American Journal of Epidemiology

Vol. 150, No. 11
Printed in U.S.A.

Determination of Risk Factor Associations with Questionnaire Outcomes: A Methods Case Study

Karen Bandeen-Roche,[1] Guan-Hua Huang,[1] Beatriz Munoz,[2] and Gary S. Rubin[3]

Increasingly in biomedical studies, health status is inferred through a series of questionnaire item responses. Challenges for analyzing associations between such responses and risk factors include multiplicity—many indicators must be combined to derive summary statements about health status, and measurement error—persons' self-report fluctuates due to causes other than substantive health changes. In order to deal with these challenges, the authors propose a strategy which comprises three methods: 1) score the item responses, then regress the score on predictors; 2) regress each item response on predictors, accounting for within-person associations; and 3) summarize and analyze the item responses jointly, using a latent variable model. The authors develop modeling and diagnostic procedures for method 3. They then show how the three-method analytic strategy can be used to solve the problem of determining which aspects of vision are associated with self-reported functioning in activities that require seeing at a distance. They demonstrate that methods 2 and 3 illuminate basic findings from method 1 by adding specificity, describing patterns as well as severities of health impairments, and identifying isolated items that relate to risk factors differentially than others. They conclude that the three-method strategy specifies how risk factors determine questionnaire-based health outcomes substantially better than any of the methods in isolation. *Am J Epidemiol* 1999;150:1165–78.

aging; discrete data; goodness-of-fit; latent class; latent variable; multivariate regression; vision

In biomedical studies, scientists increasingly infer health status through questionnaire item responses. For example, gerontologists often determine older persons' functioning through answers to questions about ability to perform routine tasks (1–3), and mental health (4), hearing (5), and pain (6, 7) specialists also commonly assess outcomes using questionnaires. Self-report is recognized as a useful health measurement both because questionnaires are often easier to administer than clinical tests and because perceived status is relevant in the design and evaluation of treatments (8). However, analyzing questionnaire responses poses substantial challenges, including: 1) multiplicity—many indicators must be combined to derive summary statements about health; and 2) measurement error—persons' self-report fluctuates not only because of substantive health changes but also because of incidental health changes and rating imprecision. This paper proposes a strategy for handling these challenges, focusing on the aim of determining risk factor relations with health outcomes.

A common approach to analyzing questionnaire outcomes for associations with risk factors is to combine item responses into scores, justify that the scores have reasonable psychometric properties (e.g., reference 9), and analyze the scores. Such "scoring analysis" is simple, reduces multiplicity, and averages out incidental fluctuations in item responses. The danger in using it exclusively is that scoring may combine indicators of distinct health processes and thus mask associations with risk factors. This paper argues the advantages of augmenting scoring with two other approaches: 1) analyzing how individual questionnaire item responses are related to explanatory variables, accounting for within-person associations ("item analysis"); and 2) jointly summarizing and analyzing item responses using latent variables. For studying health associations with risk factors, we have found parallel scoring, item, and latent variable analyses to be enlightening over any in isolation.

Received for publication April 27, 1998, and accepted for publication March 10, 1999.

Abbreviations: ADVS, Activities of Daily Vision Scale; CI, confidence interval; GHQ, General Health Questionnaire; IQR, inter-quartile range; LCR, latent class regression; MMSE, Mini-Mental State Examination; OR, odds ratio; SEE, Salisbury Eye Evaluation.

[1] Department of Biostatistics, School of Hygiene and Public Health, The Johns Hopkins University, Baltimore, MD.
[2] Dana Center for Preventive Ophthalmology, Wilmer Eye Institute, Johns Hopkins Medical Institutions, Baltimore, MD.
[3] Lions Low Vision Center, Johns Hopkins Medical Institutions, Baltimore, MD.
Reprint requests to Dr. Karen Bandeen-Roche, Department of Biostatistics, School of Hygiene and Public Health, The Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205.

This paper aims, first, to convey the main ideas underlying a latent variable method we have developed for analyzing questionnaire-based outcomes (10). The approach seeks to summarize the substantively meaningful aspects of item covariation and determines risk factor associations through regression of those aspects on covariates. It also features formal and graphical methods to determine 1) whether modeling succeeds in describing observed data; and, if not, 2) in what ways modeling fails to synthesize responses as intended. Second, this paper aims to demonstrate how scientific insights may be gained from concurrent as opposed to only one of scoring, item, and latent variable analysis, using current visual functioning data. First, we outline the visual application. Then, we detail the statistical methods to be used. Finally, we illustrate our analytic strategies with scoring, item, and latent variable analyses of the association between self-reported visual functioning and several measured aspects of vision. We conclude by highlighting benefits and potential enhancements of our analytic approach.

MATERIALS AND METHODS

Data source: The Salisbury Eye Evaluation Project

For this report, we analyzed data from the Salisbury Eye Evaluation (SEE) Project. The SEE sample has been detailed elsewhere (11). In brief, an age- and race-stratified random sample of Salisbury, Maryland residents was drawn from national Medicare eligibility lists. Eligibility criteria for inclusion in SEE were: age between 65 and 84 years as of the sampling date, community residency with better-than-invalid health, and Mini-Mental State Examination (MMSE) (12) score of ≥18. The 2,520 persons who agreed to fully participate were administered an extensive in-home interview followed within a few weeks by a 5-hour clinic examination.

Table 1 summarizes basic characteristics of persons who participated in both the home interview and clinic examination. Such persons comprised 65 percent of those eligible; differences between these and study refusals have been summarized elsewhere (13).

TABLE 1. Demographic characteristics of Salisbury Eye Evaluation (SEE) population (n = 2,520), Salisbury, Maryland, September 1993 to September 1995

Characteristic	%
Age (years)	
65–69	31.0
70–74	33.1
75–79	22.0
≥80	13.9
Sex	
Male	42.1
Female	57.9
Race	
White	73.6
African American	26.4
Education (years)	
<7	8.2
7–11	43.3
12	20.4
>12	28.1
MMSE* score	
<24	16.2
24–29	65.4
30	18.4
GHQ* depression score	
0	90.5
1–2	6.8
≥3	2.7
No. of comorbid diseases†	
0	9.6
1	21.8
2	26.5
3	20.5
4–5	17.5
≥6	4.1

* MMSE, Mini-Mental State Examination; GHQ, General Health Questionnaire.

† Out of 15 assessed by self-report, "Has a doctor ever told you that you have...": arthritis, hip fracture, back problems, myocardial infarction, angina, congestive heart failure, intermittent claudication, high blood pressure, diabetes, emphysema, asthma since age 50, stroke, Parkinson's disease, cancer in the last 5 years, vertigo.

Illustrative analytic question: What aspects of vision determine "far vision" functioning?

One SEE Project aim was to determine how different aspects of vision affect older persons' functioning. To illustrate our methodology, we narrow that aim to determining how different aspects of vision affect functioning in activities that require seeing at a distance. In the SEE Project, the Activities of Daily Vision Scale (ADVS) (14, 15) was administered to elicit subjects' ratings of their difficulty doing various vision-related activities. Each participant was asked whether he had done each activity within the last 3 months. If so, the level of difficulty of doing the activity was recorded (2 = extremely difficult; 3 = moderately difficult; 4 = a little difficult; 5 = not at all difficult), where any difficulty must have been specified as affected by vision. A participant who had not done the activity in the last 3 months because of vision was assigned a most severe level of difficulty (1 = unable to do due to vision); one who had not done the activity recently for reasons other than vision was not rated on the activity. For this paper, we analyzed responses about the five activities comprising the ADVS "far vision" subscale (table 2): reading street signs, at night

TABLE 2. Frequency distributions of Activities of Daily Vision Scale (ADVS) Far Vision items and scores: Salisbury Eye Evaluation (SEE), Salisbury, Maryland, September 1993 to September 1995

Activity ("How difficult is...")	No.† (total = 2,520)	Degree of difficulty (%) *				
		High				Low
		1	2	3	4	5
Reading street signs at night	1,927	6.7	2.0	7.5	26.4	57.4
Reading street signs in daylight	2,366	2.2	0.6	3.7	11.8	81.7
Walking down steps in daylight	2,372	1.1	0.4	2.0	8.0	88.6
Walking down steps in dim light	2,157	2.3	0.6	3.2	12.5	81.4
Watching television	2,515	0.6	0.7	1.5	7.7	89.6
Far vision score‡	2,519	0.3	1.4	3.6	16.4	78.3

* 1, do not do the activity because of vision problems; 2, extremely difficult; 3, moderately difficult; 4, a little difficult; 5, not difficult at all.

† In nearly all cases, persons not counted here reported not having done the activity in the previous 3 months for reasons other than vision problems. The highest number of cases of having no response at all recorded was 11, among all 21 questions comprising the SEE project administration of the ADVS.

‡ Number for score denotes number who could be scored by published ADVS algorithm. Scores are given as the average of available item responses, rounded to the nearest integer.

and during the day; walking down steps, in daylight and in dim light; and watching TV.

We used five "psychophysical" measures to quantify vision; these were obtained during the SEE clinic exam (16). In brief, "visual acuity" measures ability to resolve fine images, "contrast sensitivity" measures ability to distinguish shading, "glare sensitivity" measures ability to cope with glare in distinguishing shading, "stereoacuity" measures depth perception, and "visual field" testing measures range of peripheral vision as well as the presence of blind spots. Paraphrasing common wisdom as reflected in standard ophthalmic practice, the null hypothesis is that after controlling for visual acuity, no other measure influences far vision functioning.

We identified several characteristics as potentially obscuring the relation between vision and far vision functioning (table 1): age at clinic examination (years), sex (indicator of being female), race (indicator of being African-American), education (years), cognition (MMSE score), psychologic status (General Health Questionnaire (GHQ) depression score (17)), and disease burden (number of reported comorbid diseases). Each model we report included all of these as covariates.

Analytic methods for questionnaire response analysis: overview

Scoring analysis is arguably the most commonly used strategy for determining how risk factors affect a battery of self-reported health responses. It often provides a simple and effective way to summarize and report research findings. To minimize the risk of masking important associations by ineffective summary of item responses, however, we advocate that scoring analysis findings be compared and contrasted with findings derived from item and latent variable analyses. The first of these alternatives excels at accurately describing observed data; thus, it provides an excellent reference for checking statistical assumptions on which the other analytic strategies rely. However, it may yield unnecessarily imprecise inferences by retaining unreliability in individual item responses. Latent variable analyses explicitly account for item response unreliability and thus may estimate associations more accurately or precisely than the other approaches. They also give well-summarized inferences, can incorporate theory underlying item choices, and evaluate two assumptions that are implicit to the scoring approach: that subscale items measure a single construct (unidimensionality), and that risk factors are not very differentially associated with responses on different items (nondifferential measurement).

Well-practiced statistical methods are available to analyze scores on virtually any measurement scale as well as to describe relations between predictors and individual item responses, accounting for within-person associations (18–25). In contrast, latent variable regression models appropriate to questionnaire data are of much more recent origin. In the remainder of this section, we elucidate our proposed latent variable approach to determining associations between risk factors and questionnaire responses.

A method proposal: latent class regression (LCR)

Broadly, latent variable methods are designed for settings where the object of analysis is defined in theory but is hard to measure. Questionnaires in general and the SEE illustration in particular provide good examples of this; there is arguably no one adequate measure of "ability" in activities that require seeing at

a distance, so instead investigators pose questions that get at different aspects of subjects' "far vision" functioning. Latent variable modeling treats the unobserved, theoretical object (perceived "ability") as the outcome to be analyzed for relations with risk factors and the measured responses as quantities that imperfectly determine the object of interest. Inferences that involve the idealized, latent variable stand to be more accurate and precise than inferences on variables that measure the analytic object with some error. The corresponding difficulty is that one cannot hope to determine the latent variable without making assumptions about how the observed variables relate to the latent variables. A workable and good latent variable analysis must make assumptions that simultaneously serve to specify the idealized response, are scientifically interpretable and plausible, and can be checked for consistency with observed data.

Our proposed latent variable approach is designed for discrete data, such as Likert item responses. Its idea is easily applicable to polytomous or ordinal item responses; to keep the exposition simple, we will explicitly describe modeling for binary responses. The approach hypothesizes that a population of interest can be grouped into some J subpopulations, or *classes*, each of which has homogeneous outcome status (figure 1). Casting the discussion in terms of the SEE application: the presumption is that a person's functioning status can be accurately and precisely described by a categorical latent variable L, which identifies the underlying subpopulation to which the given person belongs. At first glance, it may seem overly restrictive to describe functioning categorically; this is actually quite reasonable upon reflecting that the measured outcomes are patterns of binary responses to questions about functioning. Because such patterns may involve endorsing different but similar numbers of items, the latent variable is polytomous—not necessarily ordinal. Modeling that distinguishes patterns and not only numbers of functional deficits is an appealing feature of our latent variable approach.

Our model describes risk factor associations by regressing underlying functioning, L, on predictors $x = (x_1,\ldots,x_p)$. To elucidate how, imagine for a moment that L could be observed. Then, an appropriate analysis would be a polytomous (say, logistic (26)) regression of L on x. This can be written as

$$\log\left(\frac{P_j(x)}{P_J(x)}\right) = \beta_{0j} + \beta_{1j}x_1 + \ldots + \beta_{pj}x_p, \qquad (1)$$

$$j = 1, \ldots, J-1,$$

where $P_j(x)$ is the probability that a person with covariates x is in the jth functioning class, e.g., that $L = j$. Here, β_{qj} is a log odds ratio (OR)—the (log) factor by which the odds of membership in the jth versus the Jth (reference) class compares across persons who differ by one unit on x_q holding other covariates constant. The variable L is not observable for any person; per equation 1, our latent variable model describes how the prevalences P_j of each class vary with predictor variables.

Having described how the idealized object of analysis relates to covariates, our model must specify how persons' item responses relate to their underlying status L. Suppose that Y_m stands for a person's indication of difficulty (or not) on the mth far vision activity, giving functioning responses $Y_1,\ldots,Y_5$. Our model must specify the frequencies with which members of the jth class endorse difficulty on each

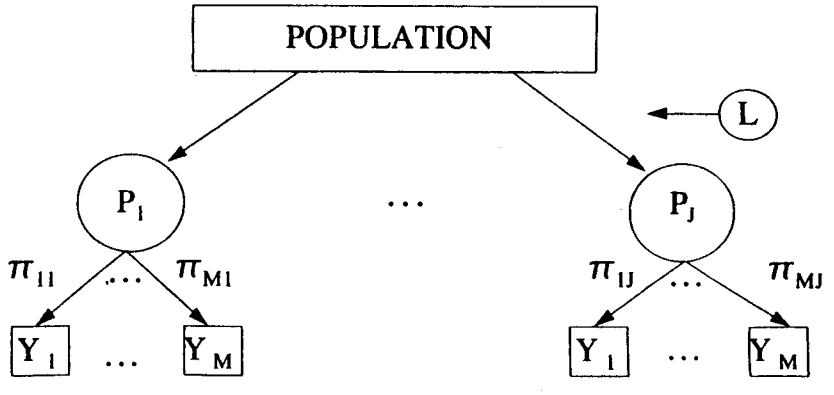

FIGURE 1. Basic model for proposed latent variable approach. The model assumes a study population to be comprised of several (J) subpopulations, each of whose members report similarly about their functioning. The defining quantities of the model are: L, the latent functioning variable that identifies the subpopulation to which a study participant belongs; P_j, the prevalence of the jth subpopulation, $j = 1,\ldots,J$; and π_{mj}, the probability that a member of the jth subpopulation reports difficulty on the mth item, $m = 1,\ldots,M$, $j = 1,\ldots,J$.

activity, $j = 1,\ldots,J$; it must also distinguish which covariate effects on the Y's are mediated through L versus not and which response associations within persons are due to the common effects of L versus not. This is too tall an order; rather, these things are achievable only if statistical conditions are imposed. In formulating conditions, our conceptual framework is that the latent variables should summarize all that is of substantive interest, and that any excess variability in the Y's should represent measurement imprecision. This mandates two conditions: that after stratifying on underlying functioning, there be 1) no response associations within persons (*conditional independence*), and 2) no direct covariate effects on reported functioning (*nondifferential measurement*). These may or may not be reasonable for any given dataset; therefore, we first presume that they are reasonable, then diagnose whether and how they may be violated.

Specifically, then, the proposed regression of self-report on covariates is:

$$P\{Y_1, \ldots, Y_5 = y_1, \ldots, y_5|x\} = \quad (2)$$

$$\sum_{j=1}^{J} P_j(x) \prod_{m=1}^{5} \pi_{mj}^{y_m}(1 - \pi_{mj})^{1-y_m},$$

where the reporting probabilities $\pi_{mj} = P\{Y_m = y_m|L = j\}$ determine the frequencies with which members of class j endorse difficulty on activities $m = 1,\ldots,5$. This formulation was originally proposed by Dayton and Macready (27). $P_j(x)$ defines a polytomous regression of L on the covariates through equation 1. The piece $\prod_{m=1}^{5} \pi_{mj}^{y_m}(1 - \pi_{mj})^{1-y_m}$ is exactly as in the standard "latent class" model (28, 29). Conditional independence mandates the $\prod_{m=1}^{5}$ form in the previous sentence, and nondifferential measurement mandates that the π_{mj} *not* vary with the covariates.

We recently laid out technical details of estimation, inference, and model checking for equation 2 (10). In brief, we estimate parameters by maximum likelihood. The number of classes must either be decided using substantive judgment or chosen empirically in a first analytic stage; if the latter, we make the choice—say, J^*—using diagnosis and likelihood ratio testing for goodness of fit to a version of equation 2 that ignores covariates (e.g., with all β's in equation 1 set equal to 0). Then, we finalize reporting probabilities and estimate regression coefficients fixing $J = J^*$, iterating between equations 1 and 2. We have developed SAS and S-PLUS macros for fitting LCR models (these are available on request from the first author).

Model checking

In latent variable analysis, one must check whether 1) models fit the data and 2) the statistical conditions appear reasonable. We accomplish the former by comparing observed versus modeled item response prevalences, for each item. We accomplish the latter using the diagnostic method of Bandeen-Roche et al. (10). For clarity, we illustrate these following reporting of results.

RESULTS: ANALYSIS OF VISION ASSOCIATIONS WITH REPORTED FUNCTIONING

Preliminary: results reporting

To report vision associations with functioning, we contrasted better versus worse vision in units: 0.3 logMAR for visual acuity, 6 letters for contrast and glare sensitivity, 0.3 log arc seconds for stereoacuity, and $\sqrt{2}$ for number of central visual field points missed. These numbers were chosen for clinical meaning. In fact, the visual acuity, contrast sensitivity, and glare sensitivity choices spanned twice their respective variables' inter-quartile ranges (IQRs), whereas the stereoacuity and visual field choices spanned approximately 60 percent of their respective variables' IQRs.

Approach 1: Scoring analysis

As a first analysis, we computed ADVS far vision scores using the published method of averaging persons' item ratings (14), then regressed the scores on vision and other variables. In this, we began by using linear regression; this proved inadequate because the far vision score distribution was severely skewed (table 2). Next we trichotomized and then analyzed the scores using ordinal logistic regression (19); however, the model's proportional odds assumption was violated for most covariates. Therefore, we present polytomous logistic regression analyses (20) of the trichotomized scores. Per equation 1, these describe associations between predictors and functioning as odds ratios, separately comparing high versus low and medium versus low functioning.

In fitted regressions, we found independent and strong associations between the five vision measures and far vision functioning after adjusting for confounding variables (table 3). A 0.3-logMAR-better visual acuity was associated with more than a twofold increase in the odds of being best- versus worst-functioning; the increase was 1.7-fold for six-letter-better contrast sensitivity; and lesser associations were observed with the other vision measures. Male gender, fewer comorbid diseases, and lower GHQ depression scores were also significantly associated with report-

TABLE 3. Scoring analysis—multiple regression of Activities of Daily Vision Score (ADVS) far vision scores on vision variables: Salisbury Eye Evaluation (SEE), Salisbury, Maryland, September 1993 to September 1995

Vision variable†	Comparison‡	OR§	95% CI§
Visual acuity (0.3 logMAR)	Best vs. worst	**2.74**	2.04, 3.68
	Mid vs. worst	**1.72**	1.29, 2.28
Contrast sensitivity (6 letters)	Best vs. worst	**1.69**	1.23, 2.32
	Mid vs. worst	**1.46**	1.06, 2.01
Glare sensitivity (6 letters)	Best vs. worst	1.39	0.97, 2.00
	Mid vs. worst	1.07	0.73, 1.56
Stereoacuity (0.3 log arcsec)	Best vs. worst	**1.25**	1.13, 1.39
	Mid vs. worst	**1.23**	1.10, 1.37
Visual field (1.4 √ letters)	Best vs. worst	1.14	0.98, 1.33
	Mid vs. worst	1.03	0.88, 1.21

* Polytomous logistic regression, adjusted for: age, sex, race, education, MMSE§, GHQ§ depression score, and number of comorbid diseases.

† Parentheses identify unit of comparison for which OR was calculated.

‡ Far vision score categories: best, 94–100; mid, 72–93.99; worst, <72.

§ OR, odds ratio; CI, confidence interval; MMSE, Mini-Mental State Examination; GHQ, General Health Questionnaire.

ing better functioning, after adjustment for vision measures (data not shown). Summarizing, we derived at least two important findings from scoring analysis:

1) *Multiple aspects of vision were associated with self-reported far vision functioning.* In itself, this finding is novel and important.

2) *ADVS reporting was associated with characteristics other than vision*, indicating that gender and psychosocial characteristics must be considered in evaluating treatments or making policy decisions based on ADVS.

Approach 2: Item analysis

We next conducted an analysis that described item-specific associations with predictors using ordinal logistic regression and accounted for response associations within persons (30). If R_{im} represents person i's rating on far vision activity m and x_{iq} are predictor variables, $q = 1,\ldots,p$, the ordinal item response model is

$$\log[\text{odds}(R_{im} > k|x_i)] = \beta_{0mk} + \beta_{lm}x_{il} + \beta_{pm}x_{ip} \quad (3)$$

for ratings $k = 1,\ldots,4$ and items $m = 1,\ldots,5$. Here, the covariate coefficients are independent of k and thus describe log ratios comparing odds of higher versus lower functioning, independently of how "higher" and "lower" are defined. They do depend on m and thus allow stronger covariate associations with some items than with others.

In practice, fitting equations 3 can involve many parameters and be both statistically and computationally complex. Hence, a key element of model building is to determine which covariates q can reasonably be modeled as having equal strengths of associations across items ($\beta_{mq} = \beta_{m'q}$ for all $m' \neq m'$). Details of the model building procedure we applied may be found in Huang, Bandeen-Roche, and Rubin (Johns Hopkins University, unpublished manuscript).

Far vision item analysis was largely consistent with scoring analysis (figure 2). However, whereas most vision covariates were similarly associated with different item responses, visual acuity was much more strongly associated with difficulty reading signs at night and during the day and watching TV than with descending steps in either type of light. Male gender and more comorbid diseases were also preferentially associated with difficulty watching TV. These findings indicate that the far vision scale has either of two shortcomings for risk factor analyses: It could be eliciting multiple functioning dimensions, or some items could be subject to differential reporting. In summary, item analysis largely confirmed the findings of scoring analysis but raised concerns about the internal validity of the far vision scale.

Approach 3: Latent class regression

Finally, we analyzed associations between far vision functioning and psychophysical vision measures using LCR. For consistency with model exposition in the Methods section, we dichotomized difficulty ratings at "no" versus "any" difficulty. A complication was that many SEE participants reported not doing various far vision activities at all and therefore were scored "missing" on those activity ratings, whereas equation 2 assumes complete item reporting. For this paper, we applied LCR to those participants who rated each far vision item for difficulty and also had no missing covariates (n = 1,643); these represented nearly 95 percent of participants who reported that they drive at night but were significantly younger, more predominantly male and white, better educated, and healthier than the complementary SEE subsample.

In initial fitting, five classes appeared to adequately describe the observed difficulty reporting patterns (likelihood ratio chi-square for fit of equation 2, ignoring covariates: 6.78 on 3 degrees of freedom, with $\pi_{24} = 0$). Table 4 displays the reporting probabilities

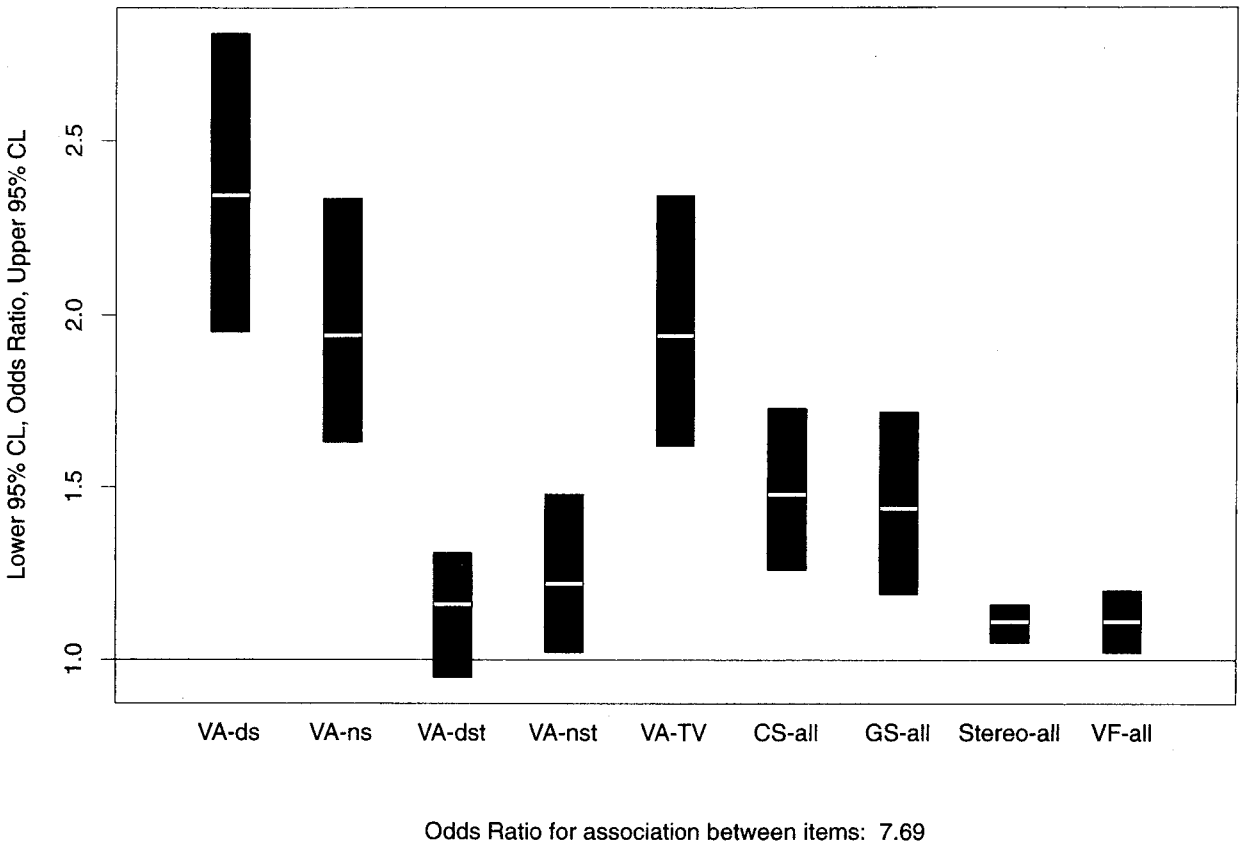

FIGURE 2. Item analysis. Regression of far vision functioning responses on vision variables: Salisbury Eye Evaluation (SEE), Salisbury, Maryland, September 1993 to September 1995. From bottom to midline to top, each "box" displays lower 95% confidence limit, estimate, and upper 95% confidence limit for (global) odds ratio describing the association between a vision measure and far vision functioning, controlling for other vision measures and age, sex, race, education, MMSE score, GHQ depression score, and number of comorbid diseases. The five leftmost boxes are for visual acuity associations (per 0.3 logMAR units) with (left to right): reading signs during the day, reading signs at night, walking down steps in daylight, walking down steps in dim light, and watching TV. The four rightmost boxes are for (left to right) contrast sensitivity (per 6 letters), glare sensitivity (6 letters), stereoacuity (0.3 log arcsec), and visual field (1.4 root-points missed) associations. For each of these latter vision variables, associations were not found to differ between items; a single odds ratio for association combining across items is displayed.

TABLE 4. Latent class regression model for Activities of Daily Vision Score (ADVS) far vision responses, reporting probabilities and class prevalences, Salisbury Eye Evaluation (SEE), Salisbury, Maryland, September 1993 to September 1995

Task	Reporting probabilities (π)*				
	Class 1 (none)	Class 2 (Nt-sign†)	Class 3 (signs)	Class 4 (steps)	Class 5 (severe)
Signs—night	<0.001	1.00	0.951	0.710	1.00
Signs—day	0.006	0.097	0.949	0.00	1.00
Steps—day	0.002	0.006	0.263	0.640	1.00
Steps—dim	0.019	0.085	0.557	0.914	1.00
Watch television	0.010	0.045	0.273	0.168	0.786
Prevalences (P)	0.569	0.235	0.109	0.061	0.026

* Following exploratory analysis, three π's (π_{12}, π_{24}, π_{45}) were set equal to 1.00 or 0.00. This was necessary to uniquely identify the model. The model fit converged to a solution with an additional three π's (π_{15},π_{25},π_{35}) equal to 1.00.

† Nt-sign, night signs.

obtained by fitting equations 1 and 2 concurrently; as an interpretive example, the bottom right value $\hat{\pi}_{55} = 0.786$ indicates that an estimated 78.6 percent of persons in functioning class 5 report any difficulty watching TV. The reporting probabilities describe the class compositions: few class 1 members report any difficulties; class 2 members report difficulty reading signs at night, but few report other difficulties; class 3 members report difficulty reading signs preferentially to other difficulties whereas class 4 members preferentially report difficulty descending steps and reading signs at night; and class 5 members report difficulty in most tasks. Notice that classes 3 and 4 were distinguished by different patterns rather than numbers of difficulties.

Our LCR analysis agreed with scoring and item analyses regarding confounding variable associations—notably indicating a strong relation between female gender and *not* being in the best functioning subpopulation—and in finding multiple vision measures to be independently associated with far vision functioning after adjusting for potentially confounding variables (table 5). More, it detected meaningful patterns of associations between the vision variables and underlying functioning. Poorer visual acuity was very preferentially associated with patterns that involved difficulty reading signs during the day versus those that did not, whereas poorer contrast sensitivity was preferentially associated with patterns that involved difficulty descending steps versus those that did not, and poorer glare sensitivity was associated very generally with not being in the most able functioning class.

Figure 3 illustrates how the LCR model describes functioning. There, worsening visual acuity appears associated with 1) a steep drop in the probability of being best functioning (widest dash), and 2) increased prevalences of being poorer functioning primarily for "signs" (smallest dash) and "severe" (solid line) disability as opposed to "night signs" (medium dash) or "steps" (dotted) disability. Also, the plots highlight remarkable gender sensitivity in reporting, such that women with 20/20 acuity (logMAR = 0.0) report functioning very similar to men with 20/40 acuity (logMAR = 0.3).

Summarizing, the latent variable analysis added several findings to those already obtained:

1) *Different vision impairments were associated in clinically meaningful ways with different patterns of far vision activity problems.* This was hinted at by item analysis but made specific by LCR, where two distinct moderately disabled functioning patterns were identified, and pattern-specific associations were described. This has immediate implications for geriatric ophthalmic practice—e.g., environmental falls might be preventable by counseling persons with contrast sensitivity loss.

TABLE 5. Multiple latent class regression of Activities of Daily Vision Score (ADVS) far vision responses on vision variables*: Salisbury Eye Evaluation (SEE), Salisbury, Maryland, September 1993 to September 1995*

Variable†	Comparison	OR‡	95% CI‡
Visual acuity (0.3 logMAR)	Severe vs. none	2.03	0.62, 6.68
	Steps vs. none	1.29	0.57, 2.93
	Signs vs. none	**3.40**	2.17, 5.31
	Nt-sign‡ vs. none	1.47	0.97, 2.22
Contrast sensitivity (6 letters)	Severe vs. none	2.28	0.80, 6.52
	Steps vs. none	**1.86**	1.07, 3.23
	Signs vs. none	1.33	0.86, 2.07
	Nt-sign‡ vs. none	1.36	0.98, 1.88
Glare sensitivity (6 letters)	Severe vs. none	2.45	0.46, 13.1
	Steps vs. none	**1.89**	1.02, 3.51
	Signs vs. none	**2.09**	1.28, 3.42
	Nt-sign‡ vs. none	1.34	0.94, 1.90
Stereoacuity (0.3 log arcsec)	Severe vs. none	1.64	0.87, 3.10
	Steps vs. none	1.06	0.80, 1.41
	Signs vs. none	1.08	0.89, 1.30
	Nt-sign‡ vs. none	1.01	0.92, 1.12
Visual field (1.4 $\sqrt{}$ letters)	Severe vs. none	2.20	0.67, 7.30
	Steps vs. none	**1.30**	1.01, 1.68
	Signs vs. none	1.14	0.92, 1.43
	Nt-sign‡ vs. none	1.01	0.87, 1.17

* Adjusted for age, sex, race, education, MMSE‡, GHQ‡ depression score, and number of comorbid diseases.

† Parentheses identify unit of comparison for which OR was calculated, scaled so that higher score = worse.

‡ OR, odds ratio; CI, confidence interval; MMSE, Mini-Mental State Examination; GHQ, General Health Questionnaire; Nt-sign, night signs.

2) *In our population-based setting, the ADVS far vision scale elicited two dimensions of functioning.* In fairness, the ADVS was not originally intended for risk factor analysis in a population-based setting but rather for evaluating effectiveness of cataract treatments, which may explain why such multidimensionality was not identified during ADVS development. Our estimated functioning profiles had good clinical face validity, suggesting separate "distance acuity" and "distance contrast" far vision dimensions similar to those reported by Rubin et al. (31).

3) *The tendency for women to report more difficulty than men was not driven by isolated items.* The gender association with functioning depended on the presence rather than the pattern of difficulty. While vision was strongly related to far vision functioning regardless of gender, women reported more difficulty of all types after accounting for vision status.

Model diagnosis

For each analytic method, we graphically checked final model fit. These diagnostic analyses indicated

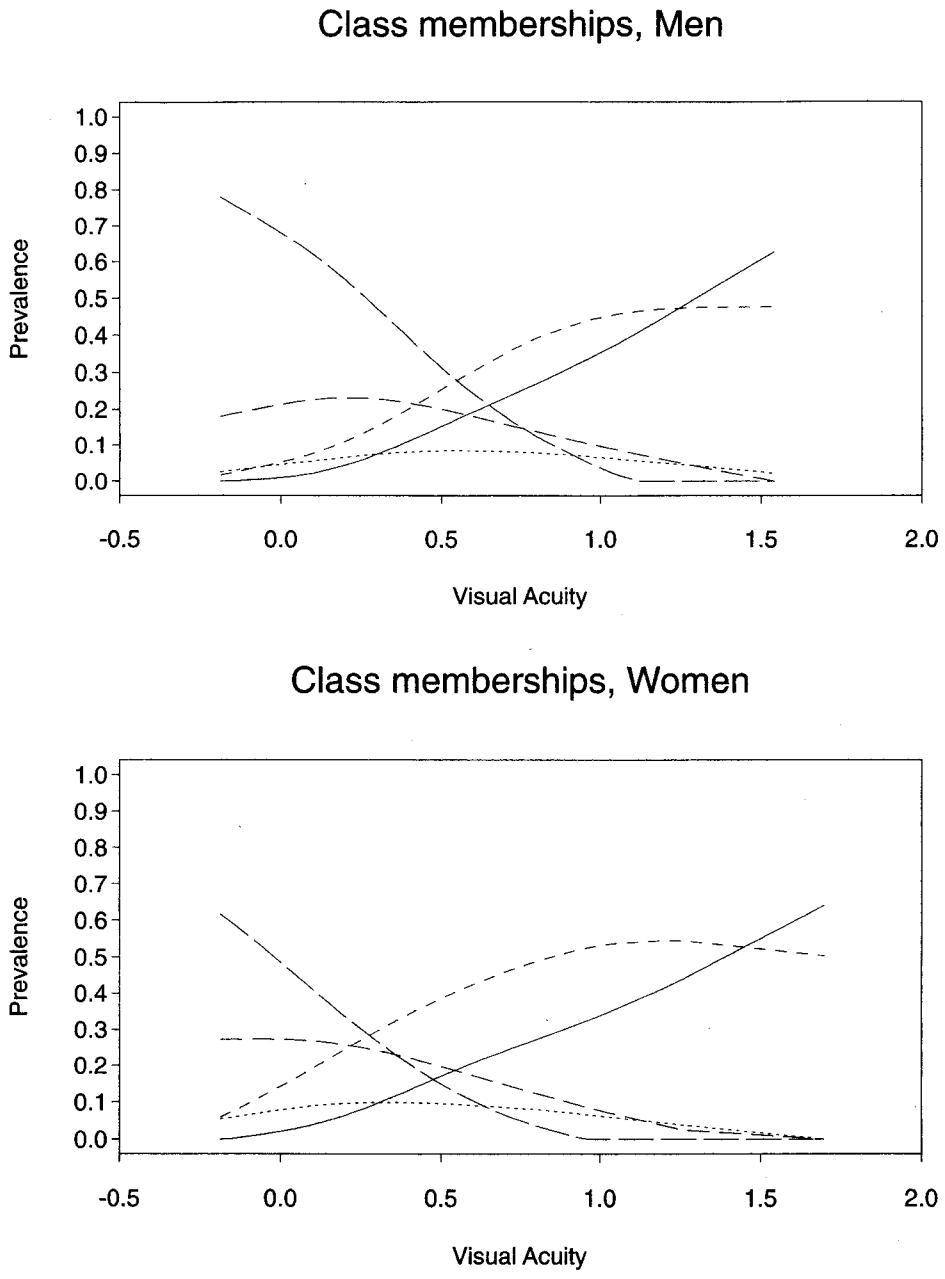

FIGURE 3. Estimated latent functioning class prevalences, by visual acuity and sex: Salisbury Eye Evaluation (SEE), Salisbury, Maryland, September 1993 to September 1995. Estimated prevalences are plotted versus acuity, separately for men (top panel) and women (bottom panel). In each panel, largest dash = class 1 ("able"); next largest dash = class 2 ("night signs"); small dash = class 3 ("signs"); dotted line = class 4 ("steps"); solid line = class 5 ("severe"). Curves were produced by fitting a smoothing spline through prevalence estimates at each visual acuity level.

that the scoring and item analyses fit the data well. Here, then, we focus on checking LCR model fit.

We began by comparing observed proportions reporting difficulty in each of the five far vision activities to those predicted by the LCR model, by sex and vision measure. First, we plotted (binary) difficulty data versus a psychophysical vision measure. Then, we described the observed relation between proportion reporting difficulty and vision by fitting a curve and its 95 percent confidence bands through the plot using the smooth.spline function with spar = 0.04 in S-PLUS (32). Next, we described the predicted relation

between proportion reporting difficulty and vision by plotting participants' fitted probabilities of reporting difficulty versus vision and then smoothing the plot exactly as for the observed data. If a model fits well, observed and predicted curves should be similar. In the display for visual acuity in men (figure 4), the observed (dashed line) and predicted (solid line) proportions agreed closely for most activities. However, the LCR model greatly underpredicted the proportion reporting difficulty of watching TV among persons with substantial acuity loss. This finding was replicated in women and across models that included fewer predictor variables. Thus, the LCR model inadequately explained the watching TV item responses.

Finding lack of fit in one item suggested that the LCR model's nondifferential measurement assumption was not satisfied. To see how one might investigate this, imagine that underlying functioning classes literally existed and that each person's underlying functioning status was known. Then, measurement would be differential if there were direct associations between predictors and item responses holding underlying functioning constant. While underlying functioning status is not known, we can estimate this by assuming an LCR model using participants' *posterior* probabilities of belonging to each (jth) functioning class:

$$\theta_j = \Pr\{L = j | x, Y_1 = y_1, ..., Y_M = y_M\} = \frac{P_j(x)\prod_{m=1}^{M}\pi_{mj}^{y_m}(1-\pi_{mj})^{1-y_m}}{\sum_{j=1}^{J}P_j(x)\prod_{m=1}^{M}\pi_{mj}^{y_m}(1-\pi_{mj})^{1-y_m}}, j = 1, ..., J. \quad (4)$$

Given posterior probability information, we propose model checking as follows (10):

1) Estimate each person's posterior probabilities of membership in classes $j = 1,...,J$, replacing parameters with estimates in equation 4;

2) Randomly assign each person to a class using his estimated posterior probabilities. (This can be done by generating a uniform random number u; then assigning the person to the class j such that $\sum_{k=1}^{j=1}\hat{\theta}_k \leq u \leq \sum_{k=1}^{j}\hat{\theta}_k$. If $u \leq \hat{\theta}_1$, assign the person to class 1.)

3) Stratifying on class assignments, examine whether there are substantial associations between predictors and reported functioning responses. If so, there is evidence of differential measurement.

Table 6 summarizes the first steps of this process—the randomized assignment to functioning classes. Multiple patterns of functioning may be reported by individuals randomized to a specific class; for the third model checking step, we grouped some such patterns together to eliminate sparse observed reporting counts. For instance, persons assigned to the "steps" class were grouped by whether or not they reported difficulty watching TV ("TV+" versus "TV–"), with each grouping subsuming several more specific reporting patterns.

To examine whether measurement appeared to be differential, we regressed the observed response groupings (table 6) on vision and other predictors, separately within each assigned class. We found large sex, race, and visual acuity associations (table 7). Roughly, among persons who reported functioning similarly on everything except watching TV, men, African-Americans, and persons with visual acuity impairments were substantially more likely to report difficulty watching TV than were women, whites, or persons without visual acuity impairments. Thus, LCR model checking added one more finding:

4) *"Watching TV" responses were associated differently with demographic and risk characteristics than were the other far vision item responses.* This clouds scoring and LCR analyses: reported gender associations may be artifactually low and acuity associations artifactually inflated. Summarizing, for risk factor analysis in a population-based setting, the "watching TV" item likely contaminates rather than improves the far vision measure and should arguably be removed from the scale.

DISCUSSION

We have detailed a three-method approach for analyzing associations between risk factors and questionnaire item responses. We were able to specify how vision contributes to far vision functioning substantially better using the multiple analyses than we could have by using any one analysis. As a general strategic recommendation: It is natural to first analyze scores for basic findings. Then, item and latent variable analyses are useful to ensure that scoring has not masked informative specificity in risk factor relations, particularly when a scale may be measuring a multidimensional construct. In this, mutually corroborating findings generally will comprise stronger scientific evidence that those that are supported by only one of the analytic approaches.

The polytomous part of the LCR model (equation 1) potentially involves many parameters. In light of multiple comparisons concerns, it is important to note that the proposed LCR formulation has a nice nested likelihood structure with respect to covariates. Specifically, the global null hypothesis of no association between

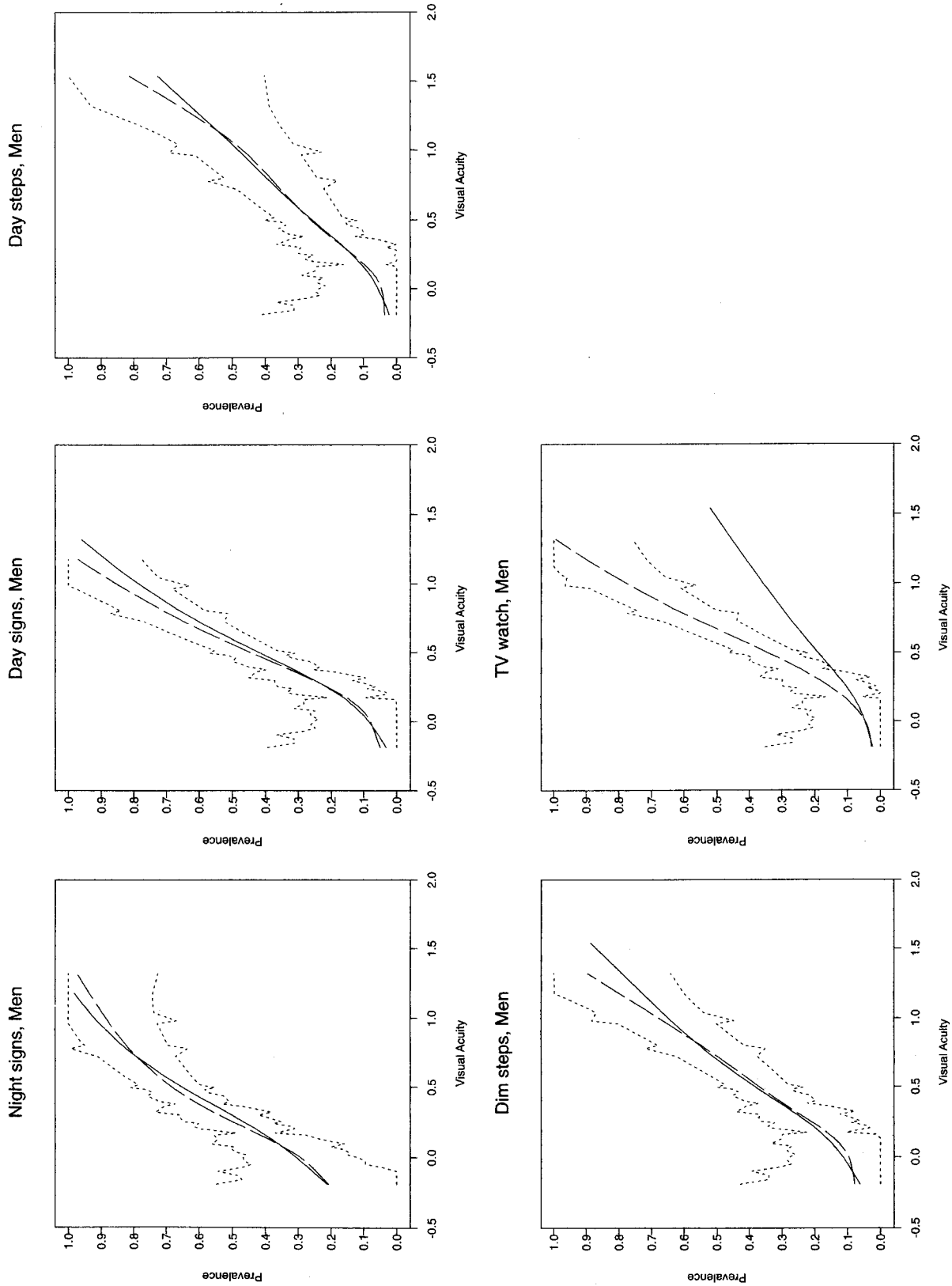

FIGURE 4. Observed and predicted item difficulty prevalences, by visual acuity, in men: Salisbury Eye Evaluation (SEE), Salisbury, Maryland, September 1993 to September 1995. Clockwise from top left, panels describe items: reading signs at night, reading signs during the day, walking down steps in daylight, watching TV, and walking down steps in dim light. In each panel: dashed line = empirical (observed); solid line = fitted from latent regression model (predicted); dotted lines = 95% confidence bands based on empirical. Observed and predicted curves constructed by fitting a smoothing spline through binary difficulty data (dashes at top and bottom of plots) and predicted item difficulty probabilities at each visual acuity, respectively.

TABLE 6. Latent class regression diagnosis—randomized class description: Salisbury Eye Evaluation (SEE), Salisbury, Maryland, September 1993 to September 1995

Class	Response groupings	No.	Composition summary
5. Severe	TV+	34	Difficulty on all activities or all except TV
	TV–	12	
4. Steps	TV+	16	Diverse. No "day sign" difficulty by a priori constraint
	TV–	80	
3. Signs	Signs + dim steps	45	Well summarized by given categories. "Other" patterns all
	TV+	45	exclude "dim steps" difficulty and subsume four individuals
	All but TV	25	with anomalous patterns including TV difficulty
	Signs only	44	
	Other	12	
2. Night signs	Night signs only	304	"Other" adds one difficulty—about equally split among day
	Other	93	signs, dim steps, TV
1. None	None	899	"Other" = one difficulty—largely night signs, TV, dim steps
	Other	34	

TABLE 7. Latent class regression diagnosis for nondifferential measurements: Salisbury Eye Evaluation (SEE), Salisbury, Maryland, September 1993 to September 1995

Class	Comparison	Variable	OR †	95% CI‡
4. Steps	TV+ vs. TV–	African American vs. white		
		Race	**5.20**	1.61, 16.8
		Education (4 years)	0.63	0.34, 1.18
		GHQ† (1 pt)	1.36	0.97, 1.90
3. Signs	Signs + dim steps vs. signs	Sex (female vs. male)	1.91	0.69, 5.27
		Comorbid (1)	1.12	0.83, 1.53
		VA† (0.3)	**0.28**	0.12, 0.63
		CS† (6)	**3.42**	1.44, 8.13
	Signs + TV vs. signs	Sex (female vs. male)	0.56	0.22, 1.40
		Comorbid (1)	**1.37**	1.02, 1.85
		VA† (0.3)	1.23	0.71, 2.16
		CS† (6)	1.51	0.69, 3.31
	All but TV vs. signs	Sex (female vs. male)	3.76	0.91, 15.5
		Comorbid (1)	0.83	0.55, 1.24
		VA† (0.3)	0.09	0.02, 0.39
		CS† (6)	2.21	0.73, 6.68

* Models were reduced one variable at a time to contain the listed variables only at per-variable criterion $p = 0.15$, to satisfy overall model significance criterion of $p = 0.05$.

† OR, odds ratio; CI, confidence interval; VA, visual acuity; CS, contrast sensitivity; GHQ, General Health Questionnaire.

outcome status and selected covariates may be tested by subtracting the –2 log likelihood of a model that includes the covariates versus that of a model that excludes the covariates. Under the null hypothesis, this statistic has a chi-squared distribution with degrees of freedom equal to $(J - 1)$ times the number of covariates in question. Such global testing restricts type I error probabilities and discourages highlighting isolated findings among many comparisons.

The LCR model (equation 2) describes measurement error in reporting through the π's: probabilities close to 0 or 1 represent homogeneous reporting among persons with the same underlying functioning whereas "middling" probabilities represent heterogeneous reporting. While this explicit accounting for measurement error is a benefit of LCR, the model's conditional independence assumption limits the sorts of measurement errors that can be described and specifically does not allow a person's errors to be correlated. Accounting for more general fluctuations in self-report is an important objective that merits further study.

In the latent variable application reported here, we handled missing item responses by excluding persons with incomplete data. While this is not ideal, we opted for such complete data analysis so as not to complicate our exemplification. Because the latent variable analyses did not contradict but rather illuminated scoring and item analyses, we believe it implausible that the additional information gleaned was primarily an artifact of restricting to complete data. In fact, it is not difficult to modify our algorithm for estimating LCR parameters to incorporate missing item responses. Such modification is ultimately critical given that missing responses are routine in questionnaire data.

The LCR fitting procedure suffers two computational difficulties. First, it may run for several hours and converge from different initializations to different solutions. For this report, we tried two very different initialization and model building strategies; the solutions were identical. We recommend such practice for LCR users. Second, a given data set may ambiguously suggest the number of latent classes. In our analysis, a four-class solution which merged the "none" and "night signs" classes appeared to fit adequately. We opted for the five-class solution for scientific reasons: there is substantial interest in the type of "early" disability that the "night signs" class might indicate, and there were many individuals who reported difficulty only in *reading signs at night* ($n = 306$). In practice, we recommend checking that fitted reporting probabilities appear consistent with the most commonly reported difficulty patterns. In our application, they did: in addition to the 306 persons already noted, 901 reported no difficulties at all, 75 reported difficulty reading signs at day and night, and so forth, consistent with the model π's.

Whereas latent variable models have a long tradition of application in the social sciences, they seem to engender skepticism among epidemiologists and biostatisticians. We acknowledge that latent variable models have dangerous power to drive scientific findings through their statistical assumptions. We find them usefui nonetheless, for several reasons. First, they force us to think about the type of measurement that a questionnaire intends and how effectively intended measurement is achieved. Second, they explicitly model the fact that similarly functioning persons do give different item responses; such modeling both may lessen analytic biases due to misclassification and supply data that are essential to improve future questionnaires. Finally, they give parsimonious summaries of complex data. The LCR model highlighted the essential feature of our outcome data—that of item reporting *patterns*. More broadly, the LCR approach not only describes item response associations with predictors but also augments scientific knowledge about which response patterns can be grouped sensibly for empirical analysis. In summary, a great deal can be learned from latent variable analyses provided that analysts evaluate model appropriateness in combination with empirical analyses.

ACKNOWLEDGMENTS

This work was supported by National Institute on Aging (NIA) Program Project P01 AG-10184-03, NIA contract N01-AG-1-2112-04, National Institutes of Health and NIA grant R01 AG-11703-01A1, and National Institutes of Mental Health grant R01-MH-56639-01A1. Dr. Bandeen-Roche is a Brookdale National Fellow.

REFERENCES

1. Katz S, Ford AB, Moskowitz RW, et al. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. JAMA 1963;185:914–18.
2. Rosow I, Breslau N. A Guttman health scale for the aged. J Gerontol 1966;21:556–9.
3. Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontologist 1969;9:179–86.
4. Robins LN, Helzer JE, Croughan J, et al. National Institute of Medicine Diagnostic Interview Schedule: its history, characteristics, and validity. Arch Gen Psych 1981;38:381–9.
5. Catlin FI. Studies of normal hearing. Audiol 1984;23:241–52.
6. Watters JM, Clancey SM, Moulton SB, et al. Impaired recovery of strength in older patients after major abdominal surgery. Ann Surg 1993;218:380–90.
7. Weiner D, Pieper C, McConnell E, et al. Pain measurement in elders with chronic low back pain: traditional and alternative approaches. Pain 1996;67:461–7.
8. Jette AM. Health status indicators: their utility in chronic-disease evaluation research. J Chronic Dis 1980;33:567–79.
9. Stewart A, Ware J. Measuring functioning and well-being. Durham, NC: Duke University Press, 1992:73–85.
10. Bandeen-Roche K, Miglioretti DL, Zeger SL, et al. Latent variable regression for multiple discrete outcomes. J Am Stat Assoc 1997;92:1375–86.
11. West SK, Munoz B, Rubin GS, et al. Function and visual impairment in a population-based study of older adults: SEE Project. Invest Ophthalmol Vis Sci 1997;38:72–82.
12. Folstein MF, Folstein SE, McHugh PR. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. J Psych Res 1975;12:189–98.
13. Munoz B, West SK, Rubin GS, et al. Who participates in population-based studies of visual impairment? The Salisbury Eye Evaluation Experience. Ann Epidemiol 1999;9:53–9.
14. Mangione CM, Phillips RS, Seddon JM, et al. Development of the "Activities of Daily Vision" scale: a measure of visual functional status. Med Care 1992;30:1111–26.
15. Valbuena M, Bandeen-Roche K, Rubin GS, et al. Self-reported assessment of visual function in a population based setting. Invest Ophthalmol Vis Sci 1999;40:280–8.
16. Rubin GS, West SK, Munoz B, et al. A comprehensive assessment of visual impairment in an older American population: SEE study. Invest Ophthalmol Vis Sci 1997;38:557–68.
17. Goldberg D. GHQ the selection of psychiatric illness by ques-

tionnaire. New York: Oxford University Press, 1972.
18. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London: Chapman and Hall, 1989.
19. McCullagh P. Regression models for ordinal data (with discussion). J R Stat Soc B 1980;42:109–42.
20. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley, 1989:216–38.
21. Morrison DF. Multivariate statistical methods. 3rd ed. New York: McGraw-Hill, 1990.
22. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–74.
23. Bryk AS, Raudenbush SW. Hierarchical linear models applications and data analysis methods. Newbury Park, CA: Sage, 1992.
24. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
25. Fitzmaurice GM, Laird NM. A likelihood-based method for analysing longitudinal binary responses. Biometrika 1993;80: 141–51.
26. Agresti A. Analysis of categorical data. New York: Wiley, 1989.
27. Dayton GB, Macready CM. Concomitant-variable latent-class models. J Am Stat Assoc 1988;83:173–8.
28. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 1974;61: 215–31.
29. McCutcheon AL. Latent class analysis. Newbury Park, CA: Sage, 1987.
30. Heagerty PJ, Zeger SL. Marginal regression models for clustered ordinal measurements. J Am Stat Assoc 1996;91: 1024–36.
31. Rubin GS, Bandeen-Roche K, Prasada-Rao P, et al. Visual impairment and disability in older adults. Optom Vis Sci 1994;71:750–60.
32. Statistical Sciences, Inc. S-PLUS for Windows user's manual, Version 3.1. Seattle, WA: Statistical Sciences, Inc, 1993.