ELSEVIER

Drug and Alcohol Dependence 41 (1996) 101–112

DRUG and ALCOHOL DEPENDENCE

Psychometric evaluation of diagnostic criteria: application to a two-dimensional model of alcohol abuse and dependence

Bengt O. Muthén

Graduate School of Education, Information Studies, UCLA, Los Angeles, CA 90095-1521, USA

Received 1 November 1993; accepted 10 January 1996

Abstract

Psychometric investigations of diagnostic criteria can be helpful in the refinement of psychiatric instruments. This paper illustrates a methodology for investigating the measurement properties of a set of diagnostic criteria. The analyses are based on a two-dimensional factor analysis model for alcohol abuse and dependence. Based on this model, the methodology shows how cutpoints for diagnoses can be evaluated and defined, with which precision the criteria measure abuse and dependence, how well abuse without dependence can be measured, if the criteria should be weighted or not, if additional criteria are needed to improve measurement, if a smaller number of criteria could be used with almost as good results, and if diagnoses can be made with reliable results. The application of the methodology to the study of alcohol abuse and dependence in general population surveys shows important implications for diagnosis and prevalence estimation based on DSM criteria.

Keywords: Diagnostic criteria; Psychometric evaluation; Factor model for alcohol abuse and dependence

1. Introduction

Psychometric investigations of diagnostic criteria can be helpful in the refinement of psychiatric instruments. Such investigations provide enhanced understanding of how criteria function and how various diagnostic schemes compare. In this paper, a set of criteria measuring alcohol abuse and dependence is studied. These criteria were created from a set of symptom items that were developed to measure the diagnostic criteria for alcohol abuse and dependence of the Diagnostic and Statistical Manual, Third Edition-Revised (DSM III-R) and the proposed DSM-IV (American Psychiatric Association, 1987, 1992). They were included in the 1988 National Health Interview Survey (NHIS88).

Muthén et al. (1993a) analyzed the dimensionality of DSM-III-R criteria for alcohol abuse and dependence from the NHIS88 and found two dimensions, one corresponding to a less severe factor measured by more prevalent symptom items than the second, more severe factor. Muthén et al. (1993b) replicated these findings in subgroups defined by age, gender, and ethnicity, although the factor patterns differed somewhat across these groups. Muthén et al. (1993c) recovered the same major dimensions for the items of the ICD-10 criteria. Muthén (1995) also recovered the two dimensions in NHIS88 data when carrying out the analysis on the individual symptom items without first categorizing them into DSM criteria. Muthén (1993) used the two-dimensional model for the DSM criteria of NHIS88 to formulate a latent variable regression model relating the dimensions to background variables including alcohol consumption, age, gender, ethnicity, and family history of alcoholism. It was found that the two dimensions related differently to the background variables; for example, family history of alcoholism had a stronger influence on the more severe factor than the less severe factor.

This paper discusses psychometric methodology for investigating the measurement properties of a set of diagnostic criteria. The analyses are illustrated with the Muthén et al. (1993a) two-dimensional factor model for the alcohol dependence and abuse criteria. The paper answers questions such as: how can cutpoints for diagnoses be evaluated, with which precision do the criteria measure abuse and dependence, should the criteria be weighted or not, are additional criteria needed to improve measurement, could a smaller number of criteria be used with almost as good results, and can diagnoses be made with reasonable sensitivity and specificity.

PII S0376-8716(96)01226-4

#66

Psychometric theory behind the factor model used by Muthén et al. (1993a) offers several ways of evaluating how well the criteria measure the factors. It introduces terms that are less common to epidemiological research, such as factor scores, information curves, and expected score functions. This paper will contribute to this theory by also adding the more familiar epidemiological terms of sensitivity and specificity of diagnoses.

The factor model used by Muthén et al. (1993a) postulates continuous factors underlying the observed dichotomous diagnostic criteria. Although this model expresses a 'dimensional', as opposed to a 'categorical' view of alcohol dependence and abuse, the model can also serve as a basis for dichotomous diagnoses using cutpoints on the continuous factors. It is important to make clear that the psychometric theory and analyses derived from this factor model can shed light on such cutpoints in two fundamentally different ways: evaluating a priori cutpoints and choosing suitable cutpoints.

The evaluation of a priori cutpoints would probably be the typical use of the psychometric procedures. Here, substantive theory specifies a set of diagnostic criteria and established conventions determine the scoring approach, i.e. the number of criteria that need to be fulfilled for a diagnosis and how the criteria should be weighted. This determines the prevalence of a certain problem and the prevalence can in turn be translated into a cutpoint for the factor in question. The psychometric properties of the diagnostic criteria and the scoring approach for making this diagnosis can then be evaluated. Switching from a purely evaluatory mode, the analysis may suggest refinements that can be made by either modifying the set of criteria or the scoring approach.

The choice of suitable cutpoints, on the other hand, involves using the psychometric techniques to find cutpoints that have good psychometric properties. This may again be carried out by a prespecified set of criteria, but here letting the psychometric procedures determine the cutpoint and thereby the prevalence. Alternatively, the set of criteria may be modified in order to enhance the psychometric properties at certain cutpoints.

There are several psychometric techniques that are useful for both of the above two analysis purposes. From a practical point of view, it is useful to distinguish between three different aims behind these techniques: studying the distribution of the latent variable values, studying the precision of the latent variable measurement, and studying the sensitivity and specificity of diagnosis.

1.1. Studying the distribution of the latent variable values

The factor model may be used to estimate factor scores for each individual based on that individual's response pattern, the response pattern in this case being the series of 0s and 1s for the diagnostic criteria (1 denoting that a criterion is fulfilled). The scores are estimates of the true values for each individual and therefore have more central importance than the criteria, which are viewed as fallible indicators of the factors. When the aim is to choose a cutpoint, it is of interest to study the distribution of these scores over all individuals in the sample and to determine if there are any 'natural breaks' in this distribution, where a minority of the individuals cluster within a clearly distinguishable and considerably higher range of factor score values than the majority of the individuals.

1.2. Studying the precision of the latent variable measurement

The factor model offers an assessment of the information function for the criteria. The information function describes how the precision in the estimation of factor scores from responses to the criteria varies as a function of the factor score value. For any given set of criteria, the information value varies with the factor score values. The information value may be thought of as the ability of the score to discriminate between two individuals with different factor values. It is a function of the quality and number of criteria involved. The notion of the information value is also applicable to scoring approaches based on summing the criteria, possibly with different weights. When the analysis aim is to choose a cutpoint, one possible candidate is the factor score value for which the information peaks.

For any given scoring approach, the model also provides an expected score as a function of the factor value. The steepness of the expected score function in a certain range of factor values describes the ability to discriminate between two individuals in that range. It is, therefore, directly related to the information value in that range of values.

1.3. Studying the sensitivity and specificity of diagnosis

The factor model offers a way to study misclassification of individuals as, say, alcohol abusers or alcohol dependent by comparing a 'correct' classification based on factor values being below or above a cutpoint with a classification based on scores from (weighted) sums of criteria. The properties of a scoring approach can thereby be assessed by means of percentage of false positives and false negatives, the sensitivity (proportion of cases diagnosed correctly), and the specificity (proportion of non-cases not diagnosed). The sensitivity and specificity values can also aid in choosing a factor score cutpoint when that is the aim of the analysis.

1.4. Dimensional versus diagnostic assessment

The above three aims of the psychometric techniques are all formulated with the idea of making diagnoses. Because the factor model uses a dimensional view of the latent variables, an alternative to making diagnoses is also available. The dichotomization based on a cut-point that is involved in the diagnosis may be avoided by instead reporting an individual's percentile on the continuous latent variable based on his/her estimated factor score value. This may be particularly warranted in situations where there is no natural break in the factor score distribution. Prevalence for different subgroups of the population can be reported and compared using the number of subgroup members above a certain percentile. This paper proposes that the alternative of using percentiles be considered in reporting results.

2. Methods

This section gives an overview of the psychometric techniques. While there is not space to give a self-contained description, the aim is to give sufficient detail so that researchers can replicate these methods on their own data, at least with the help of statistical analysts.

2.1. The factor model for binary criteria

While the factors are continuous latent variables, the criteria are binary variables. Factor analysis of binary variables requires special techniques that use non-linear regressions of the variables on the continuous, unobserved factor. This is analogous to logistic regression or probit regression of a binary dependent variable on a continuous, observed predictor. Early methodology was developed under the name latent trait theory, but item response theory (IRT) is now more commonly used. IRT has been proposed for psychiatric applications (see e.g. Duncan-Jones et al., 1986). IRT is a special form of factor analysis of binary variables in that the analysis is restricted to a single dimension. For the analyses of this paper, a more general, multiple-factor analysis was used, see e.g. Muthén (1978). For a general introduction to latent trait theory and binary factor analysis, the reader is referred to Hambleton and Swaminathan (1985), Duncan-Jones et al. (1986), and Muthén (1989).

In line with probit regression, the factor model for binary criteria describes the probability of a criterion being fulfilled or not as a non-linear function of the values of the factors. As with probit regression, the normal distribution function is used. The factor loadings can be translated to coefficients in a probit regression. A positive loading implies that for increasing factor value, the probability of fulfilling the criterion increases. Examples of loadings translated to probit coefficients are given in Table 2 under the heading 'Optimal weights'. These coefficients are obtained by dividing each loading with the corresponding residual's standard deviation, where the residual variance is obtained as 1 minus the communality (the variance contributed by the factors). The probit intercepts may be obtained as the negative of standard normal scores corresponding to the proportions, divided by the corresponding residual standard deviations. The factors are standardized to zero means and unit variances.

2.2. Studying the distribution of the latent variable values

The set of criteria gives rise to several different possible patterns of fulfilled and not fulfilled criteria. For each response pattern observed in the data, factor scores on each of the dimensions can be estimated (see e.g. Bock and Aitkin, 1981). Maximum-likelihood (ML) estimated factor scores can be computed by iterative optimization techniques. The ML theory provides an estimated standard error for a given factor score estimate. Alternatively, Bayesian estimation may be employed.

Each individual's response pattern can also be used to provide scores based on unweighted or weighted sums of the criteria. For each factor, the weighting of the criteria provided by the optimal weights is the one that maximizes the information value (see below) on that factor. The optimal weights are independent of the factor value (Hambleton and Swaminathan, 1985, p. 117). The scores may be viewed as an alternative approach to estimating the factor values. This provides a straightforward, non-iterative method which simply entails using a weighted sum of criteria scored 0 or 1. Using unit weights is a particularly simple approach. In this paper, ML estimated factor scores will not be used, but factor scores will instead be estimated using the optimal weights. Experience has shown that the two methods give very highly correlated values (see e.g. Duncan-Jones et al., 1986).

2.3. Studying the precision of the latent variable measurement

In models with a single dimension, the inverted value of the variance for a given factor score estimate represents the statistical term 'Fisher information'. Standard IRT takes this as the information value for that factor value (Hambleton and Swaminathan, 1985), describing the precision of the factor score estimation. The information value is higher with criteria that have steep probit curve slopes (large optimal weights) and increases with increasing number of criteria. The information value varies over factor values and is high in the

range of factor values corresponding to the typical values of the probit intercepts for the criteria measuring the factor. Since each probit intercept is a function of the corresponding criterion prevalence, this means that the information value for low factor values is high for a set of criteria with many highly prevalent criteria. Conversely, the information value for high factor values is high for an instrument with many less prevalent criteria.

In the present paper, the information value notion is generalized to two dimensions by considering the diagonal elements of the inverted 2 × 2 information matrix. This pertains to estimating the two factor scores simultaneously. The information curve will be plotted as a function of one of the factors, where the value of the other factor is allowed to vary as the expected value of that factor given the first. An alternative approach will be used for considering the information value for scores estimated by unweighted and weighted sums of criteria. Since this estimation is geared towards a single factor, the information value is presented corresponding to the estimation of one factor, holding the other fixed (not estimated) at its expected value. In this way, information curves will be compared for ML factor score estimates, estimates obtained by optimal weights, and estimates obtained by unit weights.

The expected score for a given weighting approach is a sigmoid shaped function of the factor value (see e.g. Hambleton and Swaminathan, 1985, p. 103). This curve can be plotted for each factor, comparing an unweighted score based on all criteria with the optimally weighted scores. The steepness of the curves is of particular interest.

2.4. Studying the sensitivity and specificity of diagnosis

This paper proposes a method based on the factor model which assesses misclassification in terms of false positives, false negatives, sensitivity, specificity, and bias in prevalence estimation. This method uses Monte Carlo simulation of factor values to generate individual responses on the criteria via the factor model. In this application, the factor model parameter values of Table 2 will be used to generate the criteria outcomes. Misclassification is assessed by comparing the classification based on the factor values (the simulated values, not estimated values) with that based on weighted and unweighted sums of criteria. The choice of cutpoints on the factors depends on assumptions of population prevalence of abuse and dependence, and the analyses will be performed under two alternative assumptions. Simulations will be performed for full-population situations as well as for sub-population situations. The subpopulation case is of interest in order to study the quality of prevalence estimation in high-risk groups. Here, a somewhat more homogeneous group than the full population is considered (factor variance of 0.5 instead of 1.0) and the factor mean of this subgroup is varied as 0.5, 1.0. 1.5, and 2.0 full-population standard deviations above the full-population factor mean. To obtain sufficient stability in the results, a sample size of 1000 will be used with 500 Monte Carlo replications.

3. Results

3.1. The two-dimensional model of Muthén et al. (1993a)

Following is a brief description of the NHIS88 and the two-dimensional model of Muthén et al. (1993a). The NHIS88 had a complex, multistage design which was both stratified and clustered with oversampling of blacks. It resulted in a national sample of 47 485 households, where for the alcohol supplement of the survey, one adult 18 years of age or older was randomly selected from each household and 43 809 individuals responded for an overall response rate in the alcohol supplement of 85.5%. Of these individuals, 22 102 were classified as current drinkers based on their alcohol behavior in the last 12 months. In their study of subgroup differences, Muthén et al. (1993b) found a somewhat different factor structure for the 18 244 whites in the sample than for the blacks and this factor solution will be the basis for the current analyses.

The 11 criteria of the two-dimensional model and their acronyms are given in Table 1. In the NHIS88, a diagnostic criterion was considered fulfilled if at least one of its symptoms was experienced at least twice in the last year. In this way, each of the 11 criteria is a binary variable scored 0 or 1, with 1 denoting that the criterion is fulfilled.

The data on the criteria for the 18 244 white, current drinkers were factor analyzed by methods for binary variables using the commercially available LISCOMP computer program (Muthén, 1987) and a weighted least-squares estimator for tetrachoric correlations. A one-factor model was clearly rejected by the data (chi-square of 3421 with 44 degrees of freedom) while a two-factor model was deemed to fit the data sufficiently well given the huge sample size (chi-square of 194 with 34 degrees of freedom); for a discussion of fit assessment with these criteria, see Muthén et al. (1993a). The two-factor model estimated from the NHIS88 data for the 18 244 white, current drinkers is given in Table 2.

The first factor was interpreted as a dimension of alcohol abuse and was found to be measured well by

Table 1
Diagnostic criteria for DSM-III-R and DSM-IV alcohol abuse and dependence and associated questionnaire items

Analysis acronym	Diagnostic criterion	Questionnaire items	Proportion*
LARGER	Drinking in larger amounts or over a longer period than the person intended.	● Ended up drinking much more than you intended to.	0.25
		● Found it difficult to stop once you started.	0.08
		● Kept on drinking for a longer period of time than you intended to.	0.13
CUTDOWN	Persistent desire or one or more unsuccessful efforts to cut down or control drinking.	● Tried to cut down or stop drinking and found you couldn't do it.	0.02
		● Wanted to cut down or stop drinking and found you couldn't do it.	0.01
TIMESPENT	Spent a great deal oftime obtaining alcohol, drinking, or recovering from drinking.	● Spent a lot of time drinking or getting over the effects of drinking.	0.03
MAJOROLE	Frequent intoxication or withdrawal symptoms when expected to fulfill major role obligations at work, school, or home.	● Stayed away from home or gone to work late because of drinking or from a hangover.	0.03
		● Gotten drunk instead of doing the things you were supposed to do.	0.04
		● Been so hung over that it interfered with doing things you were supposed to do.	0.04
HAZARD	Recurrent drinking insituations in which itis physically hazardous.	● Driven a car after having too much to drink.	0.10
		● Done things when drinking that could have caused you to be hurt.	0.07
		● Done things when drinking that could have caused someone else to be hurt.	0.04
GIVEUP	Important social, occupational, or recreational activities given up or reduced because of drinking.	● Given up or cut down on activities or interests like sports or associations with friends, in order to drink.	0.01
		● Lost ties with or drifted apart from a family member or friend because of your drinking.	0.01
CONTINUE	Continued to drink despite knowledge of a persistent or recurrent social, psychological, or physical problem that is caused or exacerbated by drinking.	● Continued to drink alcohol even though it was a threat to your health.	0.02
		● Kept drinking even though it caused you emotional problems.	0.02
		● Kept drinking even though it caused you problems at home, work, or school.	0.02
		● Had a spouse or someone you lived with threaten to leave you because of your drinking.	0.01
TOLERANCE	Tolerance.	● Found that the same amount of alcohol had less effect than before.	0.09

the criteria LARGER, drinking more or longer than intended, and HAZARD, largely involving driving after drinking too much. The second factor was interpreted as alcohol dependence and was measured well by the criteria CUTDOWN, GIVEUP, and CONTINUE, corresponding to recognition that one cannot cut down or stop drinking, giving up or reducing activities in order to drink, and drinking despite recognition that drinking was causing problems in areas of functioning. The criteria LARGER and HAZARD are seen as good measurements of the abuse dimension due to their large factor loadings on the abuse factor and their small loadings on the dependence factor. Similarly, CUTDOWN, GIVEUP, and CONTINUE are good measurements of the dependence dimension. Many other criteria have sub-

Table 1 (contd.)

Analysis acronym	Diagnostic criterion	Questionnaire items	Proportion*
		• Found that you had to drink more than you once did to get the same effect.	0.04
WITHDRAWAL	Characteristic withdrawal symptoms.	• Been sick or vomited after drinking, or the morning after.	0.09
		• Felt depressed, irritable, or nervous after drinking or the morning after.	0.10
		• Heard or seen things that weren't really there after drinking, or the morning after.	0.01
		• Found yourself sweating heavily or shaking after drinking, or the morning after.	0.03
RELIEF	Drinking to relieve or avoid withdrawal symptoms.	• Taken a drink to keep yourself from shaking or feeling sick either after drinking or the morning after.	0.01
LEGAL	Recurrent alcohol-related legal or inter-personal problems.	• Been arrested or had trouble with the police because of your drinking.	0.01

*Proportion in the calibration sample of current drinkers (n = 11 086) admitting to having had this happen at least twice in the last 12 months.

stantial loadings on the two factors but discriminate less well between them. Their loadings were also found to be less stable in cross-validation analyses. The column with proportions indicates that the abuse dimension is measured by criteria that are more common than those of the dependence dimension. The factor correlation is 0.76. Although this indicates that the two dimensions are highly correlated, the model states that abuse and dependence do not define opposite ends of the same continuum, but rather define phenomena of a distinct kind.

The two-dimensional factor model represents a departure from previous diagnostic schemes related to DSM-III-R and the proposed DSM-IV. In DSM-III-R, a person is diagnosed as alcohol dependent if at least 3 of a subset of 9 criteria have been met and some symptoms fulfil the duration criterion. The 9 criteria, however, include the criteria LARGER and HAZARD measuring the abuse dimension in the factor model. Also, abuse is diagnosed based on HAZARD and CONTINUE, the latter being a dependence dimension measure. These diagnoses therefore blur the distinction between abuse and dependence as defined in the factor model. The proposed DSM-IV diagnoses suffer from similar problems in that the dependence diagnosis also includes HAZARD and the abuse diagnosis includes GIVEUP. The consequences for prevalence estimation in the population are important. Including HAZARD in the dependence diagnosis gives a much higher prevalence than without it and requiring CONTINUE or GIVEUP for an abuse diagnosis gives a much lower prevalence than without them.

3.2. Studying the distribution of the latent variable values

It is of interest to examine the distribution of estimated latent variable values (factor scores) to determine if there are any natural breaks in this distribution that would indicate a distinction between a minority of individuals scoring high and a majority of individuals scoring at lower values. As mentioned in the Introduction, this is useful both for evaluating the efficacy of established cutpoints derived from having fulfilled a certain number of criteria, as well as for choosing cutpoints.

The estimated factor scores for the response patterns observed in the NHIS88 sample of 18 244 white, current drinkers are presented in Fig. 1, with the 40 most common patterns being represented by large filled circles (the mean and variance of the scores in Fig. 1 are not the same as those for the estimates in Table 2). Out of 2048 possible response patterns for the 11 criteria, only 325 were realized in the sample. This indicates a large degree of 'severity' ordering among the criteria, such that less prevalent criteria are not typically fulfilled for an individual when more prevalent criteria are not fulfilled. Of these 325 patterns, the 40 most prevalent patterns accounted for 96% of the observations, where the 40th pattern was observed for only 0.05% of the sample, or 10 individuals. A total of 137 of the 325 patterns had a frequency of only one.

As expected, the most common patterns give factor scores mostly positioned in the lower left-hand corner of the figure. Individuals for whom none of the crite-

Table 2
Factor model for the NHIS88 white current drinkers (n = 18 244)

Criterion	Factor loadings		Prop.	Optimal weights	
	Abuse	Dependence		Abuse	Dependence
LARGER	0.89	0.02	0.27	2.09	0.05
CUTDOWN	0.03	0.87	0.02	0.07	1.93
TIMESPENT	0.48	0.47	0.03	1.06	1.03
MAJOROLE	0.66	0.30	0.07	1.58	0.72
HAZARD	0.85	0.05	0.13	1.85	0.11
GIVEUP	0.15	0.81	0.01	0.41	2.19
CONTINUE	0.06	0.87	0.04	0.15	2.18
TOLERANCE	0.43	0.40	0.04	0.68	0.64
WITHDRAWAL	0.39	0.53	0.02	0.78	1.05
RELIEF	0.21	0.72	0.01	0.46	1.58
LEGAL	0.31	0.50	0.01	0.48	0.78
Factor correlation	0.76				

ria are fulfilled have estimated scores of zero on both factors. For the most common patterns the most prevalent criterion LARGER is fulfilled, but not together with the other major measurements of the factors, these measures being HAZARD, CUTDOWN, GIVEUP, and CONTINUE. The two common patterns at the upper right-hand corner correspond to fulfilling all criteria and all criteria but LEGAL (25 and 18 individuals, respectively).

Consider the following hypothetical example of using cutpoints on the dependence and abuse factors. The DSM-III-R diagnosis of alcohol dependence is made when at least 3 out of 9 criteria are fulfilled. Extending this to the present set of 11 criteria, we may for example require that at least 5 of the criteria have to be fulfilled for a dependence diagnosis. This results in an NHIS88 population prevalence for white current drinkers of 2.8%. This translates into a cutpoint of 3.7 for the dependence factor. Furthermore, if a cutpoint of 3.9 is chosen for the abuse factor, 11% would be diagnosed as abusers. These two hypothetical cutpoints are further considered below. They give a horizontal and a vertical line in the scatter plot of Fig. 1, and these lines define four quadrants. The bottom left quadrant contains those who are not abusers or dependent, the bottom right quadrant contains those that are abusers, but not dependent, the upper right quadrant contains those who are both abusers and dependent, and the upper left quadrant contains those who are not abusers but dependent. With the cutpoints given, it is seen that all four quadrants would have individuals in them for this sample. For example, the most common pattern in the dependence-but-not-abuse quadrant is the 40th most common pattern (a large circle in Fig. 1 with Abuse value of about 2 and Dependence value of about 4). For this pattern, only the most prevalent abuse criterion LARGER is fulfilled, while two of the important dependence criteria CUTDOWN, and CONTINUE are fulfilled.

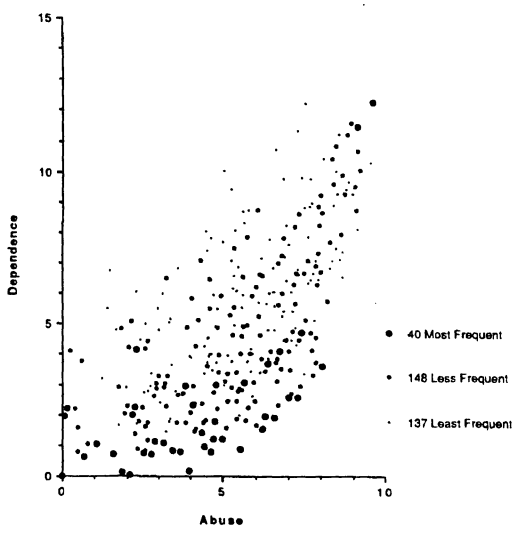

Fig. 1. Scores for the factors.

Alternatives to classifying individuals into the four quadrants are possible. For example, one might consider a classification along the 45° line that combines the two factors into one general dimension. In this way, individuals furthest out on this line, i.e. those in the upper right-hand corner, would have the most severe alcohol problems, while individuals not as far out on this line would have less severe problems and mostly abuse-related problems.

As seen in Fig. 1, there are no natural breaks in the distribution of points in the scatterplot to help in selecting cutpoints. As a result, there would be many individuals very close to any cutpoints, resulting in a case/non-case classification that is often based on a very small difference in the factor scores. This analysis outcome suggests using the above mentioned alternative to the case/non-case classification, namely describ-

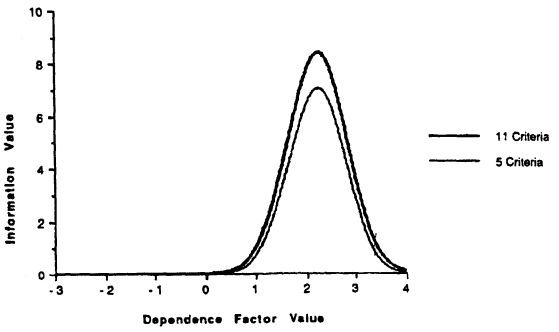

Fig. 2. Information curves for dependence factor.

ing an individual's status using the percentile corresponding to the continuous factor score.

3.3. Studying the precision of the latent variable measurement

Information curves and expected score functions indicate at which level of a latent variable the set of criteria gives good measurement precision. This psychometric analysis can shed light both on the efficacy of a given convention for using the criteria to make a diagnosis and on choosing a cutpoint.

3.3.1. Information curves

The information value in the set of 11 criteria is shown in Figs. 2 and 3 for the dependence and abuse factor, respectively. For the dependence factor, the information value peaks at a dependence factor value of 2.2. This says that the dependence factor is measured with peak precision at this factor value. If the aim of the analysis is to find a cutpoint, the value of 2.2 is therefore optimal from the point of view of information. Since the factor is given in a standard normal metric, this value corresponds to the 98.7th percentile of the dependence distribution. If a cutpoint of 2.2 is chosen, the percentage of individuals exceeding this cutpoint is therefore 1.3%. This information value result implies that the set of criteria is best at discriminating dependents from non-dependents if the population prevalence is 1.3%. Only half the information value is obtained at a factor value of about 1.5, corresponding to a prevalence of 7%.

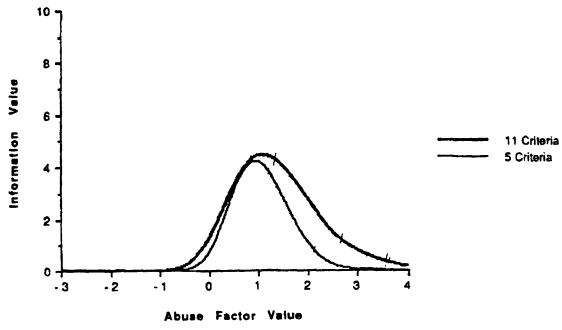

Fig. 3. Information curves for abuse factor.

3.3.2. Information for a reduced set of criteria

For comparison, an information curve is also given for a reduced set of 5 criteria, corresponding to the criteria that best define the two factors: LARGER, HAZARD, CUTDOWN, GIVEUP, CONTINUE. This addresses the question if essentially the same measurement information value could be obtained with a shorter instrument. Fig. 2 shows that the curves peak at about the same point and that rather little information is lost by using the 5 best instead of all 11 criteria. The difference in peak information values may be understood in terms of the standard error of an individual's factor score estimate, obtained as the square root of the inverted information value. The peak value for 11 criteria is 8.4, giving a standard error of 0.35, while the 5-criterion peak value is 7.0 giving a standard error of 0.38. Since the factor has a standard deviation of 1, these values are both slightly higher than 1/3 of a standard deviation and are for practical purposes almost the same.

Fig. 3 shows the corresponding information curves for the abuse factor. Relative to dependence, much less information is available on this factor. This is because the factor has fewer criteria measuring it well. The peak information value is at about 1.0 corresponding to a cutpoint with a population prevalence of 16%. Compared to the dependence curves, the information value for abuse does not drop as drastically when moving away from the value for the peak. For example, the information value is largely the same for an abuse value of 1.5, corresponding to a prevalence of 7%. The 5-criterion information value peaks at about the same point but drops more quickly for higher factor values.

3.3.3. Information for an unweighted sum of criteria

Given that the DSM-III-R diagnosis of dependence is made based on an unweighted sum of criteria, it is of interest to study the information content in such an approach. To this aim, Fig. 4 gives the information about the dependence factor for an unweighted and an optimally weighted sum of the 11 criteria. Note from Table 2 that the weighted sum for the dependence factor essentially excludes the abuse criteria LARGER and HAZARD. It should be pointed out that the information scale in this figure is not directly comparable to that of the previous figures, see the Methods section. To make this comparison, Fig. 4 also gives the information function for ordinary factor scores and it is seen that it is practically the same as that of the optimally weighted sum. Furthermore, it is interesting

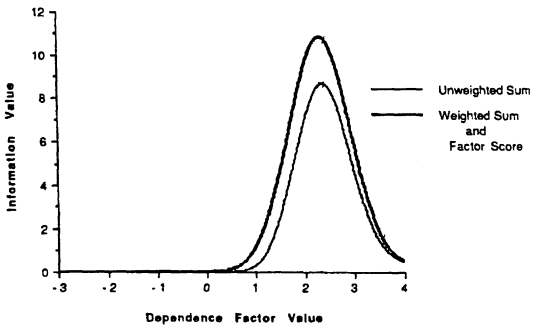

Fig. 4. Information curves for dependence factor using sums of criteria.

to note that using the unweighted sum does not give a dramatic drop in information.

3.3.4. Expected score for dependence

Another way of assessing the quality of how the set of criteria measure the factors is shown in Fig. 5. The expected score, ranging from 0 to 11, is here plotted as a function of the dependence factor value. The steepness of the curve indicates how well the score can discriminate between individuals with different factor values. For example, using the weighted score, a person with a factor value of 1.5, i.e. being in the 93rd percentile, has an expected score of about 1, while a person with a factor value of 2.5, being in the 99th percentile, has an expected score of about 8. The large change in expected score indicates a good quality in how the set of criteria measure the factors and this is a function of there being sufficient numbers of criteria with high loadings on the dependence factor. As was the case for the information curves shown earlier, Fig. 5 shows that the best discrimination can be made for factor scores just above 2.

Fig. 5 also gives another way of comparing the unweighted and weighted approach to scoring dependence. The unweighted sum is only slightly less discriminating than the weighted sum and the conclusion is that the simple, unweighted approach is not sacrificing much measurement quality.

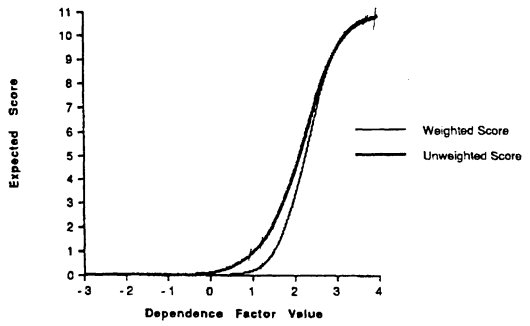

Fig. 5. Expected score for dependence: unweighted and weighted case.

The comparison of information curves and expected score functions for the unweighted and weighted score for the abuse factor will not be presented here in order to save space.

3.4. Studying the sensitivity and specificity of diagnosis

The measurement properties of a given diagnostic approach can be assessed by means of the percentage of false positives and false negatives involved in the diagnostic sensitivity (proportion of cases diagnosed correctly) and specificity (proportion of non-cases not diagnosed). The sensitivity and specificity values can also be used to choose a factor score cutpoint when that is the aim of the analysis.

3.4.1. Misdiagnosis: dependence

To study diagnosis errors, an assumption about the population prevalence of dependence has to be made. As was done above when studying the latent variable distribution, we may for example require that at least 5 of the 11 criteria have to be fulfilled for a dependence diagnosis, resulting in an NHIS88 population prevalence for white current drinkers of 2.8%. This prevalence value can be translated into the standard normal cutpoint value of 1.91 to be used in the factor score and criterion simulation method. Considering the unweighted sum of the 11 criteria and the requirement of at least 5 criteria, a comparison with the correct factor score diagnosis resulted in the misdiagnosis statistics: false positives 1%, false negatives 1%, sensitivity 67%, specificity 99%. The low degree of sensitivity is an indication of a serious problem in how the set of criteria measure the factors: only about 2/3 of the cases are diagnosed.

The use of the unweighted sum was compared to using the optimally weighted sum. The weighted sum, however, resulted in rather similar misdiagnosis statistics: false positives 1%, false negatives 1%, sensitivity 72%, specificity 99%. There is a slight improvement in sensitivity, but it does not seem important.

To check the dependence of these results on the assumed population prevalence, a value of 4.6% was also used, corresponding to fulfilling at least 4 out of the 11 criteria. The results were very similar to those for the 2.8% prevalence. Furthermore, working with simplified weights so that only CUTDOWN, GIVEUP, and CONTINUE are used with unit weights and requiring at least 2 of these 3 criteria gave a prevalence of 1.6% and again very similar misdiagnosis statistics. The conclusions about the quality of the diagnosis therefore do not seem strongly dependent on prevalence assumptions, nor on the weighting scheme, at least not when including the criteria that measure the factors well.

Table 3
Prevalence estimation in subgroups

Factor mean	Correct prevalence	Estimated prevalence	% False positives	% False negatives	Sensitivity	Specificity
0.0	0.3					
	Unweighted	0.4	0.2	0.1	59	100
	Weighted	0.4	0.2	0.1	62	100
0.5	2					
	Unweighted	2	1	0.4	70	99
	Weighted	2	1	0.4	71	99
1.0	5					
	Unweighted	8	4	1	80	96
	Weighted	7	3	1	78	97
1.5	14					
	Unweighted	21	9	2	88	89
	Weighted	18	6	2	85	93
2.0	29					
	Unweighted	42	15	2	94	79
	Weighted	36	10	3	90	86

3.4.2. Misdiagnosis: abuse

In the study of abuse misdiagnosis, three weighting schemes were considered, using unit weights for all 11 criteria, unit weights for only LARGER and HAZARD, and optimal weights. The alternative population prevalences 7.8% (at least 3 criteria fulfilled) and 11.1% (at least 2 criteria) were tried. Again, the results were very similar with 2–3% false positives and false negatives, 71–76% sensitivity, and 96–97% specificity. Optimal weighting gave the highest value of sensitivity but the difference does not seem large.

3.4.3. Misdiagnosis: abuse and dependence

In line with the discussion of Fig. 1, an individual may be diagnosed with abuse and dependence if he/she obtains factor scores falling in the upper right-hand quadrant of the bivariate distribution. Here, we consider misdiagnosis of a scheme which uses cutpoints on both axes and two optimally weighted sums, one for each of the two factors. This is compared to a simple scheme of using only one, unweighted sum of all 11 criteria. A prevalence of 1.8% was used corresponding to having at least 6 of the 11 criteria fulfilled. The misdiagnosis statistics were again very similar with about 1% false positives and negatives, 71–72% sensitivity, and 99–100% specificity. The fact that the use of two, optimally weighted scores does not improve the diagnosis is another indication of the strong correlation between the two factors.

3.4.4. Misdiagnosis: abuse but not dependence

One may also consider individuals in the lower right-hand quadrant of Fig. 1, diagnosed as abusers, but not dependent. Three schemes are considered. The first one uses unit weights and uses LARGER and HAZARD for abuse and the other 9 criteria for dependence. Both criteria need to be fulfilled for abuse and less than 3 are allowed to be fulfilled for dependence, giving a prevalence of 8.5%. The second scheme uses the same prevalence and considers the two optimally weighted sums, using two cutpoints. The third scheme uses the simple approach of the unweighted sum of all 11 criteria, requiring that at least 3 but less than 5 criteria are fulfilled with a prevalence of 5.0%. The three schemes gave similar results with false positives and false negatives in the 3–5% range, sensitivity in the 50–60% range, and specificity in the 95–97% range. This more detailed diagnosis apparently has very low sensitivity. The lowest value of 50% was obtained for the third and simplest scheme, while the highest value of 60% was obtained for the optimally weighted scheme.

3.4.5. Misdiagnosis: prevalence estimation in subgroups

It is often of interest to compute prevalence estimates for subgroups of individuals at high risk for abuse or dependence and to compare those values to those of more average groups. To study how well the set of 11 criteria accomplishes this, misdiagnosis is studied by the simulation method for subgroups that are more homogeneous than the general population and have higher factor means. In this approach it is also possible to study prevalence misestimation by applying to the subpopulations the cutpoints set for the full population (in the full population this bias is non-existent by definition).

Table 3 gives the results from this analysis letting the dependence factor mean increase from the population

mean of 0.0 to 2.0 standard deviations above this mean. Results are given for both the unweighted sum of the 11 criteria and the optimally weighted sum.

Table 3 shows that for subgroups with means close to those of the general population (mean less than or equal to 0.5), the bias in the prevalence estimates and the percentage false positives and false negatives are rather low. While specificity is very high, the sensitivity is low, only 60–70%.

For more extreme subgroups, there is an increasing amount of overestimation of prevalence. The weighted sum has smaller bias than the unweighted sum. For example, for the group with mean of 1.5, the correct prevalence of 14% is overestimated as 21% by the unweighted method while the weighted method overestimates it as 18%. Sensitivity increases considerably for groups with higher mean and is actually higher for the unweighted method than the weighted one. For specificity, however, the situation is the reverse. In summary, the new information for how the set of criteria measure the factors is given in Table 3 is that while sensitivity and specificity may be reasonably improved in high-risk subgroups, prevalence estimation may be seriously biased. Similar findings were obtained for abuse, except that the prevalence was underestimated instead of overestimated.

4. Discussion

The discussion section consists of two parts. First, the results of applying the methodology will be summarized. The discussion will then focus on how the results can be used to improve diagnosis and prevalence estimation in this application. This discussion illustrates how psychometric investigations of diagnostic criteria can help in the refinement of psychiatric instruments.

4.1. Summary

Given how the set of 11 criteria measure the factors, the ability to discriminate between individuals in the population who are and are not dependent is found to be the highest if the population prevalence is around 1.3%, corresponding to a dependence factor score cut-point in the 98.7th percentile. For dependence factor values much different than that, the information value drops dramatically. For example, for a factor value in the 93rd percentile (prevalence of 7%), only half as much information is available. For abuse, the information value is considerably lower due to the smaller number of criteria with good measurement properties. It is the highest for a prevalence of 16%, but does not diminish significantly even for half of that prevalence.

For both dependence and abuse diagnosis, the sensitivity, i.e. the percentage of cases diagnosed as such, is found to be low, around 70%. A diagnosis of a person as being an abuser but not dependent is found to have particularly low sensitivity, around 50–60%.

A large degree of bias in prevalence estimation is found for subgroups of the population at high risk for dependence or abuse. The direction of the bias is different for dependence and abuse.

Using weights based on the factor model instead of simple, unweighted sums of the criteria makes only marginal improvements in information value and sensitivity. The most important improvement is observed for subgroup prevalence estimation, although at the expense of sensitivity. Using a reduced set of the 5 best criteria in the factor model gives similar results to using all 11 criteria.

To summarize, the weaknesses in how the 11 criteria measure the factors' are the low sensitivity of diagnosis and the high bias in subgroup prevalence estimation. These weaknesses exist despite the fact that the set of criteria has rather high information values, i.e. precision of measurement of the abuse and dependence dimensions, for prevalence values that appear realistic. Refinements of optimal weighting do not significantly improve this situation. In fact, unweighted scores based on a subset of criteria that have been found to have good measurement properties in the factor model are almost as efficient.

4.2. Conclusions

The factor model gives guidance about which criteria to emphasize in abuse and dependence diagnoses. The investigation of its measurement properties, however, points to strengths and weaknesses which are seemingly contradictory. On one hand, the information values indicate a rather high ability to discriminate between individuals with different factor scores. On the other hand, the sensitivity is low and the bias in prevalence estimation is high. One explanation can be seen in Fig. 1 which describes the scatterplot for the estimated factor scores. As seen there, the choice of cut-points is difficult because there are no natural breaks in scatterplot of factor scores. Therefore, misclassification can easily occur even for scores that are estimated with rather high precision.

There are ways one could avoid these weaknesses. Three possibilities are through the use of different criteria, through the use of different psychometric techniques, and through the use of continuous rather than dichotomous diagnoses.

It is possible that the use of a greater number of more specific criteria will improve not only the information values of the set of criteria over a wider range of the factor score values, but also sensitivity and prevalence estimation. For example, more symptom items

may be included and, in addition, responses to individual items rather than combinations of items may be used. Given this possibility, it is interesting to note that the opposite approach is taken in the determination of the criteria published in the International Classification of Diseases, Tenth Revision (ICD-10) (World Health Organization, 1992) and in recent decisions on DSM-IV criteria. In both situations, several complex criteria (e.g. WITHDRAWAL/RELIEF, GIVEUP/TIMESPENT) have been combined.

Different psychometric techniques may also be used such as factor score estimation by maximum likelihood or Bayesian methods. New psychometric techniques may be drawn upon to utilize auxiliary information in the diagnosis, such as alcohol consumption, family history of alcoholism, and sociodemographic characteristics of the individual. Multiple factors can then be better estimated. This suggestion is consistent with the growing awareness in psychiatry of the need for multi-axial approaches to classification of alcohol use disorders.

For any of these approaches, the difficulty of classifying an individual exhibited by the low sensitivity and high prevalence bias could be circumvented by using percentiles corresponding to factor scores to evaluate an individual's alcohol abuse and dependence status rather than using the traditional dichotomous diagnoses.

Acknowledgements

The research was supported by grant AA 08651-01 from NIAAA for the project 'Psychometric Advances for Alcohol and Depression Studies.' I would like to thank Bridget Grant, Robert Mislevy and Hariharan Swaminathan for comments on an earlier draft. Jin-Wen Hsu, Ginger Nelson, and Li-Chiao Huang provided helpful research assistance.

References

American Psychiatric Association. (1987) Diagnostic and Statistical Manual of Mental Disorders, 3rd edition, revised (DSM-III-R). American Psychiatric Association, Washington, DC.

American Psychiatric Association. (1992) DSM-IV Options Book. American Psychiatric Association, Washington, DC.

Bock, R.D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. Psychometrika 46, 443–459.

Duncan-Jones, P., Grayson, D.A. and Moran, P.A.P. (1986) The utility of latent trait models in psychiatric epidemiology. Psychol. Med. 16, 391–405.

Hambleton, R.K. and Swaminathan, H. (1985) Item Response Theory. Principles and Applications. Kluwer-Nijhoff, Boston.

Muthén, B. (1978) Contributions to factor analysis of dichotomous variables. Psychometrika 43, 551–560.

Muthén, B. (1987) LISCOMP. Analysis of Linear Structural Equations with a Comprehensive Measurement Model. Theoretical integration and user's guide. Scientific Software, Mooresville, IN.

Muthén, B. (1989) Dichotomous factor analysis of symptom data. In: Latent Variable Models for Dichotomous Outcomes: Analysis of Data from the Epidemiological Catchment Area Program, (Eaton and Bohrnstedt, eds.). Sociological Methods and Research, Vol. 18, pp. 19–65.

Muthén, B. (1993) Covariates of alcohol dependence and abuse: A multivariate analysis of a 1988 general population survey in the United States. Acta Psychiatr. Scand. (in press).

Muthén, B. (1995) Factor analysis of alcohol abuse and dependence symptom items in the 1988 National Health Interview survey. Addiction 90, 637–645.

Muthén, B., Grant, B. and Hasin, D. (1993a) The dimensionality of alcohol abuse and dependence: Factor analysis of DSM-III-R and proposed DSM-IV criteria in the 1988 National Health Interview Survey. Addiction 8, 1079–1090.

Muthén, B., Grant, B. and Hasin, D. (1993b) Subgroup differences in factor structure for DSM-III-R and proposed DSM-IV criteria for alcohol abuse and dependence in the 1988 National Health Interview Survey. J. Nerv. Ment. Dis. (in press).

Muthén, B., Hasin, D. and Wisnicki, K.S. (1993) Factor analysis of ICD-10 symptom items in the 1988 National Health Interview Survey on Alcohol Dependence. Addiction 7, 1071–1077.

World Health Organization. (1992) International Classification of Diseases — Tenth Revision: Clinical Descriptions and Diagnostic Guidelines. World Health Organization, Geneva.