The consumer satisfaction scale with binary and ternary cutpoints.*

Cumulative Probabilities

For each binary item i in Table 1 there are two population probabilities $\pi_{i0} + \pi_{i1} = 1$ for responding in its ordered categories. Then

$$\gamma_i = \pi_{i1}$$

is the cumulative probability of responding to item i *above* cutpoint τ_i on the satisfaction scale in Figure 1.

For ternary item j in Table 1 there are three population probabilities $\pi_{j0} + \pi_{j1} + \pi_{j2} = 1$ for responding in its ordered categories. Let

$$\gamma_{j1} = \pi_{j1} + \pi_{j2} ,$$

$$\gamma_{j2} = \pi_{j2} .$$

Then γ_{jc} for c = 1,2 is the cumulative probability of responding to item j *above* cutpoint τ_{jc} on the satisfaction scale in Figure 1. The cutpoints τ_{j1} and τ_{j2} are the lower and upper bounds of the *usually* zone on this scale. The γ_{jc}, along with the γ_i, will now be modeled with respect to the cutpoints.

Logistic Links for Cumulative Probabilities

The following logistic functions link γ_i and γ_{jc} to their cutpoints τ_i and τ_{jc} and their satisfaction parameters (not shown) on the scale in Figure 1:

$$\gamma_i = \frac{\exp\{\eta_i - \tau_i\}}{1 + \exp\{\eta_i - \tau_i\}} \quad \text{for } i = 1,\ldots,14, \tag{1a}$$

$$\gamma_{jc} = \frac{\exp\{\eta_j - \tau_{jc}\}}{1 + \exp\{\eta_j - \tau_{jc}\}} \quad \text{for } j = 1,\ldots,5, \tag{1b}$$

where

η_i = the location of item i on the scale in Figure 1,

η_j = the location of item j on the scale in Figure 1.

The cutpoints τ_i and τ_{jc} for c = 1, 2 saturate the logistic model (1).

Figure 2 exhibits the commensuration of η_i and η_j. The dotted logistic curve increases (asymptotically) from zero to one and cuts the dotted ordinate into the *yes* probability (above the dotted curve) and the *no* probability (below the dotted curve). The equation of this characteristic curve is given by (1a), which transforms the cumulative probability of *no* for binary item i to the satisfaction scale. The solid logistic curves for ternary item j cut the solid ordinate into the *never, sometimes* probability (above the upper solid curve), the *usually* probability (between the solid curves), and the *always* probability (below the lower solid curve). The operating characteristics graphed by the two solid curves transform the cumulative probabilities γ_{j1} and γ_{j2} for ternary item j into the *same* metric. The equations for these characteristic curves are given by (1b) for c = 1, 2.

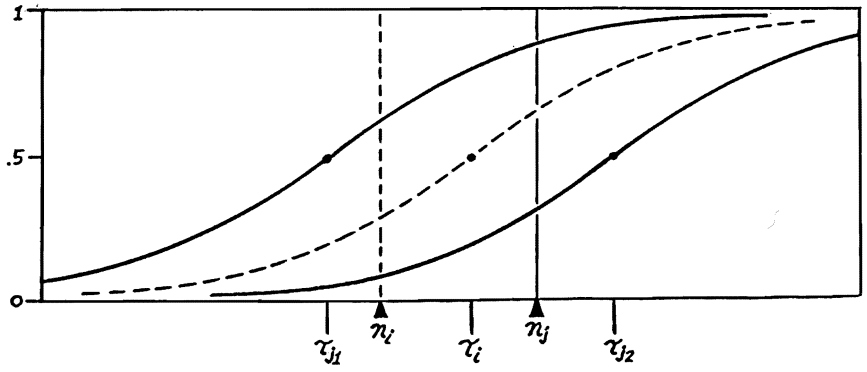

Figure 2. *Logistic operating characteristics for a binary item i and a ternary item j.*

Identification of Model (1)

The inverse functions of (1a) and (1b) are

$$\lambda_i = \eta_i - \tau_i \qquad \text{for } i = 1,\ldots,14, \tag{2a}$$

$$\lambda_{jc} = \eta_j - \tau_{jc} \quad \text{for } j = 1,\ldots,5, \tag{2b}$$

where

$$\lambda_i = \text{logit}\{\gamma_i\} = \log\{\gamma_i/(1-\gamma_i)\},$$

$$\lambda_{jc} = \text{logit}\{\gamma_{jc}\} = \log\{\gamma_{jc}/(1-\gamma_{jc})\},$$

Equations (2a) and (2b) allow us to express our η and τ parameters as linear forms in the logits. The particular linear forms used here are given by the following uniqueness conditions:

$$\tau_i = 0 \qquad \text{for } i = 1,\ldots 14, \tag{3a}$$

$$\tau_{j1} + \tau_{j2} = 0 \qquad \text{for } j = 1,\ldots,5. \tag{3b}$$

Equations (2a) and (3a) identify the cutpoints and satisfaction locations for our 14 binary items as

$$\eta_i = \lambda_i\,, \tag{4a}$$

$$\tau_i = 0\,. \tag{4b}$$

Also, summing each side of (2b) over c and using (3b) identifies the cutpoints and satisfaction parameters for our five ternary items as

$$\eta_j = \lambda_{j.} \,, \qquad (5a)$$

$$\tau_{jc} = \lambda_{j.} - \lambda_{jc} \,, \qquad (5b)$$

where

λ_{jc} = the cumulative logit for item j and cutpoint c = 1,2,

$\lambda_{j.}$ = the mean cumulative logit for item j over c = 1,2.

Equation (5a) shows that the saturated cutpoints produce item locations that are *mean cumulative logits.* Equation (5b) shows that these saturated cutpoints are *adjusted cumulative logits.*

Equations (3a) and (3b) define an *interval* scale with an arbitrary additive constant that cancels by differencing in (2a) and (2b). The origin of this interval satisfaction scale is fixed by (3a) and (3b) at the *common* mean of each item's cutpoints in Figure 2. Under the (reasonable) assumption that cutpoint means are stationary from item to item, a common scale origin, i.e., a commensuration, is maintained for the satisfaction locations η_i and η_j for i = 1,...,14 and j = 1,...,5. This stationarity assumption is much weaker than the standard assumption associated with scored item responses; namely, that ordered category values are equally spaced within and between items as well as stationary.

Cutpoint Specificity for Ordinal Ratings

Items. The item dependence of the thresholds τ_{jc} in (1b) and (2b) is based on the work of Bechtel (1991) and Bechtel et al. (1993). Unpublished research by the author has also demonstrated that survey items consistently reject the hypothesis that $\tau_{jc} = \tau_c$. This result is not surprising in view of the overly stringent assumption that there is a common set of cutpoints that sweep across several ordered multinomials generated by a set of survey items (cf. Torgerson, 1958, ch.

10; McCullagh and Nelder, 1989, ch. 5).

Samples. Common cutpoints across two compared samples characterize the well known proportional odds model for ordinal data. However, the lack of fit of this model has been observed by Landis et al. (1988). Unpublished research by the author has also found the proportional odds model to be frequently rejected when comparing two survey samples. This model failure is an important motive for using sample dependent thresholds in the present study.

The sample dependence of the thresholds in (1b) and (2b) is achieved by separate runs on the fee-for-service and managed care samples. It is easily shown that the generalized least squares procedure, when carried out on samples separately, produces MCLs and standard errors which are identical to those obtained from a single run on two samples simultaneously. This invariance follows from the block diagonality of the covariance and design matrices for the dual samples when the cutpoints are sample specific. An important practical advantage of this invariance is that large survey datasets may be handled by running each sample separately.

Unweighted Analysis: The Patient Perspective

The CAHPS core questionnaire generates item nonresponse from three sources: *skipped* items that are not applicable to a particular respondent and volunteered *don't know* and *refuse* responses to items that are not skipped (Williams et al., 1997). The varying frequencies of these nonsubstantive responses over the items in Table 1 require separate item-by-item analyses based on varying subsample sizes. Nevertheless a binary item's η_i in (4a) is commensurate with a ternary item's η_j in (5a) on the scale in Figures 1 and 2. In conventional cross tabulations the two percentages for item i are not commensurate with the three percentages for item j, nor is the mean item i rating over the scores 0,1 comparable to the mean item j rating over the scores 0,1,2.

Table 2

Satisfaction MCLs for Different Item Subsamples

Item	Fee for Service	Managed Care
Less involved than you wanted	2.18 (.305)	1.95 (.309)
Preventative care not encouraged	.72 (.154) *	.04 (.147)
Did not know history	1.41 (.183)	1.14 (.173)
Not get same day appointment	.77 (.221)	.89 (.217)
Not get appointment as soon as wanted	1.13 (.189)	.92 (.200)
Waited more than 15 minutes	.19 (.165)	.13 (.181)
Not get off-hours help	.95 (.340)	1.14 (.287)
Not get daytime help	1.78 (.262)	1.98 (.267)
Not able to see specialist	2.07 (.320)	1.77 (.289)
Specialist care did not meet needs	2.09 (.353)	2.20 (.373)
Doctors did not provide tests or treatment	2.84 (.389)	2.12 (.306)
Plan did not pay for tests or treatment	1.27 (.218)	†1.86 (.278)
More paperwork than reasonable	.64 (.201)	*2.27 (.606)
Got no information from customer service	2.43 (.426)	*1.37 (.323)
Listened carefully	1.37 (.168)	1.05 (.154)
Explained things	1.75 (.193)	1.41 (.170)
Respected what you said	1.60 (.182)	1.24 (.162)
Spent enough time	1.05 (.154)	.84 (.147)
Respected by office staff	1.74 (.192)	1.54 (.182)

Note. Standard errors are in parentheses. A star (*) and a dagger (†) indicate significance levels beyond .05 and .10 (two-tailed).

Table 2 presents MCLs calculated from (4a) and (5a) for the 19 items in Table 1 by type of care. These MCLs measure satisfaction from the patient perspective, i.e, from *only* those respondents actually experiencing the health care problem or interaction posed by a given CAHPS item. Among the 192 fee-for-service enrollees

the subsample sizes for these 19 items ranged from 43 to 192. For the 187 managed care enrollees these 19 subsample sizes ranged from 32 to 186. Item-by-item (standard normal) z tests reveal three *fee-for-service* versus *managed care* differences to exceed the .05 level of significance, with two of these items favoring the former group and one favoring the latter. One other item reached a marginal significance level, with *managed care* satisfaction exceeding that for *fee-for-service.*

Using the standard errors in Table 2 a z test may also be used to contrast different item satisfactions within a group of enrollees. However, a within-group z test will understate the significance level due to its spuriously large difference variance that neglects an inter-item covariance. This (unknown) covariance stems from common respondents shared by the two different item subsamples from, say, *fee-for-service* enrollees. Nevertheless these within-group contrasts, along with the between-group contrasts, attest to the commensurability of the values in Table 2 on the scale in Figures 1 and 2.

Weighted Analysis: The Provider Perspective

Shrinkage for Item Nonresponse

Neutral coding. Nonsubstantive responses do not arise for the five ternary items in Table 1. However, they are widespread among the 14 binary items which, due to their nature, solicit sizable *not applicable* response frequencies. Because the size of the *not applicable* count signals the incidence of positive and negative responses to a binary item, the suggestion here is to weight this item accordingly in the calculation of its MCL. The weighting is carried out by coding *not applicable*, *don't know,* and *refuse* responses as a neutral response. Thus, *Yes/No* in Table 1 is extended to *Yes/Neutral/No*, Figure 1 is replaced by Figure 3, and the ternary equations (1b),(2b), (3b), and (5) now apply to all 19 items in Table 1.

The neutral coding of nonsubstantive responses is not new. Perhaps the most well known precedent for ternary scales is the University of Michigan's *Index of Consumer Expectations* (Katona, 1975; 1979), which is a United States leading economic indicator. Here *pessimistic, neutral* and *optimistic* response options are presented by telephone. Subsequently, response codes for *don't know, not ascertained* and all other nonsubstantive replies are set equal to the code of the neutral category (Curtin, 1972).

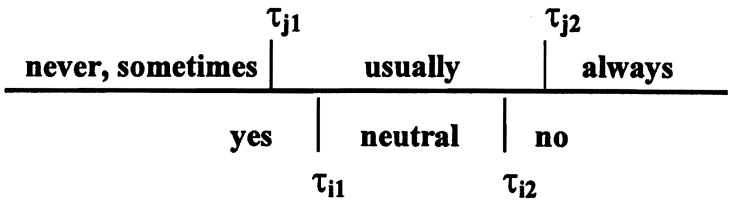

Figure 3. *The consumer satisfaction scale with item-specific ternary cutpoints.*

Bounds on MCLs. Replacing j with i, equation (5a) gives a *Yes/Neutral/No* MCL. This ternary MCL is now shown to be protected by limits in the event of excessive flows into its neutral category. These flows occur when high-frequency nonsubstantive responses to item i are coded as *neutral.*

Referring to Figure 3, let π_{i0}, π_{i1}, and π_{i2} be the probabilities of responses *yes, neutral,* and *no* to item i. Now, we hold constant the ratio of *no* to *yes* responses. That is, we let $\pi_{i2} = \kappa\pi_{i0}$, where κ is a fixed positive constant, and study the behavior of item i's MCL λ_i over a varying probability π_{i1} in the neutral category. This MCL is

$$\lambda_{i\bullet} = \tfrac{1}{2}(\lambda_{i1} + \lambda_{i2})$$

$$= 1/2 \left(\log \frac{\pi_{i1} + \kappa\pi_{i0}}{\pi_{i0}} + \log \frac{\kappa\pi_{i0}}{\pi_{i0} + \pi_{i1}}\right). \qquad (6)$$

Because $\pi_{i0} + \pi_{i1} + \kappa\pi_{i0} = 1$, it follows that $\pi_{i0} = (1-\pi_{i1})/(\kappa + 1)$. Substituting this latter expression into (6) gives

$$\lambda_{i\bullet} = \tfrac{1}{2} \log \frac{\kappa\pi_{i1}(\kappa+1) + \kappa^2(1-\pi_{i1})}{1+\kappa\pi_{i1}}, \qquad (7)$$

which, for fixed κ, is a function of the middle probability π_{i1} only. The range of the MCL function in (7) is traversed by letting the argument π_{i1} travel over its domain $0 < \pi_{i1} < 1$. Hence, the bounds on this range, given by substituting 0 and 1 for π_{i1} in (7), are

$$\tfrac{1}{2} \log \kappa \quad \text{for } \pi_{i1} = 1, \qquad (8a)$$

$$\log \kappa \quad \text{for } \pi_{i1} = 0. \qquad (8b)$$

The expression (8a) gives the limit for $\lambda_{i\bullet}$ as the probability of *neutral* approaches one, and (8b) gives the limit for $\lambda_{i\bullet}$ as this middle probability approaches zero.

Table 3 illustrates the behavior of a ternary MCL for a (hy-

pothetical) sample of 10,000 when $\kappa = 3$, i.e., when three times as many *no* as *yes* responses are given to item i vis-à-vis varying numbers of *neutral* responses. The top value $\lambda_{i.} = 1.0981$ is very close to 1.0986, which is given by (8b) for $\kappa = 3$. The bottom value $\lambda_{i.} = .5494$ is very close to the bound of .5493, which is given by (8a) for $\kappa = 3$. If $\kappa = 1/3$, i.e., if three times as many *yes* as *no* responses are given in each frequency distribution, then the bounds in (8a) and (8b) and the MCL values in Table 3 reverse sign. Hence the approach of this MCL toward zero is symmetric on its positive and negative sides as increasing probability π_{i1} flows into the *neutral* category.

When this increased flow stems from nonsubstantive responses, the corresponding shrinkage of the MCL toward zero serves as a correction, or down weighting, of its value due to the smaller substantive subsample for item i. That is, this adjustment for nonsubstantive responses downgrades the importance of item i due to a lower incidence of positive and negative responses in the total sample. However, equations (8a) and (8b), as illustrated in Table3, show that the MCL is protected from overcorrection. Thus the 3-to-1 favorable odds for item i will preserve an MCL of no less than .5493 when only a subsample of 4 in 10,000 give positive and negative responses.

Table 3 shows that the bounds (8a) and (8b) protect a satisfaction MCL from overcorrection even when there is a massive transfer of nonsubstantive responses into its neutral category. This shrinkage for lower substantive incidence, i.e., for item nonresponse, is not possible with standard integer scoring of ordered response categories. For example, assigning 0, 1, 2 to the categories in Table 3, one finds that the mean rating implodes to 1 with excessive flows into the neutral category.

The binary CAHPS items. The bounds (8a) and (8b) show that for $\kappa > 1$ the ternary MCL shrinks over the interval

$$(\tfrac{1}{2} \log \kappa, \log \kappa)$$

Table 3
The Ternary MCL When $\kappa = 3$

Yes	Neutral	No	MCL
2499	4	7497	1.0981
2000	2000	6000	.8959
1500	4000	4500	.7670
1000	6000	3000	.6750
500	8000	1500	.6049
1	9996	3	.5494

as the probability of a neutral response increases from zero to one. This interval widens with the value of κ, which is the ratio of the *no* to *yes* response probabilities. Hence, items with more extreme *Yes/No* splits (larger κs) will have larger MCL corrections, i.e., larger binary logits for the *Yes/No* format will have larger nonresponse shrinkage when their MCLs are calculated in the *Yes/Neutral/No* format.

The effect of coding nonsubstantive responses as *neutral* is graphed in Figures 4 and 5 for the 14 binary items in Table 1. For both the *fee-for-service* and *managed care* enrollees the relation between binary and ternary satisfactions is linear with slope greater than one due to the above effect. In each figure an item's correction is the vertical distance from its plotted point to the equiangular line. This shrinkage of a subsample (binary) MCL toward the lower bound of a whole-sample (ternary) MCL is analogous to Longford's (1999) shrinkage of a local area sample mean toward the national sample mean.

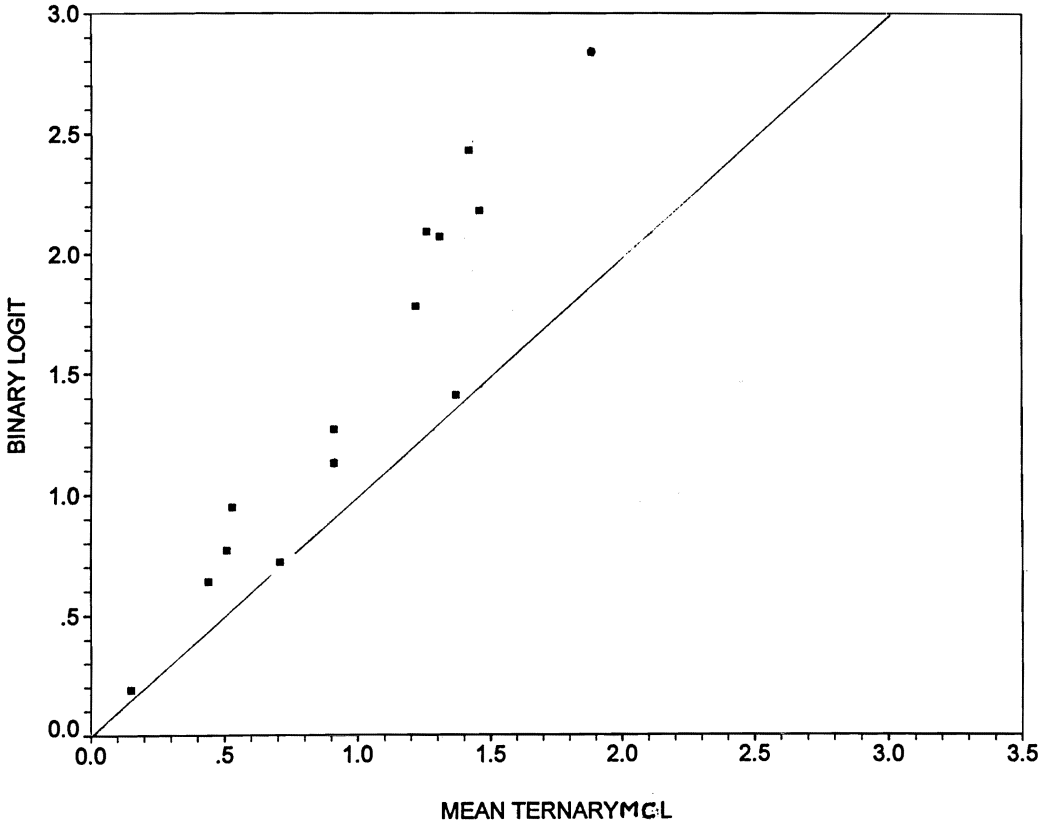

Figure 4. *Plot of binary versus ternary satisfaction for fee-for-service enrollees.*

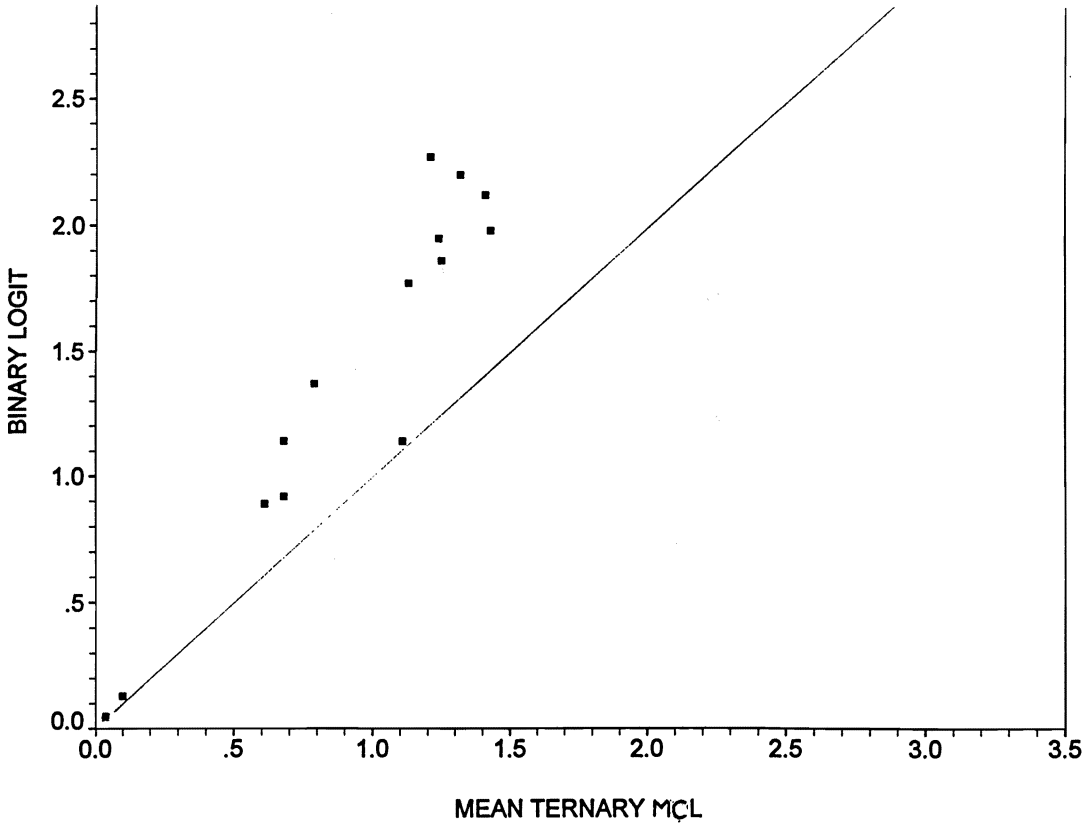

Figure 5. *Plot of binary versus ternary satisfactions for managed care enrollees.*

Case Weighting for Unit Nonresponse

Sampling weights. Having completed the protocols of our 379 respondents by neutrally coding their nonsubstantive responses, we are now in a position to handle unit (enrollee) nonresponse by means of case weighting. As noted above, the population frame for these field-test data consisted of health plan enrollees in California, Michigan, and Connecticut. This frame included a subpopulation of *managed-care* enrollees stratified over these three sites and a subpopulation of *fee-for-service* enrollees stratified over Michigan and Connecticut only (Williams et al., 1997).

In this study the total sample was evenly split between enrollees covered by each type of health plan, and it is necessary to weight the cases in our subsamples of *fee-for-service* and *managed care* enrollees to adjust for their different inclusion probabilities. The sampling weight for each case is the inverse of its probability of inclusion from its population frame assuming 100% response. Therefore, in each of the following analyses the sampling weights used by Williams et al. (1997) required a poststratification adjustment for unit (enrollee) nonresponse.

Poststratification weights. The subsample of 379 respondents described above consists of 192 *fee-for-service* enrollees and 187 *managed care* enrollees. The 192 *fee-for-service* sampling weights were adjusted to the known Michigan and Connecticut frame counts from which these 192 cases were drawn (Cox, 1997). Similarly, the 187 *managed care* sampling weights were adjusted to their known Michigan, Connecticut, and California population totals. With these poststratified case weights, the following analyses of the 192 and 187 respondents became representative of the *fee-for-service* and *managed care* site frames.

Results

As already noted, neutral coding of nonsubstantive responses gives item commensuration on the scale in Figure 3. Hence ternary equation (5a) has been used to calculate all nineteen weighted MCLs in Table 4. These values are presented by type of care. They measure

satisfactions as social utilities from the provider perspective, i.e., from completed samples that downweight each MCL by the number of respondents *not* experiencing the problem posed by its CAHPS item.

In Table 4 item-by-item (standard normal) z tests show seven care-type differences to exceed the .02 level of significance. Five of these items favor the fee-for-service group and two favor the managed-care group. Three items reached the .04 level, with fee-for-service satisfaction exceeding that for managed care in each of these cases. All item locations in Table 4, except for *preventive care not encouraged* in the managed care sample and *waited more than 15 minutes* in both samples, differ from the satisfaction scale origin at the .000 level of significance.

Table 4 reveals considerable variation over items and care types in social utilities of health care. The first column in Table 4 shows fee-for-service satisfaction ranging from .15 for *waited more than 15 minutes* to 1.94 for *doctors did not provide tests or treatment.* The highest MCL in the second column is observed for *respected by office staff,* whereas *preventative care not encouraged* has the lowest satisfaction value of -.18.

The results in Table 4 confirm those in Table 2, and they reach higher levels of statistical significance due to the larger, completed samples constructed by neutral coding. The standard errors of the MCLs in Table 4 are the square roots of corresponding diagonal elements in a consistent estimator of the covariance matrix of our MCLs. The standard errors of the first 14 items in Table 4 are considerably lower than those of their binary counterparts in Table 2. Moreover, the off-diagonal elements of the MCL covariance matrix provide the covariances that are missing in the analysis exhibited in Table 2. These estimated covariances will generate more precise z tests for contrasting different item satisfactions within the same (fee-for-service or managed-care) sample. (See paragraph above Table 2.)

Table 4

Social Utility MCLs for the Complete Samples

Item	Fee for Service			Managed Care	
Less involved than you wanted	1.53	(.191)	*	.91	(.166)
Preventative care not encouraged	.72	(.154)	*	-.18	(.146)
Did not know history	1.46	(.182)	*	.88	(.157)
Not get same day appointment	.49	(.144)		.57	(.146)
Not get appointment as soon as wanted	.83	(.145)		.61	(.142)
Waited more than 15 minutes	.15	(.134)		.20	(.135)
Not get off-hours help	.58	(.192)		.61	(.161)
Not get daytime help	1.23	(.169)		1.31	(.167)
Not able to see specialist	1.30	(.185)		1.13	(.174)
Specialist care did not meet needs	1.35	(.209)		1.25	(.202)
Doctors did not provide tests or treatment	1.94	(.236)	†	1.30	(.178)
Plan did not pay for tests or treatment	.82	(.146)	*	1.36	(.185)
More paperwork than reasonable	.49	(.138)	*	1.55	(.427)
Got no information from customer service	1.62	(.262)	†	.92	(.207)
Listened carefully	1.37	(.168)	†	.91	(.149)
Explained things	1.72	(.190)	*	1.05	(.154)
Respected what you said	1.55	(.179)	*	1.01	(.153)
Spent enough time	1.05	(.155)		.67	(.142)
Respected by office staff	1.77	(.196)		1.65	(.190)

Note. Standard errors are in parentheses. A star (*) and a dagger (†) indicate significance levels beyond .02 and .04 (two-tailed).

Marginal Archival Data

An important property of the MCL is that this descriptive statistic and its standard error can be calculated from the marginal distribution of item i alone. This item separability is essential for marginal archival data and/or large survey datasets where case-by-case response protocols are not available for reasons of privacy or practicality. It is also essential for incomplete datasets, where item nonresponse causes subsample sizes to differ from item to item. For example, using the marginal standard errors in Table 2, a *fee-for-service* versus *managed care* (standard normal) z test was constructed for each CAHPS item. As noted previously, these standard errors may also be used to contrast different item satisfactions within a group of enrollees. However, this conservative within-group z test understates significance level due to its spuriously large difference variance that neglects an inter-item covariance. This (unknown) covariance stems from common respondents shared by two different item subsamples from, say, *fee-for-service* enrollees. This problem does not arise in comparing MCLs for the same item in two different samples, each plagued by item nonresponse. Thus in Table 2 the z contrast between samples is not conservative because the two item subsamples, having no respondents in common, generate marginal standard errors without an accompanying covariance term.

For complete datasets the standard errors, such as those in Table 4, may be found marginally or as square roots of the diagonals of a consistent estimator of the covariance matrix of our MCLs. The off-diagonals of this matrix provide the covariances that are missing in any marginal item analysis. These estimated covariances generate more precise z tests for contrasting different item satisfactions within the same (*fee-for-service* or *managed-care*) sample. As the present paper emphasizes, however, use of the MCL is not confined to complete datasets. In practice the user is free to decide between varying subsample MCLs like those in Table 2 or complete sample MCLs such as those in Table 4. In health care applications these two options represent patient and provider perspectives.

Conclusions

The preceding analyses have illustrated the following advantages of MCLs over status-quo rating methods. These advantages suggest that MCLs should replace mean item ratings in health care questionnaires.

Inter-Item Commensuration of MCLs

In conventional item analysis binary and ternary items like those in Table 1 are scored 0,1 and 0,1,2 respectively. However, mean item ratings over the doubly censured intervals [0,1] and [0,2] are not commensurate and, therefore, do not allow inter-item comparisons. In contrast, Figure 1 shows a very different conceptualization that represents ordered response labels as successive *intervals* rather than successive *integers*. The binary and ternary MCLs in (4a) and (5a) *are* fungible, even though they are coded by different numbers and types of response labels. Due to their common metric these MCLs allow statistical contrasts between binary and ordinal items. They also may be averaged to construct a multi-item indicator that is located on the same numerical scale as the MCLs themselves. This aggregation produces an indicator whose standard error is much lower than that of a single-item MCL. Thus a very small average perceptual change may be detected at an extremely high level of significance.

The commensuration property of MCLs allows flexibility in tailoring response labels to each item in a multi-item indicator. For example, as in Table 1, certain items may compel graded response options, whereas other items may dictate binary responses. It should be emphasized, however, that the common metric for MCLs should *not* be confused with the concept of unidimensionality in psychological test theory (Samejima, 1969; Drasgow, 1995; van der Linden, 1996). Again referring to Tables 1 and 4, we can meaningfully say "Patients are more satisfied with the respect they receive (ternary) than with their preventive care (binary)." These two items would appear to be on different health care dimensions, but their common satisfaction metric is very useful in health care policy and research.

Protections on MCLs

Grouping of response categories. Over the years there has been much discussion and debate as to how many response categories should be used for rating scales (Jones, 1960: Ramsay, 1973; Cox, 1980; McCullagh, 1980; Krosnick and Fabrigar, 1997). A practical and theoretical desideratum would be, of course, that the number of categories is irrelevant to the measures constructed from a given questionnaire. In short, these measures should be invariant over varying numbers of response categories that might be used to obtain them. From both applied and scientific points of view, this property is essential for two measuring devices, i.e., two different response formats should produce the same results.

In the present paper invariance of item MCLs is preserved when the four categories *never, sometimes, usually, always* are collapsed to three as shown in Table 1. This response grouping is justified on both theoretical and empirical grounds. In discussing the invariance of cumulative logit models, McCullagh (1980, p. 116) notes that

> ...varying conditions could mean a redefinition of the response categories, grouping or merging of the categories or the splitting of categories. Hence the parameter or parameters of interest should not depend *for their interpretation* on the actual response categories involved although the estimate will in general be affected. This property permits testing the consistency of various sources of information and, if warranted, combining information from the separate sources. All the models advocated in this paper share the above property: log-linear models...do not.

More recently, Crouchley (1995, p. 491) has reiterated this major feature of cumulative logit models, namely, that their parameters "are not affected by the grouping, merging, or splitting of response categories." The same invariance of Thurstonian "successive intervals" scales was emphasized earlier by Jones (1960, p.11) who, in citing a variety of studies, stated:

> It may be concluded that when the same stimulus items are presented to different random samples from the same popu lation, results of scaling by the method of successive categories are identical over changes in the number of categories on the rating form and changes in descriptive-phrase labels of the categories. Neither of these forms of invariance is found for statistics computed from arbitrary integers assigned to the categories.

Bounds for item nonresponse. Health care items like *not able to see specialist* have high skip frequencies in questionnaires because they are not applicable to many respondents. This nonresponse count, which indicates neither positive nor negative item evaluation, is neutrally coded here to generate a whole-sample MCL as a social utility. The preceding analysis shows that these MCLs are protected from excessive shrinkage for item nonresponse. This protection stems from a lower bound, even with large nonresponse flows into a neutral category. (See Table 3.) The shrinkage of the MCL toward this limit adjusts it for lower incidence of positive and negative responses in the population. Conventional mean ratings, in contrast, are not protected by a lower bound. Therefore, they implode to the value assigned to the neutral category itself when there are excessive nonresponse flows into this category.

Affected Utility versus Social Utility

Inter-item commensuration of MCLs, along with their built-in protections, allow the dual patient and provider analyses developed in the present study. The patient perspective is achieved for each item by using *only* the subsample of respondents who are *affected* by the health care problem or interaction posed by a questionnaire item. (See Table 1.) Despite varying subsample sizes over different items, the item MCLs are scaled in a common metric. (See Figures 1 and 2.) This inter-item commensuration allows affected patient utilities to be compared across different aspects of health care. (See Table 2.)

The provider perspective is given by the whole-sample analysis in Table 4 which is made possible by neutral nonresponse coding.

As shown in Figures 4 and 5, this coding corrects each affected patient MCL for its subsample size. This correction gives a new MCL that is adjusted for the frequency of positive and negative item responses in the total sample. Thus an affected patient utility based on a handful of respondents is downgraded to a lower social utility for the entire sample (cf. Arrow, 1951; Luce and Raiffa, 1957, ch. 14; Coombs, 1964, ch. 18). Importantly, this latter MCL is protected by a lower bound that prevents excessive shrinkage, even in the case of health care items that rarely elicit positive and negative responses. (See Table 3.)

Tables 2 and 4 show that from both patient and provider perspectives the perceived quality of health care is greater for fee-for-service plans than managed care plans in the field test population studied here. These tables illustrate a dual template for health care evaluation that is potentially useful for consumers and suppliers alike.

Acknowledgements

This project was supported by grant number R03 HS09550-01 from the Agency for Healthcare Research and Quality (AHRQ). The data for the present report was provided by the Research Triangle Institute (RTI). The analysis of these data was carried out with *PerceptiMetrics*® software supplied by the Florida Research Institute. The author thanks Sonia Holowatsch for the associated computer runs. The views and conclusions herein are solely those of the author and do not represent AHRQ or RTI.

References

Agency for Healthcare Research and Quality. (1996). *Technical Overview of Consumer Assessment of Health Plans (CAHPS).* Silver Spring, MD: AHRQ Publication Number 97-R013.

Andrews, F.M. & Withey, S.B. (1976). *Social Indicators of Well-Being.* New York: Plenum Press.

Arrow, K.J. (1951). Social Choice and Individual Values. New York: Wiley.

Bechtel G.G. (1977). A model for monitoring consumer satisfaction. In H.K. Hunt (Ed.), *Conceptualization and Measurement of Consumer Satisfaction and Dissatisfaction* (pp. 187-214). Cambridge, MA: Marketing Science Institute.

Bechtel, G.G. (1978). Life-quality and consumer satisfaction: A measurement model and some survey results. In F.D. Reynolds & H.C. Barksdale (Eds.), *Marketing and the Quality of Life* (pp. 42-50). Chicago: American Marketing Association.

Bechtel, G.G. (1983). Consumer satisfaction with foods and their attributes. In R.L. Day & H.K. Hunt (Eds.), *International Fare in Consumer Satisfaction and Complaining Behavior* (pp. 69-74). Bloomington, IN: Foundation for the School of Business, Indiana University.

Bechtel, G.G. (1991). Probabilistic dimensionality: A study of confidence and intention. In D.R. Brown & J.E.K. Smith (Eds.), *Frontiers in Mathematical Psychology: Essays in Honor of Clyde Coombs* (pp. 80-109). New York: Springer-Verlag.

Bechtel G.G., Vanden Abeele, P., & De Meyer, A-M. (1993). The sociotropic aspect of consumer confidence. *Journal of Economic Psychology.* **14,** 615-633.

Clogg, C.C. (1979). Some latent structure models for the analysis of Likert-type data. *Social Science Research,* **8**, 287-301.

Coombs, C.H. (1964). *A Theory of Data.* New York: Wiley.

Cox, B.G. (1997). Weighting survey data for analysis. Presentation for the *1997 International Conference on Health Policy Research,* pp.1-22. Alexandria, VA: American Statistical Association.

Cox, E.P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research,* **17**, 407-422

Crouchley, R. (1995). A random-effects model for ordered categorical data. *Journal of the American Statistical Association,* **90**, 489-498.

Curtin, R.T. (1972). Index construction: An appraisal of the index of consumer sentiment. In *1971-72 Surveys of Consumers* (pp. 253-261). Ann Arbor, MI: Institute for Social Research, University of Michigan.

Drasgow, F. (1995). Special issue: Polytomous item response theory. *Applied Psychological Measurement,* **19,** 1.

Gore, A. (1993). *Improving Customer Service: Accompanying Report of the National Performance Review.* Washington, DC: Office of the Vice President, U.S. Government Printing Office, Superintendent of Documents.

Jones, L.V. (1960). Some invariant findings under the method of successive intervals. In H. Gulliksen & S. Messick (Eds.), *Psychological Scaling: Theory and Applications* (pp. 7-20). New York: Wiley.

Katona, G. (1975). *Psychological Economics.* New York: Elsevier.

Katona, G. (1979). Toward a macropsychology. *American Psycholo gist,* **34**, 118-126.

Krosnick, J.A. & Fabrigar, L.R. (1997). Designing rating scales for effective measurement in surveys. In L.E. Lyberg et al. (Eds.), *Survey Measurement and Process Quality* (pp. 141-164). New York: Wiley.

Landis, J.R., Miller, M.E., Davis, C.S., & Koch, G.G. (1988). Some general methods for the analysis of categorical data in longitudinal studies. *Statistics in Medicine,* 7, 109-137.

Levy, S. & Guttman, L. (1975). On the multivariate structure of well-being. *Social Indicators Research,* **2**, 361-388.

Longford, N.T. (1999). Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society, Series A*, **162 (No. 2)**, 227-245.

Luce, R.D. & Raiffa, H. (1957). *Games and Decisions.* New York: Wiley.

Mann, N.R. (1994). W. Edwards Deming 1900-1993. *Journal of the American Statistical Association,* **89**, 365-366.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B,* **42 (No. 2)**, 109-142.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models (2nd ed.).* London: Chapman & Hall.

Ramsay, J.O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika,* **38**, 513-532.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Number 17*, **34** (Number 4, Part 2).

Torgerson, W.S. (1958). *Theory and Methods of Scaling*. New York: Wiley.

van der Linden, W.J. & Hambleton, R.K. (Eds.) (1996). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.

Williams, V. S. L., Burnbauer, L., Lubalin, J., Sweeny, S. F., Ardini, M. & Chromy, J. (1997). *Adult Privately Insured Survey (API) Field Test Analysis Report*. Washington, DC: Research Triangle Institute.

JOURNAL OF OUTCOME MEASUREMENT®, 4(3), 667-680

Maintaining Instrument Quality While Reducing Items: Application of Rasch Analysis to a Self-Report of Visual Function

Craig A. Velozo

Jin-Shei Lai

Trudy Mallinson

Ellyn Hauselman

While efficiency has been of concern in the measurement of health care outcomes, little attention has been devoted to methods that achieve efficient, shortened instruments that have good psychometric properties. The purpose of this study was to show how Rasch analysis could be used to reduce the number of items in an instrument while maintaining credible psychometric properties. This approach was applied to the Visual Function-14 (VF-14), a self-report of 14 vision-dependent activities, designed to measure the need for and outcomes of cataract surgery. An instrument which contained the VF-14 plus an additional 10 items that were developed for the study (VF-24) was administered to sixty-one patients (73.7±9.5 years) about to undergo extracapsular cataract removal at one of two surgical centers. Rasch analysis (BIGSTEPS) of the VF-14 showed a number of limitations to the original instrument, including: 1) unequal use of the five rating

Requests for reprints should be sent to Craig A. Velozo, Department of Occupational Therapy, University of Florida, P.O. Box 100164, Gainesville, FL 32610-0164

categories, 2) ceiling effect, 3) several other gaps where patient abilities did not match with item difficulties, and 4) sets of items that appeared redundant, (i.e., having the same calibration level). To resolve the first three of these problems, the rating scale was converted to a three-point scale and BIGSTEPS was run with all 24 items. (10 additional items added to the VF-14 designed to "fill in" the gaps). The conversion to a three-point scale and the increase in items resulted in some improvement in the matching of item difficulty to patient ability, as evidenced by a slight decrease in gaps. The addition of items resulted in improvements in person separation (2.55 to 2.99) and Cronbach's alpha (.83 to .91) but did not substantially reduce the ceiling effect and furthermore resulted in an increase in item redundancy. The final practical improvement undertaken was to reduce the number of items while attempting to maintain the psychometric qualities of the instrument as a whole. Three criteria were used in deciding to remove items: 1) high mean square, 2) low mean square and 3) items having similar calibrations. In addition, if an analysis showed that the removal of an item substantially decrease person separation, that item was retained for further analyses. Relative to the original VF-14, the resulting VF-10 showed less redundancy of items while person separation (2.20) and Cronbach's alpha (.89) remained relatively intact. The study demonstrates that Rasch analysis, while effective in elucidating the metrics of an original instrument, can also be useful in designing modifications of instruments that are both efficient and psychometrically sound.

Introduction

The most recent era of health measurement has been characterized as that of "psychometric efficiency" (McHorney, 1997, p. 744). This has been exemplified by the current "short" generic health scales such as the Duke Health Profile-17, Medical Outcome Scale (MOS) Short-form 36 and MOS Short-form-12, having evolved from the large multi-item instruments of the 1970s (McHorney, 1997; Parkerson, Broadhead, and Tse, 1990; Stewart, Sherbourne, Hays, Wells, Nelson, Kamberg, et al., 1996; Ware, Sherbourne, 1992; Ware, Kosinski and Keller, 1996). Shorter instruments were developed as the result of several demands such as the costs of large scale clinical trials studying the broad spectrum of health care and the burden severely ill patients face when taking lengthy surveys (McHorney, 1997, p. 744-745).

Due to the shear volume, cataract surgery is an area of health care that demands efficient outcomes measurement. More than 2 million cataract extractions are performed each year on Medicare beneficiaries (Health Care Financing Administration, 1995). In 1994, Medicare payment amounted to

over $1.4 billion for cataract removals and related procedures (Health Care Financing Administration, 1995). In an effort to curb the spiraling costs of cataract surgery, in 1995, Bruce C. Vladeck, administrator of the Heath Care Financing Administration (HCFA), stated "Surgery would not be considered necessary if the individual is able to function normally or if the condition can be corrected with eyeglasses" (Health Care Financing Administration, 1995).

In response to HCFA's request, the Visual Function - 14 (VF-14) was developed through a Patient Outcomes Research Team (PORT) project funded by the Agency for Health Care Policy and Research (AHCPR). This self-report of 14 vision-dependent activities was designed to measure the need for and outcomes of cataract surgery (Steinberg, et al., 1994). The popularity of this instrument is reflected in it being adopted by the American Academy of Ophthalmology as one of their primary instruments for measuring outcomes (NEON, 1997).

Psychometric studies of the VF-14 have been based on true-score test theory methodology. That is, these studies have focused on traditional measures of reliability and validity. For example, Steinberg et al. (1994) in a study of 766 patients undergoing cataract surgery, showed that the VF-14 produced a Cronbach's alpha of .85. In addition, Cassard, et al. (1995), in a study of 552 patients who had undergone cataract surgery in one eye, showed that the VF-14 produced an interclass correlation for reproducibility (between 4 and 12 months post operatively) of .79.

With the onset of modern test theory, psychometric evaluations of the VF-14 based on Item Response Theory (IRT) methodologies, such as Rasch analysis, could provide valuable information. For example, using Rasch analysis, the match of VF-14 items to the anticipated population could be elucidated. Furthermore, consistent with the need to have efficient health measurement systems, IRT methodologies, such as Rasch analysis, could be used as a basis to reduce the number of items in an instrument while attempting to maintain credible psychometric properties.

The purpose of this study was threefold: first, to identify the psychometric properties of the VF-14 using Rasch analysis. Second, to determine if the psychometric properties of the VF-14 could be improved with the addition of items. Third, from the set of combined items, attempt to remove items to create an instrument comparable to the VF-14 psychometrically, but with fewer items. Overall, the objective of this study was to achieve a shortened instrument that was well-targeted to the anticipated

population and had acceptable psychometric properties.

Methods

Instrument

The VF-14 is a questionnaire designed to measure deficits in vision-dependent activities caused by cataract (Steinberg, et al., 1994, p. 634). It consists of 14 questions that are a subset of those commonly asked by ophthalmologists when considering a patient for cataract surgery (Steinberg, et al., 1994, p. 631). The set of 14 vision-dependent activities comprising the VF-14 instrument are presented in Table 1.

Table 1

Vision-Dependent Activities Comprising the VF-14

1. Reading small print
2. Reading a newspaper or book
3. Reading a large-print book or newspaper or the numbers on a telephone
4. Recognizing people
5. Seeing steps, stairs, or curbs
6. Reading traffic, street or store signs
7. Doing fine handwork
8. Writing checks
9. Playing games such as bingo
10. Taking part in sports such as bowling
11. Cooking
12. Watching television
13. Daytime driving
14. Nighttime driving

The question for each activity is presented as, "Do you have any difficulty, even with glasses (activity listed above)," whereby the patient answers "yes" or "no". If the answer is "yes" a follow-up question is presented as, "If yes, how much difficulty do you currently have?" There are 5 response categories for this follow-up question: *no difficulty, a little difficulty, a moderate amount of difficulty, a great deal of difficulty* and *unable to do activity.*

Subjects

Instruments which contained the VF-14 plus an additional 10 items that were developed for the study (VF-24; see Table 2) were administered to sixty-one patients from two surgical centers just prior to extracapsular cataract removal. Sixty-nine percent of patients were female and 31% percent were male. The average age of the sample was 73.7±9.5. Fifty-one percent of the patients were receiving cataract surgery for the first time, 28% for the second time (on the second eye) and for 21% of the sample this data

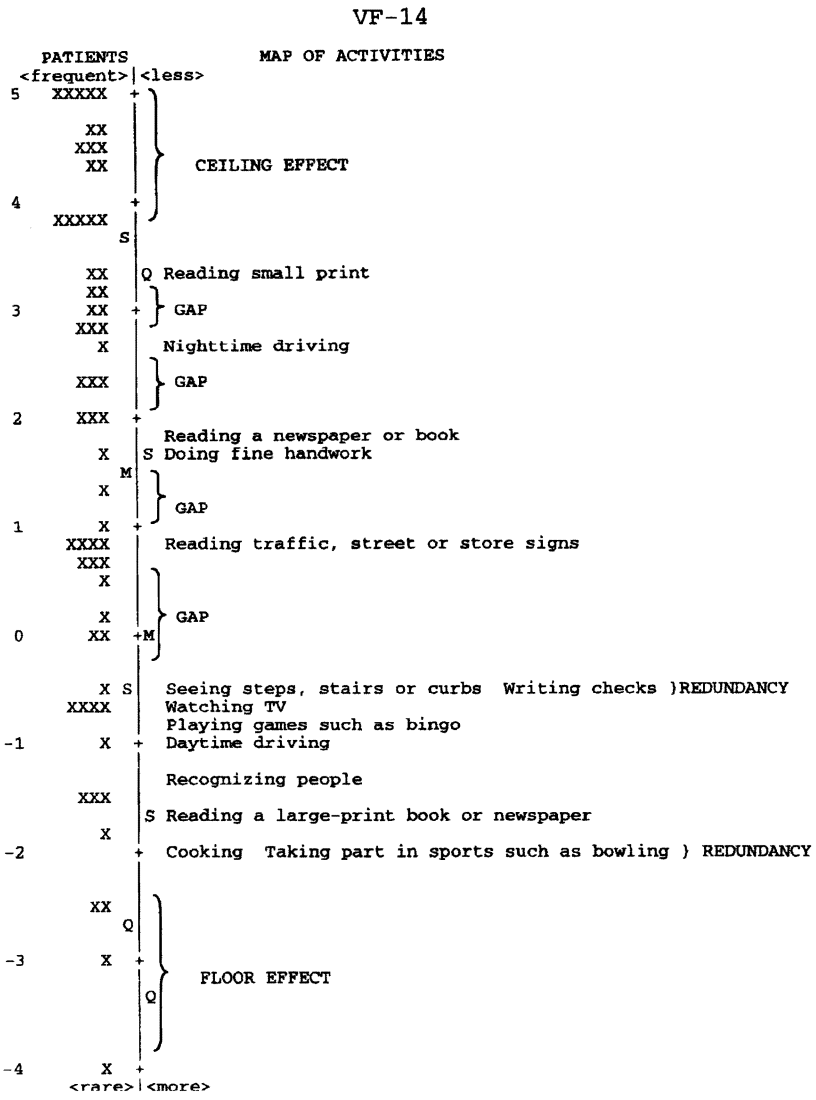

Figure 1. Map of person ability measures against the mean of item difficulty calibrations for the VF-14. The scale is presented in log equivalent units (logits) with the more able persons (X's on the left side) and more difficult items (labeled on the right side) appear toward the top of the figure and the less able persons and less difficult items appear toward the bottom of the figure. *Gaps* are indicated where items are separated by more than 2 standard errors. Similarly, *ceiling effects* and *floor effects* are indicated where person abilities are not matched well with item difficulties at the top and bottom of the scale, respectively. *Redundancies* are indicated where items are at the same calibration level. Abbreviations: M=mean, Q=1 standard deviation, S=2 standard deviations.

was not available.

Results

Initial findings from the BIGSTEPS analysis (Wright and Linacre, 1998) indicated that while the VF-14 originally had five rating categories, all five categories were not being used with similar frequencies. Each of the three lowest categories, *unable to do activity, a great deal of difficulty* and *a moderate amount of difficulty* had only 2-10% of the responses, in contrast to the two highest categories, *a little difficulty* and *no difficulty* which had 21-57% of the responses. These findings suggest that the respondents had a tendency to report little to no difficulty with visual function. Therefore, the three lowest categories were combined for all further analyses. Conceptually, the new set of categories could be thought to represent *a moderate amount of difficulty, a little difficulty* and *no difficulty.*

The change in the rating scale resulted in slight improvements in person and item separation but virtually no change in Cronbach's alpha. Person separation improved from 2.44 to 2.55 and item separation increased from 4.12 to 4.48. Cronbach's alpha decreased from .84 to .83. Patients with extreme scores increased from 5 to 6. This increase resulted in the appearance of one patient with a minimum extreme score that was not evident with the 5-point scale analysis.

While the VF-14 showed slight improvements in person and item separation with the change in the rating scale structure, the instrument continued to show limitations in measurement qualities. Figure 1 presents a map comparing VF-14 item difficulty measures to patient ability measures. First, the instrument showed a ceiling effect with sixteen percent of the patients (12/61) having measures that are at least two standard deviations higher than the average measure for the most difficult item, *Reading small print.* Five of these patients showed extreme maximum scores or scores that were not measured by the instrument. Second, there appeared gaps where items were separated by at least 2 standard errors and therefore in most cases did not separate patients into different abilities (see Figure 1). For example, there are seven patients who showed ability measures located between the item *Reading traffic, street or store signs* and the set of items *Seeing steps, stairs or curbs* and *Writing checks*. It should be noted that gaps are not the only indications of the limitations of the VF-14 in differentiating patients, but also the "piling up" of patients at the same calibration level. For example, at .80 logits there were four patients with the same ability level as indicated by 4 "X's". Since all four of these patients were at the same

calibration level, the instrument did not differentiate their abilities. Finally, in spite of there being a fairly small numer of items, there are two sets of items that appeared to be redundant; items that were at the same calibration level (see Figure 1). *Seeing steps, stair, curbs* was at the same calibration level as *Writing checks* and *Cooking* is at the same calibration level as *Taking part in sports such as bowling* (see Figure 1). It should be noted that there was only a slight floor effect with 1.6 percent (1/61) of the patients scoring two standard deviations below the mean calibration of the easiest item.

To resolve the first two of these problems, 10 additional items were added that were designed to "fill in" the gaps. Then BIGSTEPS was run with the total 24 items. We refer to the VF-14 in combination with these ten items as the VF-24. Table 2 presents the ten additional items included in the VF-24. Analysis of all 24 items showed that relative to the VF-14, the VF-24 showed improvements in person separation (from 2.55 to 2.99) and improvements in Cronbach's alpha (.83 to .91). With the increase in the number of items, as expected, item separation decreased (from 4.48 to 3.72). In addition, with the increase in items, the number of patients with extreme scores dropped from 6 to 3.

Figure 2 presents the map from the analysis of the VF-24. The original VF-14 items used are in normal text, while the ten newly developed items are in bold/italic text. With the increase in items, there was an associated slight decrease in the number of gaps, from 4 with the VF-14 to 3 with the VF-24. The ceiling remained essentially the same, increasing slightly to 21.3 percent (13/61) of the sample, but there was an apparent increase in redundancy. That is, there were many more items that were at the same level of difficulty. The VF-14 showed 2 sets of 2 items that were redundant (see Figure 1) while the VF-24 showed 6 sets of 2-3 items that were redundant (see Figure 2). It should be noted that there were no patients that scored more than 2 standard deviations below the mean of the easiest item on the instrument. In conclusion, relative to the VF-14, the VF-24 showed some improvement in the matching of item difficulty to patient ability, as evidenced by a slight

Table 2

Ten Additional Functional Activities Comprising the VF-24

1.	Seeing moving objects at night	6.	Writing address
2.	Seeing steps at night	7.	Read clock
3.	Walk on different surfaces	8.	See dials on radio
4.	Read a watch	9.	Identify coins
5.	Read a tape measure or ruler	10.	Avoid bumping into things

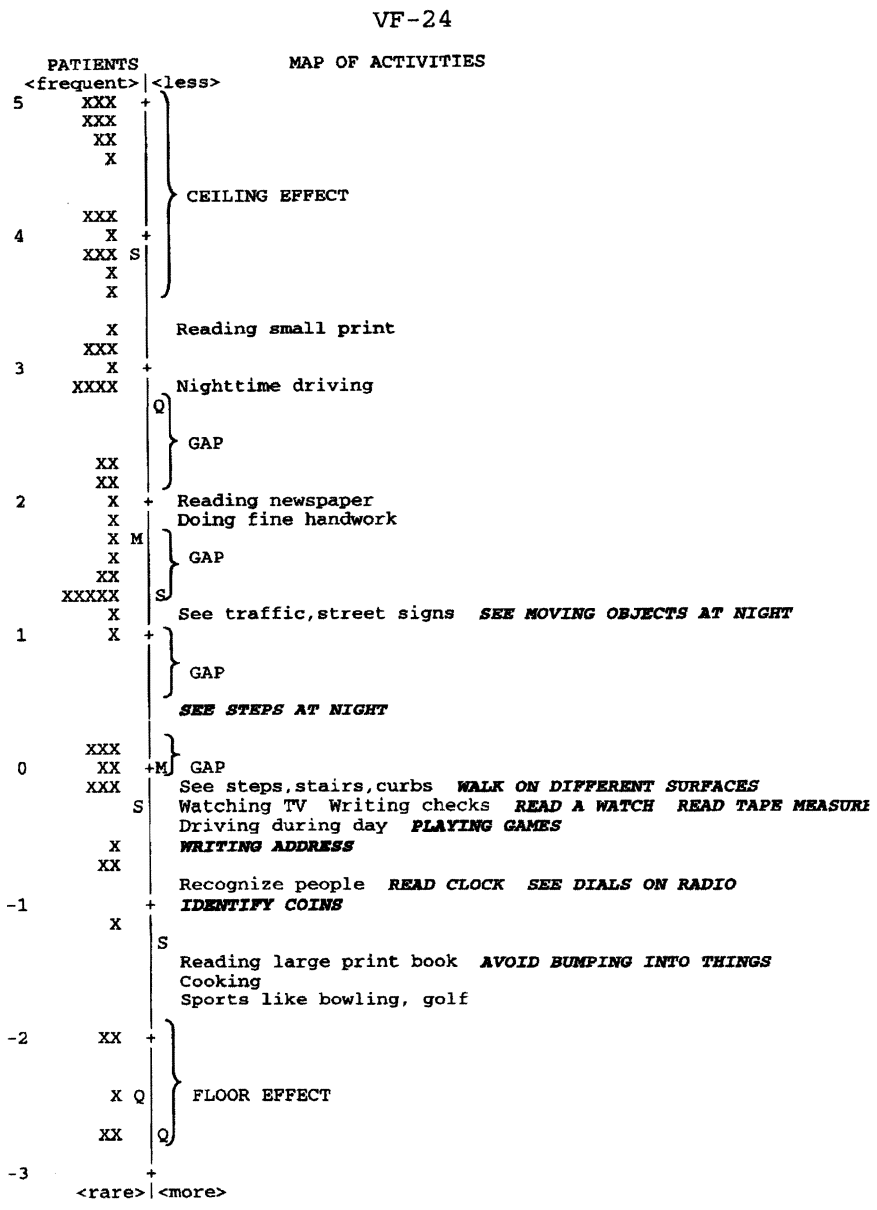

Figure 2. Map of person ability measures against the mean of item difficulty calibrations for the VF-24. See Figure 1 for explanations of symbols and labels. The original VF-14 items used are in normal text while the ten newly developed items are in bold/capital/italic text.

decrease in gaps, but there still remained a ceiling effect and there was an increase in the redundancy of items.

While the ceiling effect was still a concern, from a practical point of view, there seemed little that could be done to eliminate this effect. Pragmatically, if items were added, the study would have to be re-run with a new sample and comparisons across samples could lead to considerable misinferences. Furthermore, there did not seem to be many common activities that were more challenging than *Reading small print* which was the most difficult item on the instrument. Therefore, the next practical improvement we could make in the instrument was to reduce the number of items to decrease redundancy while attempting to maintain the psychometric qualities of the instrument as a whole.

Three criterion were used in deciding to remove items: 1) high mean square, 2) low mean square and 3) items having similar calibrations. In addition, if an analysis showed that the removal of an item substantially decreased person separation, that item was retained for further analyses. Following a series of analyses, we arrived at a ten-item instrument (VF-10) which is presented as part of Figure 3.

Figure 3 represents the original map of the VF-14 and the VF-10 resulting from the above procedure. The map of the VF-10 indicates items from the original VF-14 in normal text and the two items in underline/bold/italic text, *See moving objects at night* and *See steps at night*, that were from the VF-24. There are several differences between the maps of the two instruments. First, the VF-10 presents a slightly better spread of patient ability measures. Ability measures for the VF-14 range from –4 to 5 logits while ability measures for the VF-10 range from –5 to 5 logits. Second, in addition to the increase in range of ability measures, the VF-10 shows a better differentiation of patients below 0 logits. This is indicated by the VF-10 map displaying patients in this range with different ability measures (not "piling up" at a particular calibration), while the VF-14 map shows 2-4 patients at -.75, -1.5 and –2.5 logits. Third, by design, the VF-10 shows less redundancy of items. That is, the VF-10 shows fewer items at the same or similar calibration levels. It should be noted that the redundancy of items (*See traffic, street sign* and *See moving objects at night*) at approximately 0.5 logits was unavoidable. When either of these items was removed, there was a considerable reduction in person separation. Finally, with the reduction in items represented by the VF-10, person separation decreased from 2.55 to 2.20, item separation increased from 4.48 to 5.07 and Cronbach's alpha increased from .83 to .89.

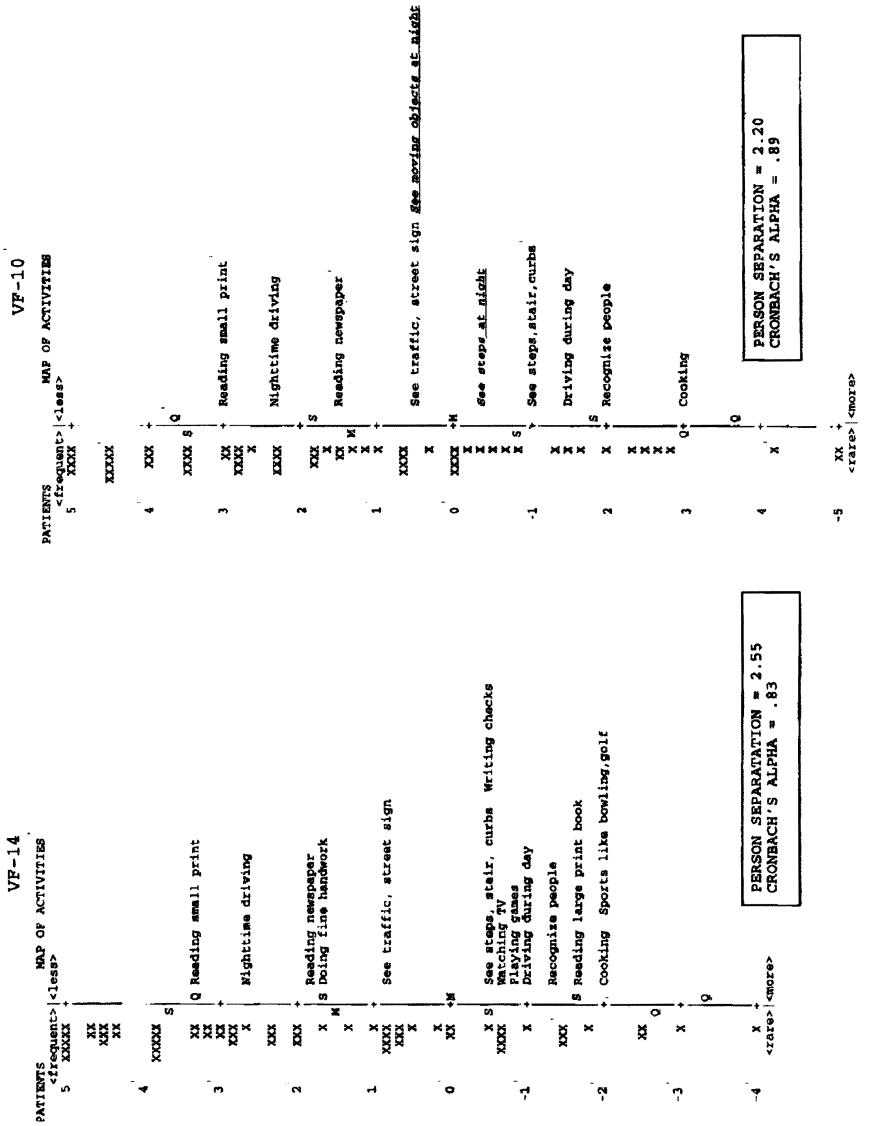

Figure 3. Map of the VF-14 and map of the VF-10 (reduction of the VF-24; see text for explanation). The map of the VF-10 indicates items from the original VF-14 in normal text and the one item in bold/italic text, *See steps at night*, which was from the VF-24. See Figure 1 for explanations of symbols and labels.

As expected, the map of the VF-14 continues to show the ceiling effect and gaps, since the item-reduction procedure was not intended to reduce either of those characteristics.

Discussion

The general purpose of this study was to demonstrate how Rasch analysis could be used to reduce the number of items in a self-report of visual function while monitoring and maintaining psychometric qualities of that instrument. Previous psychometric studies have focused on the internal and test-retest reproducibility of the VF-14 (Cassard, et al., 1995; Steinberg, et al., 1994). While the reliability statistics generated from these studies can be compared to existing published criterion, they provide little direction for making improvements in the instrument. In contrast, IRT analyses, such as Rasch, furnish information about an instrument (e.g., person/item calibrations and fit statistics) that can be used to make logical modifications to achieve improved precision and efficiency in measurement.

The original development protocol of the VF-14 involved modification of the items by an 11-member National Advisory Panel of ophthalmologists and optometrists (Steinberg, et al., 1994). In this modification phase, the panel was asked to do the following: 1) identify the items on the list that they believed best reflected the full spectrum of functional limitations experienced by patients with cataract: 2) delete or combine items from the list that they believed constituted functionally equivalent tasks from a vision perspective, and 3) specify relevant functional activities that had not yet been identified (Steinberg, et al., 1994). The reliability studies performed on the VF-14 did little to verify whether or not the national panel was successful in accomplishing the above goals.

Rasch analysis, on the other hand, could have been useful in addressing each of the issues that were the focus of the national panel review. The determination of whether or not the items of the VF-14 covered the full spectrum of limitations experienced by patients with cataract can be evaluated by the comparison of item difficulty calibrations to person ability measures. If *functionally equivalent* tasks can be operationally defined as items that have similar difficulty calibrations, once identified, these items can be either combined or deleted (we chose to delete equivalent items). Finally, while no analysis can provide *unidentified relevant* activities, Rasch analysis can identify *gaps* where item-difficulty calibrations do not match person-ability measures. These gaps can provide direction in developing new items that can better differentiate patients

of similar functional vision abilities.

Based on Rasch analysis, the findings of this study suggest that the national panel did not accomplish their first objective, that is to develop a bank of items which successfully captured the full spectrum of functional limitations experienced by patients with cataract. Prior to surgery, 16% (12/61) of the patients in this study reported being able to accomplish reading small print, the most difficult item on the instrument. While it is possible that the sample in this study was biased, showing extraordinarily high visual function, the study findings suggest that the items of the VF-14 do not reflect the full spectrum of functional limitations. It is possible that visual deficits that are detected with clinical measures, such as glare, are not being captured by the functional measures derived from the VF-14 (Carta, Braccio, Belpoliti, Soliani, Sartore, Gandoflfi and Maraini, 1998; Grover, Alexander, Choi and Fishman, 1998).

The second objective of the panel, to combine or delete functionally equivalent tasks, was the central focus of this study. Rasch analysis of the VF-14 showed that several items were at the same calibration level (e.g., *Steps, stairs, and curbs* and *Writing checks*). While qualitatively, these items are very different, quantitatively, they appear to be of similar challenge for individuals with cataract. In assessing visual functional ability, from a quantitative perspective, one of these items will not contribute any more information than the other. Therefore, for the sake of efficiency, only one of these items is necessary to be included in the instrument.

The final objective of the national committee was to specify relevant functional activities that had not yet been identified. From a measurement perspective, the most relevant activities are not those that are unidentified, but those that contribute to differentiating patients of similar abilities. The area where items were not differentiating patients was at the high end of the scale where 16% of the patients appeared to have the same visual functional ability and, in effect, appeared to have no functional visual difficulties. As stated earlier, while this may indicate that these patients are not having difficulties in visual functioning, it may suggest that there are areas of visual functional deficits that are not being addressed by the instrument, for example in the area of glare.

The overall purpose of this study was to demonstrate that items of an instrument can be reduced while the psychometric qualities of an instrument are maintained. While a circumscribed set of qualifiers were used in this study (e.g., Cronbach's alpha, person separation), it is apparent that

these values can be maintained as items are removed from an instrument. The purpose of this study was not to condone the removal of items from instruments; it is clear, that such procedures can decrease measurement precision (McHorney, 1997). By combining techniques such as computerized adaptive technology along with IRT methodologies, it may be unnecessary to remove items from an instrument in order to achieve efficiency. These methodologies will allow for the use of large banks of items whereby both precision and efficiency of measurement are accomplished, not by reducing items, but by selectively presenting items that are matched to the ability levels of respondents (Bjorner and Ware, 1998; McHorney, 1997; Velozo, Kielhofner and Lai, 1999).

Until contemporary testing methodologies are more widely accepted, the preference for instruments with relatively few items is likely to continue. Efficiency continues to be a major concern in the measurement of outcomes in health care. True-score psychometric approaches which focus on the entirety of an instrument, provide little direction for identifying what items to retain in the shortened forms of these instruments. In addition to allowing the evaluation of original instruments, IRT methodologies, such as Rasch analysis, can furnish empirical support for retaining or deleting of items to achieve shortened forms of these instruments. Since performance on measures such as the VF-14 may be used to allow or deny health care services, it becomes critical that our instruments or their derivatives are as sensitive as possible in measuring consumers health status, thus providing accurate information as the basis for these important decisions.

Acknowledgement

Funding for the above study was provided by Medirisk of Illinois, Inc.

References

Bjorner, J. V., and Ware, Jr., J. E. (1998). Using modern psychometric methods to measure health outcomes. *Medical Outcomes Trust Monitor, 3*, 12-16.

Carta, A., Braccio, L., Belpoliti, M., Soliani, L., Sartore, F., Gandolfi, S. A., and Maraini, G. (1998). Self-assessment of the quality of vision: association of questionnaire score with objective clinical tests. *Current Eye Research, 17*, 506-511.

Cassard, S. D., Patrick, D. L., Damiano, A. M., Legro, M. W., Tielsch, J. M., Diener-West, M., Schein, O. D., Javitt, J. C., Bass, E. B., and Steinberg, E. P. (1995). Reproducibility and responsiveness of the VF-14: an index of functional impairment in patients with cataracts. *Archives of Ophthalmology, 113*, 1508-1513.

Grover, S., Alexander, K. R., Choi, D. M., and Fishman, G. A. (1998). Intraocular light scatter in patients with choroideremia. *Ophthalmology, 105*, 1641-1645.

Health Care Financing Administration, (1995). Medicare policy proposed for eye surgery. [Online]. Available: http://www.hcfa.gov/scripts/WEBINATOR.EXE?cmd=view&id=3487bf7f42&rr&s0&e0&ls&db=db&disp=all&grsz=10&rv=806&arg=cataract+surgery#tag1 [1998, July 21].

McHorney, C. A. (1997). Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. *Annals of Internal Medicine, 127*, 743-750.

National Eyecare Outcomes Network (NEON) (1997). [Online]. Available: http://www.eyenet.org/member/features/neon.html [21, July, 1998].

Parkerson, G. R., Jr., Broadhead, W. E., and Tse, C. K. (1990). The Duke Health Profile. A 17-item measure of health and dysfunction. *Medical Care, 28*, 1056-1072.

Steinberg, E. P., Tielsch, J. M., Schein, O. D., Javitt, J. C., Sharkey, P., Cassard, S. D., Legro, M. W., Diener-West, M., Bass, E. B., Damiano, A. M., Steinwachs, D. M., and Sommer, A. (1994). The VF-14: an index of functional impairment in patients with cataract. *Archives of Ophthalmology, 112*, 630-638.

Stewart, A. L., Sherbourne, C. D., Hays, R. D., Wells, K. B., Nelson, E. C., and Kamberg, C. J. (1992). Summary and discussion of MOS measures. In Stewart, A.L.,Ware, J.E., ed. Measuring functioning and well-being: The Medical Outcomes Study Approach, Chapel Hill, NC: Duke University Press, 345-371.

Velozo, C. A., Kielhofner, G., and Lai, J-S. (1999). The use of Rasch analysis to produce scale-free measurement of functional ability. *American Journal of Occupational Therapy, 53*, 83-90.

Wright, B. D., and Linacre, J. M. (1998). A User's Guide to BIGSTEPS Rasch-Model Computer Program. Chicago: MESA Press.

JOURNAL OF OUTCOME MEASUREMENT®, 4(3), 681-705

Measuring disability: application of the Rasch model to Activities of Daily Living (ADL/IADL)

T. Joseph Sheehan, Ph.D.

Laurie M. DeChello

Ramon Garcia

Judith Fifield, Ph.D.

Naomi Rothfield, M.D.

Susan Reisine, Ph.D.

University of Connecticut School of Medicine & University of Connecticut

Requests for reprints should be sent to T. Josep Sheehan, University of Connecticut School of Medicine, 263 Farmington Ave., Farmington, CT 06030

The original version of this article contained errors in the tables.

Please see JOM V5, N1, pages 839 to 863 for a corrected version.

JOURNAL OF OUTCOME MEASUREMENT®, 4(3), 706-720

Payment and Provider Profiling of Episodes of Illness of Clinical Illnesses Involving Rehabilitation

Norbert Goldfield

Richard Averill

Jon Eisenhandler

3M Health Information Systems

John S. Hughes

West Haven Veterans Affairs Medical Center, CT

John Muldoon

National Association of Children's Hospitals and Related Institutions

Barbara Steinbeck

Farah Bagadia

3M Health Information Systems

Adapted from Goldfield, N., Averill, R., Eisenhandler, J., Hughes, J.S., Muldoon, J., Steinbeck, B., Bagadia, F. "The Prospective Risk Adjustment System" published in Goldfield N. Physician Profiling and Risk Adjustment. Aspen Publishers, 2nd edition, 1999.

Requests for reprints should be sent to Norbert Goldfield, M.D., 3M Health Information Systems, 72 Laurel Park, North Hampton, MA 01060.

The prevailing trend in American health care finance for the last two decades and likely for the forseeable future, is the movement from a system based on fee for service payments to one dominated by capitation arrangements. At the core of this change is a shifting of risk from payers to providers. Such a fundamental transition is not without its difficulties. This is exemplified by some of the problems experienced by the Health Care Financing Administration (HCFA) as it has begun to encourage beneficiaries to move from traditional fee for service Medicare to capitated HMOs. The most recently published regulations involving risk adjustment of rates of payment for managed care organizations (MCOs) did not find much favor as the statistical power was poor and the clinical meaningfulness of the risk adjustment was highly problematic. Specifically, the risk adjustment explained less than 10% of the variation in costs (compared to approximately 30% when DRGs were first implemented). From a clinical perspective, the methodology only used hospitalization data for adjustment of capitation rates; that is, anyone who was not hospitalized was assumed to be healthy! It is known that payment for rehabilitation services is among the most difficult to understand and predict of all medical care. This article will summarize the development of a new risk adjustment methodology which should be particularly useful for payment and monitoring of episodes of clinical conditions that involve rehabilitation. The article will conclude with directions for future research.

The risk adjustment system discussed in this paper, the Clinical Risk Groups System (CRGS), is a prospective capitation risk adjuster which provides an estimate future health care costs. It will do this by assigning each individual a single capitation risk adjustment category based on an analysis of the medical history and of health care services rendered during a specific period of time. This assignment will be sensitive to the relative severity of the illness and to the presence of multiple conditions.

There are many challenges facing the development of the Clinical Risk Groups System. Current prospective risk adjustment models tend to have limited explanatory power as measured by R^2. This is not surprising. Even discounting the inherently unpredictable effects of random events, new or previously unreported illnesses, uncertainty in disease progression, etc. the environment in which episode technology must function is a difficult one. Clinical risk groups systems must be able to sort through massive amounts of data and accurately estimate a person's future health care expenditures without being overly sensitive to variations in data which stem from factors other than individual health status. This has led Newhouse, et al., (1989) to suggest that the maximum explanatory power of prospective risk adjustment models is an R^2 of about 20%.

If risk adjustment technologies are to succeed, they must focus on the

diagnostic procedure and other information recorded in the medical records or claims of the population which they are trying to predict. They must be able to differentiate between the factors associated with high and low future costs. To do this they should be able to:

- Identify what is important and what is not. For example, minor trauma generally will have minimal impact upon future resource requirements and generally should be ignored.
- Distinguish between diagnoses or groups of diagnoses associated with high and low costs in immediate future. For example, an individual needing rehabilitation services for a recent cerebrovascular accident will usually require more resources than an individual with vertigo.
- Differentiate between less and more severe cases of the same illness. Within a given diagnosis or set of similar diagnoses, there are likely to be significant differences in resource requirements between an individual in the early stages of an illness and an individual in a more advanced stage. For example, a multiple sclerosis patient who is paraplegic will probably require fewer resources in the immediate future than a multiple sclerosis patient that is diabetic, in the advanced stages of disease, who has significant circulatory problems.
- Define the relationship between multiple diagnoses. Individuals may have more than a single recorded diagnosis. Many of these diagnoses will have no effect upon future health care costs, others will indicate differing levels of severity of some underlying condition, and still others will indicate the presence of additional diseases. The technology should be able to distinguish, at the individual level, the importance of each diagnoses relative to an individual's health status. It should be sensitive to the whole constellation of an individual's diagnoses as well as the time frame in which they occurred. For example, patients who have a cerebrovascular accident and then have subsequent transient ischemic attacks after the CVA will likely require more services than a CVA, at the same level of initial severity as the previous patient, who did not have subsequent TIAs. Furthermore, a risk adjustment system that is cognizant of clinical events potentially important for rehabilitation would not incorporate TIAs that occur prior to the CVA.

Efforts to date, primarily ACGs (Ambulatory Care Groups), DCGs (Diagnostic Cost Groups), and their various implementations, have had some success addressing these issues. They have grown increasingly sophisticated and have experienced improvements in predictive power as they have

developed solutions to the problems posed by prospective risk adjustment. While there are many important differences between ACGs and DCGs, both use one of two classification strategies. The first and simpler approach is to place an individual into a single group. The second, and more complex approach, is to assign each individual a unique weight based on membership in multiple groups. The latest commercial version of ACGs uses the first methodology and assigns each individual to one of up to 83 ACGs (John Hopkins University, 1998). The 83 ACGs represent the combination of underlying diagnostic and demographic categories. The simpler form of DCGs is similar in that an individual is assigned a single weight based on the most costly DCG in which the individual has membership (Ellis, et al.,1996). DCGs with HCC (Hierarchical Coexisting Conditions) adopt the second, more complex approach (Ellis, et al., 1996). It allows membership in multiple groups with individual weights equaling the sum of their group weights (albeit subject to some limitations as to which combinations of groups will be allowed) to which the individual belongs. ACGs, or more precisely ADGs, an interim step in ACG assignment, have also been used to experiment with multiple group membership and the prediction of health care utilization in a comparative study commissioned by the American Society of Actuaries (Dunn, et al.,1995). By allowing membership in multiple groups, the multiple group methodologies in effect create an extremely large number of distinct groups with the number of groups equal to the number of allowed combinations.

The Clinical Risk Groups System is a single category methodology. It assigns only a single clinically based risk class. Where it differs from other single category models, is that it offers a far more detailed clinically based classification system. It addresses the problems of differences in severity and multiple group membership by creating specific groups to reflect those differences where and whenever feasible. In effect, it adopts a multiple group methodology by explicitly identifying those groups it wishes to create rather than allowing the groups to interact freely.

This approach has its strengths and weaknesses. The strengths of the Clinical Risk Groups System are threefold. First, its detailed classification system is sensitive to differences between a greater number of illnesses. Second, it differentiates between the relative severity of illnesses. Third, by assigning an individual to only a single risk class, it makes no assumptions about the mathematical nature of the interaction between the identifiable clinical factors which may influence an individual's future health care needs. Its strength however, is also its weakness. A detailed classifica-

tion system can produce cells which are overly sensitive to their underlying data and can become overspecified especially as many cells will be populated by relatively few cases.

Data Preparation

The Clinical Risk Groups System is a categorical model which predicts future health care utilization using clinical information derived from individual health claim histories. In its simplest form, the Clinical Risk Groups System works by reviewing an individual's history of medical claims, identifying pertinent clinical information, and assigning a single clinically based Risk Adjustment Category or RAC. RACs reflect the presence or absence of a chronic illness or illnesses, the relative severity of those illnesses, and where appropriate the individual's age and sex. An individual's predicted utilization is based solely on RAC assignment with each RAC being assigned a single weight.

RAC assignment is a reasonably simple process. It occurs in two stages. The first is data preparation. The second is RAC assignment. The Clinical Risk Groups System requires individually linkable data for all contacts an individual has with a health care provider. For the data analysis which was used in developing the system, exposure issues also played a significant role. For every contact Clinical Risk Groups System needs diagnoses, procedures, dates of service, site of service, and type of provider. The data are then split into procedure and diagnostic categories.

Procedure categories, called, Episode Procedure Categories or EPCs, collapse ICD-9-CM, CPT-4, and HCPCS procedure codes into comprehensive categories. Most EPCs, such as diagnostic tests and procedures, evaluation and management codes, etc., are inherently ambiguous and are not used. Others, such as chemotherapy, dialysis, etc. provide considerable insight into health status and future utilization and are retained.

ICD-9-CM diagnosis codes are collapsed into clinically meaningful diagnostic categories called Episode Diagnostic Categories or EDCs. EDCs are grouped by body system or Major Diagnostic Category, MDC. Some body systems are assigned multiple MDCs if some of the EDCs are judged to be significantly different from clinical perspective from other EDCs in the same MDC or to facilitate manipulation of the data. The grouping of EDCs within MDC is hierarchical in order of clinical significance.

EDCs are further classified as acute or chronic. Acute EDCs are categorized as minor or moderate acute. A minor acute is an acute EDC which

has no sequelae and is not indicative of any underlying debility or health status (e.g., a minor fracture or upper respiratory infection). A moderate acute EDC is an acute EDC which has either sequelae or is indicative of an underlying health problem (e.g., an intercranial hemorrhage) which is not identified on any claims. Chronic EDCs are classified as Dominant Chronic, Moderate Chronic, Minor Chronic or Chronic Manifestations. Chronic EDCs are defined as follows:

- Dominant Chronic diseases (EDCs) are serious chronic illnesses which dominate an individual's consumption of health care resources over time and usually result in the progressive deterioration of an individual's health and often times lead to, or significantly contribute to, an individual's debility and/or death. Multiple Sclerosis and Emphysema are examples of dominant chronic EDCs.
- Moderate Chronic diseases (EDCs) are illnesses which, though less severe than Dominant Chronic illnesses, tend to dominate an individual's consumption of health care resources and may also lead to, or significantly contribute to, an individual's debility and/or death. Vertigo and Peripheral Vascular Disease are examples of moderate chronic EDCS.
- Minor Chronic diseases (EDCs) are illnesses which are of an extended, frequently lifelong, duration. These are minor illnesses which occur in otherwise healthy individuals. While they tend to be serious in their advanced stages, they can usually be managed effectively and at relatively low cost throughout an individual's life with minimal effect upon the utilization of health services. Headache and Osteoarthritis are examples of minor chronic EDCs.
- Chronic Manifestations are diagnoses (EDCs) which include an underlying chronic diagnoses in their definition. For example, a diagnosis of Diabetic Neuropathy, indicates the presence of diabetes at an advanced level. Chronic Manifestations will be used to create that chronic diagnosis. In addition, They will be used to severity adjust (level) those diagnoses if they are present.

Chronic EDCs can also be created indirectly. That is, some acute EDCs (e.g., CVA or Intercranial Hemorrhage) can create chronic EDCs (Hemiplegia if this is a sequelae or otherwise History of Cerebrovascular Disease). The acute EDC is retained and is available for severity adjustments later in the process of RAC assignment (e.g., the impact of a second AMI). Some acute EDCs will conditionally create a chronic EDC based on recurrence over a defined period. Some chronic diagnoses can create other chronic

diagnoses as discussed earlier. Some procedures, EPCs, (e.g., organ transplants) can create chronic EDCs (e.g., history of a transplant).

All diagnoses associated with an inpatient admission are retained. Only those outpatient EDCs which are reported on two separate occasions are kept. If an EDCs is associated with a hospital admission, is produced by a procedure, or is an important indicator of health status which may not receive active treatment (e.g., blindness) it will be retained based on only a single occurrence is sufficient.

Assigning RACS

The first task of the Clinical Risk Groups System is to sort through what may be for some individuals literally hundreds of diagnoses and procedures which may make up an individual's claim history and develop a description which best characterizes the individual's health status. The collapsing of data into EDCs and EPCs is the first step in data reduction. The next step is to identify chronic EDCs and select only the clinically relevant ones.

Primary Chronic Disease (PCD)

All chronic (including dominant, moderate chronic, and minor) EDCs have been placed into a hierarchy within their respective MDCs based on their severity and likely impact upon the future consumption of resources. Only one chronic EDC will be selected for each MDC. That EDC will be called the Dominant Active Treatment Chronic Episode Diagnosis Category or PCD. If an individual has more than a single chronic EDC from an MDC the following hierarchy will be used:

- Dominant Chronic EDC which is an inpatient primary diagnosis which has occurred in the last year.
- Dominant Chronic EDC which was treated on an outpatient basis in the last year with the first and last treatment dates being at least ninety days apart.
- Any Dominant Chronic EDC.
- Moderate Chronic EDC which is an inpatient primary diagnosis which has occurred in the last year.
- Moderate Chronic EDC which was treated on an outpatient basis in the last year with the first and last treatment dates being at least ninety days apart.
- Any Moderate Chronic EDC.
- The most significant Minor Chronic EDCs as determined by a hierar-

chy of EDCs within each MDC.

In the event more than one EDC meets the criteria and could be selected the EDCs are arranged hierarchically and the most senior one is selected.

After being identified PCDs are adjusted for severity. For each chronic EDC there is a leveling structure with up to four levels based on severity and likely impact upon future resources will be defined. The leveling structure of chronic EDCs within the same MDC tend to be similar, but not identical, for all chronic EDCs in any given MDC. Variations in leveling structure address the specific clinical characteristics of different EDCs. The four levels will generally adhere to the following structure:

PCD with few if any symptoms.

PCD with minor symptoms.

PCD with moderate symptoms.

PCD with major or extreme symptoms.

The severity adjustment is based on the presence or absence of other EDCs from the same or other MDCs and selected EPCs. In order to avoid the possibility of double counting if an EDC from another MDC is used to level a PCD, that EDC (the one being used to level) can not be used as the PCD for its own MDC. The EDCs and EPCs used in the leveling matrix will themselves be subject to modification or inclusion based on specified rules. For example, an EDC with the first and last treatment dates being at least ninety or one hundred and eighty days apart will often produce a higher level than the same EDC which does not meet these criteria. The theory behind this is that all things being equal an EDC which persists or recurs over an extended period is more clinically significant than one which does not. Similarly, an EDC or EPC which has been noted within the last six months of data may receive a higher level than if it has not been recorded within that time period.

RACs

Every individual is assigned a RAC (Risk Adjustment Category). Unique RACs are assigned for all chronic illnesses, combinations of chronic illnesses, and specified conditions. In addition there are specific RACs for people without chronic conditions. RACs reflect severity adjustments and may also reflect an individual's demographic characteristics.

The RAC number has seven digits. The first digit indicates the individual's health status. There are eight statuses.

Healthy
Moderate Acute
Single Chronic
Multiple Chronic
Three or more Dominant Chronics
Metastatic Malignancies
Catastrophic Illnesses and Conditions

RAC Assignment

Once an individual has his/her PCDs assigned, RAC assignment takes place. RAC assignment is done hierarchically with RACs assigned in reverse order of status.

Catastrophic RACs are assigned first. These RACs can be either procedure or diagnosis based. They include dialysis, HIV Disease, TPN, mechanical ventilation, history of allogenic bone marrow transplant, history of major organ transplant (heart, lung, liver, or pancreas), outpatient gastrostomy, quadraplegia, and persistent vegetative state. Assignment is done hierarchically. Therefore a person on dialysis who has also had a major organ transplant would be assigned to the dialysis group. Each of the catastrophic groups is further refined by severity structure unique to itself.

Metastatic Malignancies are assigned next. Individuals in this status include all those with evidence of metastases, multiple malignancies, or recurring malignancies. They are assigned to groups based on a primary malignancy and severity adjusted based on malignancy related EDCs and the presence of other EDCs.

Individuals with three or more Dominant Chronic PCDs or an explicitly named combination of three PCDs are assigned next. Group assignment is hierarchical with combinations selected in a specific order. For example, if a person has congestive heart failure, emphysema, chronic renal failure, and some other dominant chronic PCDs, the first three PCDs will always be used to form the RAC as that combination comes first in the hierarchy. Cases are then assigned to one of six severity levels based on the underlying severity levels of the PCDs which caused them to be assigned to the group. These assigned severity levels are subject to modification based on the presence of specified EDCs or EPCs.

Pairs of diseases that are dominant or moderate represent the next status. For example, in a pair consisting of a dominant chronic PCD and a

moderate chronic PCD, the level of the dominant chronic PCD receives more weight in assigning severity than the level of the moderate chronic PCD. These assigned severity levels are subject to modification based on the presence of specified EDCs or EPCs. This status also includes groups consisting of two and more than two minor chronic PCDs with severity level 1. These groups are further refined by the sex and age.

Individuals with only a single PCD (regardless of type or severity level) or a single PCD which is not a minor chronic PCD with severity level 1 and one or more minor chronic PCDs with severity level 1 are assigned to the next status, Single Chronic Illnesses. The next statuses are those with a minor chronic, acute illnesses or healthy.

For illustrative purposes, consider the following example of how RAC assignment changes as diagnoses change.

A sixty-eight year old female with a few office visits for minor acute illnesses (e.g., EDC 658, Non-Bacterial Infections—Minor) would be assigned to Healthy Female, Aged 65–79. If she had a single outpatient diagnosis from EDC for Cerebrovascular Disease, a moderate acute EDC, she would still be assigned to the healthy RAC because single outpatient occurrences of a diagnoses are ignored. If she had more than one outpatient diagnosis from EDC for Cerebrovascular Disease or if the diagnosis was a primary diagnosis on a hospital admission and if the last reported instance of that diagnosis was no more than six months prior to the end of the analysis period, she would be assigned to the EDC for Cerebrovascular. If the patient had cerebrovascular disease which resulted in a hemiplegia it is likely that the patients future costs and mortality are higher as compared to a patient with only cerebrovascular disease. There is a separate and distinct EDC for hemiplegia.

If she had also received care for diabetes (EDC 315, a dominant chronic illness) but had no other symptoms, she would be assigned to a doublet consisting of diabetes and cva. If the diabetes became worse and she was treated, with rehabilitation, of an a/k amputation she would be assigned to a higher level of severity for this doublet. If her cardiovascular problems worsened and she had a several diagnoses of angina (EDC 125), the cardiovascular PCD would be angina. Her RAC assignment would become a triplet, Diabetes, CVA and Angina. It is entirely possible that this patient might receive rehabilitation services for any and/or all three of the diseases within this triplet. If there were additional diagnoses, e.g., asthma (EDC 78) or some psychiatric problems (e.g., Depression, EDC 929), these would

either be ignored or, if clinically appropriate, raise the patient to a more severe level.

If the patient's renal failure worsened and she required dialysis on at least two separate occasions at least thirty days apart, she would be considered a catastrophic case or more correctly a person at high risk for future high costs. She would be assigned to RAC 7010300, Dialysis with Diabetes, Level 3. The assignment of level 3 would reflect the presence of CHF.

The above example provides a graphic case of why it is important to take into account all key clinical illnesses that a patient may have. While small in number these consume a large percentage of dollars within the health care system. Recently published Medicare data indicates that 9.8 percent of Medicare enrollees consume 68.4 percent of expenditures; conversely 73.6 percent of Medicare enrollees consume only 8.6 percent of Medicare expenditures.

Data

The Clinical Risk Groups System has been under development for several years. It was completed in early 1999. Initial development is based on the Medicare Standard Analytical File, SAF, a five percent sample of Medicare patients with linkable data and encrypted individual identifiers to protect the privacy of beneficiaries. Two years of clinical data derived from fee for service Medicare claims (1991 and 1992) are being used to project total charges in the third year (1993). To be included in the analysis an individual had to have two full years of Medicare fee for service coverage and a third full year of coverage unless the reason for a shortened third year was death. Individuals with HMO coverage or whose claims indicated the possible presence of another primary payer were excluded from the analysis.

For the purposes of the analysis the data have been adjusted. First, total annual charges have been capped (cases kept but not allowed to exceed a predetermined level) at $100,000 for Single Chronic illness, $150,000 for Multiple Chronic Illnesses, $250,000 for Three or More Dominant Chronic illness, $250,000 for Metastatic Malignancies, and $500,000 for Catastrophic Cases. Second, the charges of individuals who died in 1993 have been annualized. The model, albeit with the aforementioned capping of catastrophic costs and annualization of the charges of people who died, has an R^2 that explained 18.3 percent of the validation in total charges.

For illustrative purposes, examples of conditions for each status relevant for rehabilitation are provided in Figures 1–6. The examples follow the patient example provided above. In the same table, one can see that mortality has a fairly unambiguous positive relationship with severity level. The differences between different levels combinations of different diseases is particularly interesting. The six levels of the combination reflects the collapsing of sixteen possible cells. The lower levels of the combination roughly correspond to the lower levels of the single groups and the higher levels to the higher levels of the single groups.

Directions for Future Research

The research that developed this risk adjustment system is pursuing several different lines of research:

- Retrospective analyses are being performed. The data presented in this paper only pertain to future consumption of resources or mortality. It is important to validate the classification system for retrospective analyses
- Mini-episodes of illness need to be identified. Thus, while year long

# of Patients	% of Death Dep Yr.	Avg Paid	RAC Description
279	6.81	7,909	DC Single Chronic Acquired Hemiplegia
472	8.90	9,572	DC Single Chronic Acquired Hemiplegia
113	12.39	11,205	DC Single Chronic Acquired Hemiplegia
133	9.77	12,695	DC Single Chronic Acquired Hemiplegia

Figure 1.

episodes of care for hemiplegia are important, providers will also find

# of Patients	% of Death Dep Yr.	Avg Paid	PRAC Description
1,041	5.38	6,151	DC Single Chronic-history of CVA
1,271	9.91	7,792	DC Single Chronic history of CVA
364	9.89	8,402	DC Single Chronic history of CVA
451	19.96	11,515	DC Single Chronic history of CVA

Figure 2.

useful a classification system which includes mini-episodes such as a

# of Patients	% of Dealth Dep Yr	Avg Paid	PRAC Description
15,945	2.90	5,215	DC Single Chronic Diabetes
5,066	2.70	6,546	DC Single Chronic Diabetes
4,060	3.84	8,242	DC Single Chronic Diabetes
2,778	5.26	10,679	

Figure 3.

# of Patients	% of Death Dep Yr.	Avg Paid	PRAC Description
317	8.20	8,859	CVA & DM
746	8.18	11,541	CVA & DM
694	9.51	14,082	CVA & DM
736	13.04	15,095	CVA & DM
697	15.64	16,860	CVA & DM
1,123	21.10	26,257	CVA & DM

Figure 4.

# of Patients	% of Death Dep Yr	Avg Paid	PRAC Description
164	12.20	11,628	CHF-DM-CVA
450	17.78	17,157	CHF-DM-CVA
439	18.68	21,300	CHF-DM-CVA
521	24.18	26,252	CHF-DM-CVA
507	24.26	31,398	CHF-DM-CVA
909	28.60	39,710	CHF-DM-CVA

Figure 5.

# of Patients	% of Death Dep Yr	Avg Paid	PRAC Descritpion
461	5.64	13,389	Acquired Quad
255	6.67	22,997	Acquired Quad
154	13.64	35,914	Acquired Quad
398	16.33	47,509	Acquired Quad

Figure 6.

hospitalization for a CVA together with 90 days post discharge. This would include the vast majority of resources consumed for rehabilitation.

New data elements need to be included in the risk classification system. The risk adjustment system described in this paper uses only information present on the claims form. The data elements most easily incorporated in the next five years include the name of the pharmaceutical and outpatient laboratory results. The reason for the inclusion of these two variables is that pharmaceuticals and laboratory results are often capitated

out by the insurance or MCO. The pharmacy benefit management firm and laboratory company typically have this data on file. The reader is likely interested in knowing when functional health status will be readily available. In part the answer is political. That is, if and when this information is routinely collected by federal and/or state government for coverage of inpatient rehabilitation, home care and/or nursing home facilities, it will be linkable to claims data and other data elements. Otherwise, it will be up to providers to insist that functional health status be routinely collected on patients.

Conclusion

This preliminary analysis shows that severity levels vary considerably with illnesses and that disease interaction is disease or disease combination specific. While further analysis is in order, it is reasonable to state that if prospective rate adjustment methodologies are to improve their predictive power they need to be sensitive to disease specific relationships. Methodologies, such as the Clinical Risk Groups System, which take this into account will do as well, and probably better than alternative methodologies, which fail to address disease specific relationships. Taking severity and disease specific relationships, as described in this paper, is particularly important in the understanding of episodes of illnesses important for rehabilitation medicine.

References

Dunn, D. L., Rosenblatt, A., Taira, D. A., Latimer, E., Bertko, J., Stoiber, T., Braun, P., and Busch, S. (1995, December 21). A Comparative Analysis of Methods of Health Risk Assessment: Final Report. American Society of Actuaries, Chicago, Illinois.

Ellis, R. P., Pope, G. C. L. I., Ayanian, J. Z., Bates, D. W., Burstin, H., Dayhoff, D., Rensko, A. M., Dawes, T., and Ash, A. S. (1996, April 21). Diagnostic Cost Group (DCG) and Hierarchical Coexisting Conditions (HCC) Models for Medicare Risk Adjustment. Health Economic Research, Inc., Waltham Massachusetts.

Government Accounting Office (1995, November 8). Medicare Managed Care: Growing Enrollment Adds Urgency to Fixing HMO Payment Problem. GAO/HEHS-96-21.

John Hopkins University (1998). The Johns Hopkins University ACG Case-Mix Adjustment System, Version 4.1. Johns Hopkins University, Baltimore, Maryland.

Newhouse, J. P., Manning, W. G., Keeler, E. B., and Sloss, E. M. (1989). Adjusting Capitation Rates Using Objective Health Measures and Prior Utilization. *Health Care Financing Review*, *10*(3), 41-54.