Detecting measurement confounding in epidemiological research: construct validity in scaling risk behaviours: based on a population sample in Minnesota, USA

Kathryn Dean, Nagi Salem

Abstract

***Study objective*—The purpose of this study was to construct and validate a behavioural risk index scale to determine if the scaled variable could be used to study possible latent dimensions of risk behaviour.**

***Design*—Data from the Minnesota Behavioral Risk Factor Surveillance System (BRFSS) were used to construct and validate a behavioual risk index using item response theory methods.**

***Participants*—A representative sample of 3303 Minnesota adults 18 years of age and older.**

***Results*—Massive evidence was found against the construct validity of the index scale as a measure of risk behaviour. Seven commonly studied risk behaviours could not be scaled into a valid construct of health behaviour for either men or women. Tests of scalability, homogeneity, and item independence were rejected. In addition, item bias was found for all of the items in relation to important exogenous variables, especially age and education.**

***Conclusions*—The risk behaviours do not represent sufficiently similar types of phenomena to form an additive scale of health related risk taking. Not only do the practices fall on different, undefined dimensions of behaviour, subgroup differences in risk taking would be hidden by data reductions summing the behavioural practices into additive scales. The findings indicate that the behaviours have quite distinct meanings that should be studied separately.**

(*J Epidemiol Community Health* 1998;**52**:195–199)

Population Health Studies, Copenhagen, Denmark
K Dean

Minnesota Department of Health Minneapolis, Minnesota, USA
N Salem

Correspondence to:
Dr K Dean, Population Health Studies, Ribegade 6 st tv, DK-2100 Copenhagen, Denmark.

Accepted for publication
5 June 1997

While it is known that behavioural habits are relatively stable, it is also known that health related beliefs and attitudes often are not reflected in the behavioural practices of people.[1–3] In recent years there has been a growing tendency to construct index scales to represent concepts used by researchers and policy analysts. The possibility of classifying the adult population according to levels of risk has generated interest in risk scales. It would be useful to know if there exists special tendencies toward health promotion or health related risk behaviour in certain people that can be identified and studied in population health data. Research to test for possible latent dimensions of health related risk behaviour not readily measured by asking people about their health beliefs is needed before it can be determined if the population can be validly classified according to risk behaviour. Successful scaling of information on health related behavioural practices into valid index measures would provide evidence suggesting personal health related behavioural perspectives that could be studied to obtain new knowledge for health protection. A prerequisite for achieving this goal is to assure the validity of any measures constructed to study latent risk tendencies.

Whether or not there exists latent tendencies toward health related risk, it is important to evaluate the validity of scaling questions on health related habits from behavioural risk factor data. Inconsistencies and weaknesses in the research evidence about risk behaviours related to health and longevity have directed increasing attention to methodological problems that must be resolved to obtain sound knowledge about behavioural risks.[4–6] The problems are created by systematic errors, biases, and confounding that can overwhelm statistical variation and must be detected to obtain sound results in risk factor research.[7 8] The purpose of this project was to construct and validate a health behaviour index scale to determine if it could be used as a valid measure of behavioural risk.

The construction of index scales is based on the idea that they represent some type of underlying variable that cannot be measured directly. Methods for the validation of measurement scales were developed in the field of psychometrics. Evaluating construct validity tests the extent to which theoretical concepts account for the empirical results obtained from using an index constructed to measure the concept.[9 10] Construct validity is generally considered in relation to criterion validity, a type of validity where other variables that are considered measures of the theoretical concept under study are used along with methods for testing construct validity to validate an index measure. According to Rosenbaum,[11] when an index scale predicts criterion variables, several conditions must hold to reasonably conclude that the construct is measuring the latent characteristic: higher values on the criterion variable/s are associated with more of the latent characteristic; the index scale measures the latent characteristic rather than systematically measuring anything else; and, the values obtained with the

index scale predict the criterion variable/s because the items included in the scale measure the latent characteristic and not for some other reason.

Methods

DATA

The Minnesota Behavioral Risk Factor Survey, conducted in collaboration with the US Centers for Disease Control is designed to monitor risk behaviours and expand knowledge about the impact of behavioural practices on health. Random digit dialling and telephone survey methods are used to conduct monthly interviews with a representative sample of the population over 18 years of age. In the context of this system the Minnesota BRFSS is unique in terms of the diversity of the data available and the relatively large sample size.

The 1992 Minnesota BRFSS data were used in this study to construct and validate a behavioural index. The data were obtained from 3033 adult residents of Minnesota. Questions potentially suitable for constructing a positive health behaviour index were identified. Seven items developed from the data were constructed into a health behavioural index by adding the number of positive categories, creating risk scores ranging from 0 to 7. The seven items included in the index were:

(1) Seat belt use (1 = always/nearly always, 0 = sometimes/seldom/never)

(2) Physical activity (1 = non-sedentary lifestyle, 0 = inactive/sedentary lifestyle)

(3) Overweight based on body mass index (1 = normal weight, 0 = overweight)

(4) Fruit and vegetables (1 = 5 or more servings per day, 0 = less than 5 servings per day)

(5) Smoking (1 = never/former, 0 = current smoker)

(6) Alcohol consumption (1 = none/less than 60 drinks per month, 0 = more than 60 drinks per month/binge drinkers)

(7) Drinking and driving (1 = never, 0 = at least once in prior month)

Six exogenous variables, with categories shown, were included in the validation of the risk behaviour scale to test for item bias and group dependency:

(1) Age (18–24, 25–34, 35–44, 45–54, 55–64, 65 and over);

(2) Education (11 years or less, high school graduate, attended college, college graduate or graduate school);

(3) Employment (wage earner or self employed, unemployed, homemaker, student, retired);

(4) Household composition (one adult, more than one adult, one adult with one or more children, more than one adult with one or more children);

(5) Marital status (married or unmarried couple, divorced or separated, widowed, never married);

(6) Area of residence (Twin Cities metro, greater Minnesota).

VALIDATION PROCEDURES

Two approaches to the validation of measurement scales developed in the field of psychometrics, classic psychometric methods and methods based on Item Response Theory, are widely used to construct and validate index measures. The validation of index scales in investigations of population health and behaviour have traditionally been based on classic psychometric measurement procedures. In recent years important limitations of these methods have become recognised.[12–14] The statistical models used to construct and validate scales in the classic psychometric theoretical paradigm are based on assumptions that the scores are linear functions of latent variables and that the scores are independent of measurement errors. Both scores and errors are assumed to be normally distributed. The consequences of violating these assumptions are difficult to determine and often not considered.

Another limitation of the classic methods has to do with the inability to separate the characteristics of the subject and the characteristics of the index because with these methods the reliability and the validity of the index are defined in relation to the particular group used to develop the instrument.[12] This makes scales constructed in the classic psychometric framework group dependent on the subjects used in the construction of the scale. Psychometric scale construction is traditionally concerned with the measurement of ability or other person attributes that are considered relatively stable characteristics of people that may be normally distributed in general populations. This investigation concerns personal behavioural practices that have been associated with statistical risk for specific diseases or mortality. Distributions of risk behaviours are often skewed in population samples. The population groups that display greater proportions of risk behaviours may contribute disproportionately to measurement error, bias, and confounding that can overwhelm statistical variation. The need to assure the second and third conditions mentioned above for testing construct validity is readily apparent.

Some of the problems encountered in using classic psychometric scale validation techniques are avoided by using procedures based on the alternative paradigm. Item Response Theory does not build on assumptions derived from the theory of normal distributions. In Item Response Theory, probabilities of specific responses to the items of a scale are considered general functions of item parameters and either person parameters or latent person variables characterising the variable the scale intends to measure.[13] A scale is considered adequate if items are independent given the summary index measure. An important part of the validation process is to assure that the characteristics of the members of the sample used to validate the scale do not influence the results in any systematic way. Independent behavioural items and item bias[15] for subgroups of the population are important considerations in the construction of behavioural scales.

Item Response Theory methods were selected for the validation of the health risk behaviour index. The methods used include techniques for testing construct validity, item

Table 1 Distribution of health risk behaviour among Minnesota residents by sex, 1992

	Men		*Women*	
*Behaviour risk score**	*Number*	*%*	*Number*	*%*
0	1	0.1	0	0.0
1	10	0.7	4	0.2
2	74	5.0	20	1.1
3	189	12.8	133	7.3
4	319	21.6	303	16.6
5	387	26.2	529	28.9
6	338	22.9	576	31.5
7	157	10.6	263	14.4
Total	1475	100.0	1828	100.0

* Scores refer to number of behavioural risks reported; seat belt use; physical activity; weight based on BMI; consumption of fruits and vegetables; smoking behaviour; alcohol consumption; drinking and driving.

KEY POINTS

- Behaviour practices associated with statistical risk for poor health outcomes cannot be automatically combined into index scales.
- Invalid measures contribute to variable confounding and biased results in epidemiological research.
- It is important to evaluate the consequences of constructing index scales from data collected from population samples.

independence, (both with regard to other items included in the scale and to exogenous variables that would be included in multivariate analyses), and item bias. Methods based on the Rasch model[16] were used to test the scalability of the behavioural items. The model assesses unidimensionality (the scale measures only a tendency toward health protective behaviour and not something else), and monotonicity ("correct" responses to each item are more common among the respondents with a greater tendency toward health behaviour). The Rasch model is a binary linear logistic model that tests for the monotonicity and independence of the items scaled into the hypothesised construct. Conditional likelihood tests proposed by Anderson[17] were used to test homogeneity, or the extent to which the items measure the hypothesised attribute regardless of the characteristics of particular sub-populations.[14] Methods used to validate or check the fit of the tests of scalability were suggested by Rosenbaum[18] to test the conditional independence of the items given the scale. If the scaled items are measuring an underlying latent tendency, all items should be bivariately correlated with each other, but no longer correlated when they are examined in relation to the scale, which should collect the variation due to the latent construct. Contingency table methods measuring Goodman-Kruskal's partial gamma coefficients for ordinal level variables and χ^2 for nominal variables were used to test for item bias.[13 19]

Results

Table 1 shows the distributions of the health behaviours. Only one man and none of the women had a zero score. Most of the sample members fell in the score categories 4 to 6, with 10% of the men and 15% of the women obtaining the maximum score of 7.

The validation analysis found that the seven behavioural items could not be scaled into a valid measure of health behaviour for either men or women. The Rasch likelihood ratio test result for the female scale was z=7.9, p=0.248, while the test result for men was z=41.8, p=0.000. By itself, these results would indicate construct validity for women, but clear and strong rejection of the hypothesis of scalability for men. However, the procedures used for checking the model found massive evidence against construct validity for both men and women. All conditional likelihood ratio tests of the model taking into account subgroups of the population were rejected. Likewise, item bias was found for all of the items in relation to important exogenous variables, especially age and education. This is illustrated for physical activity on the male scale and alcohol consumption on the female scale in tables 2 and 3.

Negative correlations between some of the item pairs found for both men and women illustrated that items in the risk scale represented different types of phenomena rather than tapping a unidimensional risk construct. For both men and women the low risk items on

Table 2 Item bias for the male physical activity behaviour factor included in an addititive index of health behaviour, Minnesota BRFSS, 1992

Item	χ^2	*Degrees of freedom*	*p Value (exact)*	*Gamma*	*p Value (exact)*	*Significant bias*
Age	52.7	10	0.000	−0.24	0.000	Yes
Educational attainment	15.7	6	0.020	0.14	0.000	Yes
Employment status	9.1	5	0.095			No
Household composition	6.0	6	0.415			No
Marital status	51.3	6	0.000			Yes
Area of residency	10.4	2	0.002	−0.18	0.000	Yes

Table 3 Item bias for the female drinking behaviour factor included in an addititive index of health behaviour, Minnesota BRFSS, 1992

Item	χ^2	*Degrees of freedom*	*p Value (exact)*	*Gamma*	*p Value (exact)*	*Significant bias*
Age	121.7	10	0.000	0.54	0.000	Yes
Educational attainment	21.5	6	0.005	−0.24	0.000	Yes
Employment status	33.8	6	0.000			Yes
Household composition	7.1	6	0.303			No
Marital status	87.1	6	0.000			Yes
Area of residency	5.2	2	0.075	0.18	0.013	Yes

Table 4 Result of testing the conditional independence for item pairs in the male behavioural risk factor index

	Seat belt t	*Physical activity x*	*Overweight w*	*Fruit and vegetables f*	*Smoking k*	*Drinking l*	*Drinking and driving v*
t		−0.07	−0.11	−0.10	0.02	0.19**	0.36**
		0.13	0.14	0.03	0.14	0.28**	0.19
x			0.13	0.11	0.00	−0.39**	−0.15
			0.19**	0.02	−0.08	−0.48**	−0.49**
w				−0.03	−0.29**	−0.09	−0.02
				−0.12	−0.41**	−0.23**	−0.50**
f					0.06	0.05	−0.14
					0.04	−0.01	−0.50**
k						0.25**	0.30
						0.23**	−0.06
l							0.96**
							0.94**
v							

The value presented is the conditional gamma correlation coefficient for item pairs given the score after removing the respective items. Significant at p value ⩽ 0.01. ** Significant at p value ⩽ 0.001.

smoking and weight were negatively correlated. Additionally, for women not being overweight was negatively correlated with the low risk items on alcohol consumption and drinking and driving, as were the low risk item categories on physical activity and alcohol consumption. Some of the items were not correlated at all, as for example weight and consumption of fruit and vegetables. As shown in tables 4 and 5, the Rosenbaum procedure, testing the conditional independence of the item pairs given the scale, confirmed and further explicated the problems with the scale. The tables present the correlations for the item pairs controlling for the scale values. Additional negative correlations between items emerged after the scale was included in the model. At the same time, a number of item pairs (low alcohol consumption and not driving with alcohol, low alcohol consumption and seat belt use, non-smoking and low alcohol consumption; and for women, not being overweight with both physical activity and seat belt use) are positively correlated after controlling for the scale, indicating that the items represent variation in the data that is not tapped by the scaled construct.

Following these results, the question remained about the possibility of achieving acceptable data reductions by combining some of the variables into smaller groupings that might represent some dimension of behaviour and that could be used for more parsimonious analysis of the larger data set. Separating the pairs of items that were negatively correlated or not correlated into smaller scales it was found that for men, physical activity, fruit and vegetable consumption, and not being overweight could be combined, and that non-smoking, seat belt use, and not driving with alcohol could also be combined. For women, fruit and vegetable consumption, non-smoking, and seat belt use could be combined. However, even for these three item scales, the serious item bias problems would preclude their use in multivariate analysis, and no assumptions could be made about their meaning.

Discussion

In the analyses conducted in this project to validate a behavioural risk index scale it was found that seven commonly studied health related behaviours could not be validly scaled into an index variable representing a dimension of positive/negative health related behaviour. If the items reflected an underlying dimension of health behaviour/risk taking, all item pairs would be positively correlated at the bivariate level, with the correlations disappearing when examined with the scale included as the measure of risk behaviour. Not only did some of the behaviours represent variation not tapped by the scale, the negative correlations found between some of the item pairs show that fundamentally different dimensions are represented by the items included in the scale.

In addition to the evidence against an underlying behavioural risk dimension that could by studied by including the seven behaviours in an additive scale, item bias was found

Table 5 Result of testing the conditional independence for item pairs in the female behavioural risk factor index

	Seat belt t	*Physical activity x*	*Overweight w*	*Fruit and vegetables f*	*Smoking k*	*Drinking l*	*Drinking and driving v*
t		−0.17	−0.14	−0.04	0.23**	0.28**	−0.23
		−0.01	0.14	0.05	0.28**	0.36**	−0.46
x			0.23**	0.02	−0.02	−0.60**	−0.23
			0.39**	0.00	−0.09	−0.65**	−0.50**
w				−0.19**	−0.41**	−0.26	−0.49
				−0.29**	−0.56**	−0.39**	−0.78**
f					0.19**	0.16	0.49
					0.19**	0.21	0.14
k						0.35**	−0.02
						0.44**	−0.29
l							0.92**
							0.94**
v							

The value presented is the conditional gamma correlation coefficient for item pairs given the score after removing the respective items. Significant at p value ⩽ 0.01. ** Significant at p value ⩽ 0.001.

for all of the items in relation to important exogenous variables. The consistent strong item bias found in relation to age and education suggests that the behaviours generally represent quite different phenomena for people with different levels of education and at different ages.

It must be concluded that the risk behaviours do not represent sufficiently similar types of phenomena to form an additive scale of health related risk taking. Not only do the practices fall on different, undefined dimensions of behaviour, subgroup differences in risk taking would be hidden by data reductions summing the behavioural practices into additive scales. Thus, no evidence was found to suggest potential for constructing a valid risk index that could be used for classifying the population on a risk dimension. On the contrary, important policy relevant differences would probably by hidden or confounded in research attempting to use risk indices of this sort. The findings in this study illustrate the importance of evaluating the consequences of constructing index scales from data collected from population samples.

Funding: this research was supported by a grant from the US Centers for Disease Control and Prevention and by the Minnesota Department of Health.
Conflicts of interest: none.

1 RUHBC (Research Unit in Health and Behavioural Change). *Changing the public health*. New York: John Wiley, 1989.
2 Dean K. Influence of health beliefs on lifestyle: what do we know? *European Monographs in Health Education Research* 1984;**6**:127–49.
3 Blaxter M. *Health and lifestyles.* London: Routledge, 1990.
4 Anonymous. Population health looking upstream. *Lancet* 1994;**343**:429–30.
5 Dean K, Kreiner S, McQueen D. Researching population health: new directions In: Dean K, ed. *Population health research: linking theory and methods*. London: Sage Publications, 1993.
6 Rose G. Sick individuals and sick populations. *Int J Epidemiol* 1985;**14**:34–8.
7 Taubes G. Epidemiology faces its limits. *Science* 1995;**269**: 164–9.
8 Dean K, Holst E, Kreiner S, Schoenborn C, Wilson R. Measurement issues in research on social support and health. *J Epidemiol Community Health* 1994;**48**:201–6.
9 Cronbach L, Meehl P. Construct validity in psychological tests. *Psychol Bull* 1955;**52**:281–302.
10 Cronbach L. Test validation, educational measurement. In: Throndike R, ed. *Educational measurement*. Washington, DC: American Council on Research in Education, 1971.
11 Rosenbaum P. Criterion-related construct validity. *Psychometrika* 1989;**54**:625–34.
12 Hambleton R, Swaminathan H, Rogers H. *Fundamentals of item response theory.* Newbury Park: Sage Publications, 1991.
13 Kreiner S. Validation of index scales for analysis of survey data: the symptom index. In: Dean K, ed. *Population health research: linking theory and methods*. London: Sage Publications, 1993.
14 Holst C. *Item response theory*. Copenhagen, Danish Institute of Educational Research, 1995.
15 Camilli G, Shepard L. *Methods for identifying biased test items*. Thousand Oaks: Sage Publications, 1994.
16 Rasch G. *Probabilistic models for some intelligence and attainment tests*. Vol 1. Studies in mathematical psychology. Copenhagen: Danish National Institute for Educational Research,1960.
17 Andersen E. A goodness of fit test for the Rasch model. *Psychometrika* 1973;**38**:123–40.
18 Rosenbaum P. Testing the conditional independence and monotonicity assumption of item response theory. *Psychometrika* 1984;**49**:425–35.
19 Kreiner S. Analysis of multidimensional contingency tables by exact conditional tests, *Scandinavian Journal of Statistics* 1987;**14**:97–112.