PSYCHOMETRIKA—VOL. 64, NO. 1, 9–24
MARCH 1999

# INDCLAS: A THREE-WAY HIERARCHICAL CLASSES MODEL

IWIN LEENEN, IVEN VAN MECHELEN AND PAUL DE BOECK

KATHOLIEKE UNIVERSITEIT LEUVEN

SEYMOUR ROSENBERG

RUTGERS UNIVERSITY

A three-way three-mode extension of De Boeck and Rosenberg's (1988) two-way two-mode hierarchical classes model is presented for the analysis of individual differences in binary object × attribute arrays. In line with the two-way hierarchical classes model, the three-way extension represents both the association relation among the three modes and the set-theoretical relations among the elements of each mode. An algorithm for fitting the model is presented and evaluated in a simulation study. The model is illustrated with data on psychiatric diagnosis. Finally, the relation between the model and extant models for three-way data is discussed.

Key words: binary data, clustering, hierarchical classes, individual differences, set-theoretical relations, three-way three-mode data.

## 1. Introduction

Hierarchical classes models (De Boeck & Rosenberg, 1988; Van Mechelen, De Boeck & Rosenberg, 1995) have been developed to study structural relations in a two-way two-mode array with binary cell entries. Depending on the application, the first mode may refer to objects, situations, persons, and so on; the second mode may refer to attributes, responses, items, and so on; and the binary relation defined by the two-way two-mode array, may be "has" (e.g., if the modes are objects and attributes), "elicits" (if the modes are situations and responses), "solves" (if the modes are persons and items), and so on. In the present paper a three-way three-mode extension of the hierarchical classes model is introduced. The third mode may refer to judges (according to whom object $i$ either has or does not have attribute $j$), individuals (for whom situation $i$ either elicits or does not elicit response $j$), ages (at which child $i$ either solves or does not solve item $j$), and so on. In this paper, we will refer to elements of the first mode as objects, to elements of the second mode as attributes, and to elements of the third mode as sources.

The development of the three-way three-mode extension was motivated by the need to model individual differences in the use of a common consensual hierarchical classes structure. In line with the two-way model, the three-way extension also represents the set-theoretical relations among the elements of each of the three modes. The new model is called the INDCLAS (*in*dividual *d*ifferences hierarchical *clas*ses) model.

In section 2, the theory of the INDCLAS model is outlined. In section 3 an associated algorithm is presented. In section 4, the algorithm is evaluated in a simulation study. Section 5 presents an application to real data. Section 6 discusses the relation of the new model with extant models for three-way data.

The research reported in this paper was partially supported by NATO (Grant CRG.921321 to Iven Van Mechelen and Seymour Rosenberg).

Requests for reprints should be sent to Iwin Leenen, Department of Psychology, Tiensestraat 102, B-3000 Leuven, BELGIUM. E-mail: Iwin.Leenen@psy.kuleuven.ac.be

0033-3123/1999-1/1997-0549-A $00.75/0

© 1999 The Psychometric Society

## 2. Theory

Several variants of the two-way two-mode hierarchical classes model have been distinguished (Van Mechelen et al., 1995). This paper takes the original disjunctive hierarchical classes (HICLAS) model as proposed by De Boeck and Rosenberg (1988) as a starting point (which in the remainder of this paper will be referred to as the disjunctive HICLAS model). For the reader's convenience, we will first briefly recapitulate this model. Next, we will introduce the new three-way INDCLAS model. Finally, the variants of the HICLAS and INDCLAS models will be discussed.

### *De Boeck and Rosenberg's (1988) HICLAS Model*

A HICLAS model approximates an $n_1$ (objects) $\times$ $n_2$ (attributes) binary data matrix $D$ by an $n_1 \times n_2$ binary matrix $M$ which can be decomposed into an $n_1 \times r$ binary matrix $S$ and an $n_2 \times r$ binary matrix $P$, where $r$ denotes the rank of the model. $S$ (*resp.* $P$) is called the object (*resp.* attribute) bundle matrix: The rows of $S$ (*resp.* $P$) refer to the objects (*resp.* attributes) and the $r$ columns of $S$ (*resp.* $P$) define $r$ (possibly overlapping) binary clusters, called bundles, of objects (*resp.* attributes). The bundle pattern of an element (an object or an attribute) is the set of bundles it belongs to.

As a guiding example we use the hypothetical matrix $M$ of Table 1. The bundle matrices of a disjunctive HICLAS model for this matrix are presented in Table 2.

We now introduce some notation: For an element $x$ of either of the two modes, $M(x)$ denotes the set of elements from the other mode that $x$ is associated with in $M$. For example, for the element $e$ in Table 1, $M(e) = \{3, 5, 7\}$.

TABLE 1.
Hypothetical Two-Way Two-Mode Array

| | Attributes | | | | |
|---|---|---|---|---|---|
| Objects | *a* | *b* | *c* | *d* | *e* |
| 1 | 0 | 1 | 1 | 1 | 0 |
| 2 | 0 | 1 | 1 | 1 | 0 |
| 3 | 0 | 1 | 1 | 0 | 1 |
| 4 | 1 | 0 | 0 | 1 | 0 |
| 5 | 0 | 1 | 1 | 0 | 1 |
| 6 | 0 | 0 | 0 | 0 | 0 |
| 7 | 1 | 1 | 1 | 1 | 1 |

TABLE 2.
Disjunctive Hierarchical Classes Model for the Data in Table 1

| | Object Bundles | | | | Attribute Bundles | | |
|---|---|---|---|---|---|---|---|
| Objects | I | II | III | Attributes | I | II | III |
| 1 | 0 | 1 | 0 | *a* | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 | *b* | 0 | 1 | 1 |
| 3 | 0 | 0 | 1 | *c* | 0 | 1 | 1 |
| 4 | 1 | 0 | 0 | *d* | 1 | 1 | 0 |
| 5 | 0 | 0 | 1 | *e* | 0 | 0 | 1 |
| 6 | 0 | 0 | 0 | | | | |
| 7 | 1 | 1 | 1 | | | | |

A HICLAS model represents three types of structural relations in $M$:

*Association.* The association relation is the binary relation between the set of objects and the set of attributes as defined by the 1-entries of the array $M$. The disjunctive HICLAS model represents the association relation by the following association rule:

$$m_{ij} = \bigoplus_{b=1}^{r} s_{ib} p_{jb} \qquad (\forall i = 1..n_1,\ \forall j = 1..n_2), \tag{1}$$

where $\oplus$ denotes the Boolean sum. For example, from the model in Table 2, it can be derived that Object 3 is associated with Attribute $b$ because both elements belong to Bundle III.

*Equivalence.* An equivalence relation is defined among the elements of each mode. Two elements $x$ and $y$ are equivalent iff $M(x) = M(y)$. Sets of equivalent objects (attributes) are called object (attribute) classes. In the disjunctive HICLAS model, equivalent elements have identical bundle patterns. In the example in Table 1, Objects 3 and 5 are associated with the same set of attributes, and, hence, those objects are equivalent and have identical bundle patterns in the HICLAS model of Table 2. On the attribute side, $b$ and $c$ are equivalent elements.

*Hierarchy.* Hierarchical relations are defined among the elements (or classes) of each mode. An element (class) $x$ is hierarchically below another element (class) $y$ iff $M(x)$ is a proper subset of $M(y)$. In the disjunctive HICLAS model hierarchical relations are represented in the corresponding bundle matrix in that $x$ is hierarchically below $y$ iff the bundle pattern of $x$ is a subset of the bundle pattern of $y$. For example, in Table 1, Object 1 is hierarchically below Object 7; consequently, the bundle patterns of Objects 1 and 7 are in a subset-superset relation in the HICLAS model of Table 2. Equivalence and hierarchy relations will be called set-theoretical relations. A decomposition of $M$ that represents the set-theoretical relations in $M$ is said to be set-theoretically consistent (with respect to $M$).

Any matrix $M$ can be decomposed into an $S$ and a $P$ in line with (1), the decomposition in a disjunctive HICLAS model being such that $r$ is minimal. This minimal number $r$ is called the Schein rank of $M$. Kim gives an equivalent definition of the Schein rank of a Boolean matrix $X$ as the minimum number of cross-vectors whose Boolean sum is $X$, a cross-vector (Kim, 1982) being a matrix $T$ for which $t_{ij} = a_i b_j$ with $a_i$ and $b_j$ either 0 or 1 (for all $i$ and $j$). Note that the Schein rank of a Boolean matrix can be shown to be smaller than or equal to its row rank and its column rank. As illustrated by De Boeck and Rosenberg (1988), decompositions (1) of $M$ into an $S$ and $P$ exist that do not represent the set-theoretical relations in $M$. De Boeck and Rosenberg proved that for any such decomposition, another decomposition in the same rank can be found that does represent both the association and set-theoretical relations in $M$, though. Van Mechelen et al. (1995) further proved that a sufficient condition for a set-theoretical decomposition to be unique (upon a permutation of the bundles) is the existence of an $r \times r$ identity submatrix in $M$, or, equivalently, each bundle specific class (i.e., each class of elements belonging to a single bundle) in both bundle matrices being nonempty.

De Boeck and Rosenberg (1988) also proposed a comprehensive graphic representation of the structural relations in $M$ as represented in a disjunctive HICLAS model. For example, Figure 1 is a graphic representation of the HICLAS model of Table 2. The objects are drawn in the upper half and the attributes in the lower half of the representation. The association relation can be read as follows: An object is associated with an attribute iff they are connected with each other by a downward path of lines and a zigzag. Equivalent elements are enclosed by a box, which represents a class, and hierarchical relations between classes are indicated by straight lines between the respective boxes. Note that the attribute hierarchy is to be read upside down.

If a bundle pattern does not apply to any element, the corresponding class is empty. Empty classes are most often omitted from the graphic representation, unless the class is at the bottom of the hierarchy. The number of bottom classes equals the rank of the model. Elements that do not belong to any bundle, as Object 6, are drawn in an isolated box that is not connected to the remainder of the structure. It may be noted that Table 2 and, consequently, the original array $M$ can be completely reconstructed from Figure 1.

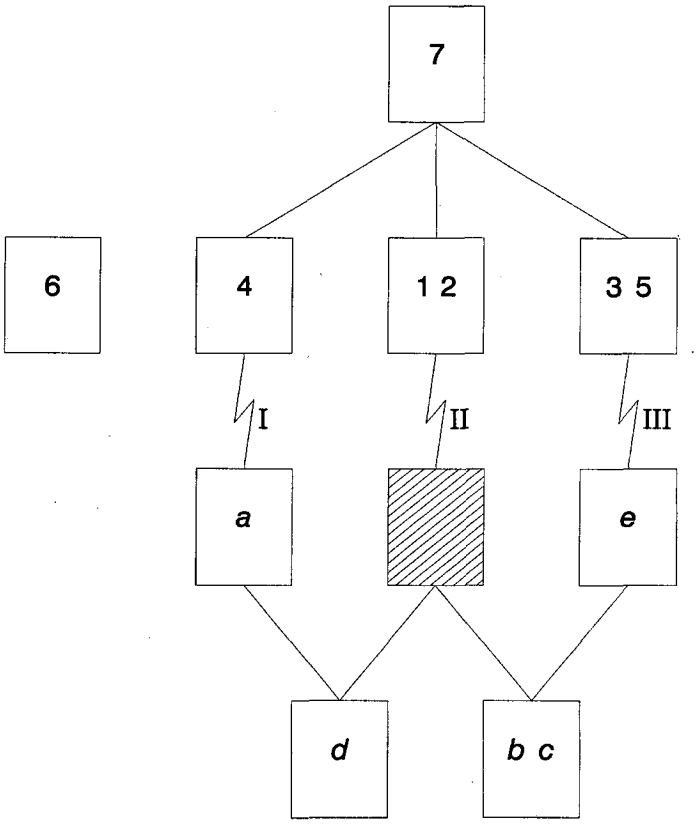



FIGURE 1.
Graphic representation of the hierarchical classes model in Table 2.

*A Three-way Three-mode Hierarchical Classes model: INDCLAS*

INDCLAS models imply a decomposition of an $n_1$ (objects) $\times$ $n_2$ (attributes) $\times$ $n_3$ (sources) binary array $M$ into an $n_1 \times r$ binary object bundle matrix $S$, an $n_2 \times r$ binary attribute bundle matrix $P$ and an $n_3 \times r$ binary source bundle matrix $I$, with $r$ the rank of the model.

The hypothetical array $M$ in Table 3 serves as the guiding example in this section. Table 4 presents the bundle matrices of a disjunctive INDCLAS model for these data. Extending the notation of the first subsection, $M(x)$ denotes the set of couples (of elements of the two other modes)

TABLE 3.
Hypothetical Three-Way Three-Mode Array

| Source *A* | | | | | | Source *B* | | | | | | Source *C* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Attributes | | | | | | Attributes | | | | | | Attributes | | | | |
| Objects | *a* | *b* | *c* | *d* | *e* | Objects | *a* | *b* | *c* | *d* | *e* | Objects | *a* | *b* | *c* | *d* | *e* |
| 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| 2 | 0 | 1 | 1 | 1 | 0 | 2 | 1 | 1 | 1 | 1 | 0 | 2 | 1 | 1 | 1 | 1 | 0 |
| 3 | 0 | 1 | 1 | 1 | 0 | 3 | 1 | 1 | 1 | 1 | 0 | 3 | 1 | 1 | 1 | 1 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 4 | 1 | 0 | 0 | 1 | 0 | 4 | 1 | 0 | 0 | 1 | 0 |
| 5 | 0 | 1 | 1 | 1 | 1 | 5 | 0 | 1 | 1 | 1 | 1 | 5 | 0 | 1 | 1 | 1 | 0 |
| 6 | 0 | 0 | 0 | 1 | 1 | 6 | 0 | 0 | 0 | 1 | 1 | 6 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 |

TABLE 4.
Disjunctive INDCLAS Model for the Data in Table 3

| | Object Bundles | | | | Attribute Bundles | | | | Source Bundles | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Objects | I | II | III | Attributes | I | II | III | Sources | I | II | III |
| 1 | 0 | 1 | 0 | *a* | 1 | 0 | 0 | *A* | 0 | 1 | 1 |
| 2 | 1 | 1 | 0 | *b* | 0 | 1 | 0 | *B* | 1 | 1 | 1 |
| 3 | 1 | 1 | 0 | *c* | 0 | 1 | 0 | *C* | 1 | 1 | 0 |
| 4 | 1 | 0 | 0 | *d* | 1 | 1 | 1 | | | | |
| 5 | 0 | 1 | 1 | *e* | 0 | 0 | 1 | | | | |
| 6 | 0 | 0 | 1 | | | | | | | | |
| 7 | 0 | 0 | 0 | | | | | | | | |

that $x$ is associated with in $M$. For example, for element $a$ in Table 3 $M(a) = \{(2, B), (3, B), (4, B), (2, C), (3, C), (4, C)\}$.

The association relation and the set-theoretical relations of equivalence and hierarchy are represented by an INDCLAS model as follows:

*Association.* The association relation is the ternary relation between the set of objects, the set of attributes and the set of sources as defined by the 1-entries of the array $M$. In the disjunctive INDCLAS model, this relation is represented by the matrices $S$, $P$ and $I$ by the following rule, which generalizes (1):

$$M_{hjk} = \bigoplus_{b=1}^{r} s_{hb} p_{jb} i_{kb} \qquad (\forall h = 1..n_1,\ \forall j = 1..n_2,\ \forall k = 1..n_3). \tag{2}$$

An entry $M_{hjk}$ equals 1 if and only if there is a bundle to which object $h$, attribute $j$ *and* source $k$ belong. In the example of Tables 3 and 4, Object 4 is associated with Attribute $a$ by Source $C$, because all three elements belong to Bundle I.

*Equivalence.* Two elements $x$ and $y$ of the same mode are equivalent iff $M(x)$ equals $M(y)$. Equivalent objects (*resp.* attributes, sources) constitute an object (*resp.* attribute, source) class. The disjunctive INDCLAS model represents the equivalence relations in that equivalent elements must have identical bundle patterns. For example, in Table 3, objects 2 and 3 are equivalent and have identical bundle patterns in the INDCLAS model of Table 4.

*Hierarchy.* An element $x$ is hierarchically below an element $y$ iff $M(x)$ is a proper subset of $M(y)$. A disjunctive INDCLAS model directly represents the hierarchical relations among elements in the respective bundle matrices in terms of subset-superset relations between bundle patterns. In Table 3, for example, Object 1 is hierarchically below Object 2 and Source $A$ is hierarchically below Source $B$, which is reflected in the bundle matrices of Table 4.

The association rule in (2) is equivalent with a decomposition of $M$ into $r$ three-way cross-vectors whose Boolean sum equals $M$, a three-way cross-vector being any three-way array $T$ with $t_{hjk} = a_h b_j c_k$ with $a_h$, $b_j$ and $c_k$ either 0 or 1 (for all $h$, $j$ and $k$). A disjunctive INDCLAS model implies a cross-vector decomposition of $M$ such that $r$ is minimal. Generalizing the notion of Schein rank to $n$-way arrays—similar to Kruskal's (1977, 1989) generalized rank of a (real) $n$-way array—this minimal rank $r$ will be called the (generalized) Schein rank of $M$.

Extending a proof of De Boeck and Rosenberg (1988), it further holds that for any $S$, $P$ and $I$ that are a decomposition of $M$ according to (2), an $S^*$, $P^*$ and $I^*$ of the same size exist that represent the set-theoretical relations in $M$. To prove this latter assertion, we define an $n_1 \times n_1$ hierarchy matrix $U$ with $u_{hh'} = 1$ if object $h$ is hierarchically above or equivalent to object $h'$, and $u_{hh'} = 0$ otherwise. Similarly, an $n_2 \times n_2$ hierarchy matrix $V$ and an $n_3 \times n_3$ hierarchy matrix $W$ are defined for the attributes and the sources, respectively. From the definition of $U$, $V$ and $W$, it follows (a) that $S^* = US$, $P^* = VP$ and $I^* = WI$ yield a decomposition (2) of $M$ and (b) that $S^*$, $P^*$ and $I^*$ represent the relations of hierarchy and equivalence in $M$.

Finally, generalizing the proof by Van Mechelen et al. (1995) on the uniqueness of a two-way HICLAS model, it holds that the existence of a $r \times r \times r$ superidentity subarray in $M$, or, equivalently, nonemptiness of the bundle specific classes in all three bundle matrices, is a sufficient (though not necessary) condition for an INDCLAS model of $M$ to be unique.

In line with the INDSCAL (Carroll & Chang, 1970) terminology, the hierarchical classes model for objects and attributes in INDCLAS can be considered a *group structure*, which is a consensual model for all sources. Figure 2 shows the graphic representation of the group structure for the INDCLAS model of Table 4; it may be read as a two-way HICLAS graphic representation. Furthermore, the hierarchical structure of the sources can be considered the *source structure*, similar to the *source space* in INDSCAL. Figure 3 graphically represents the source structure for the INDCLAS model of Table 4: Both Source $A$ and Source $C$ are hierarchically below Source $B$, whereas $A$ and $C$ themselves are not hierarchically related. This source structure implies that source $B$ considers any distinction made by source $A$ or by source $C$.

A unique feature of INDCLAS is that it may be given a single overall graphic representation that combines the group structure and the source structure. Figure 4 contains such a representation for the model of Table 4. Compared to the graphic representation of the group model, each zigzag now includes a box that contains the (classes of) sources that belong to the corresponding bundle. The association relation can be read as follows: An object $h$ is associated with attribute $j$ by source $k$ if and only if object $h$ and attribute $j$ are connected with each other by a path that goes via source $k$. For example, there exists a path from the object class $\{2, 3\}$ to the attribute class $\{a\}$ via $B$ and $C$ and hence, Objects 2 and 3 are associated with Attribute $a$ by Source $B$ and $C$. For Source $A$, however, there exists no such path and hence, 2 and 3 are not associated with $a$ by $A$.

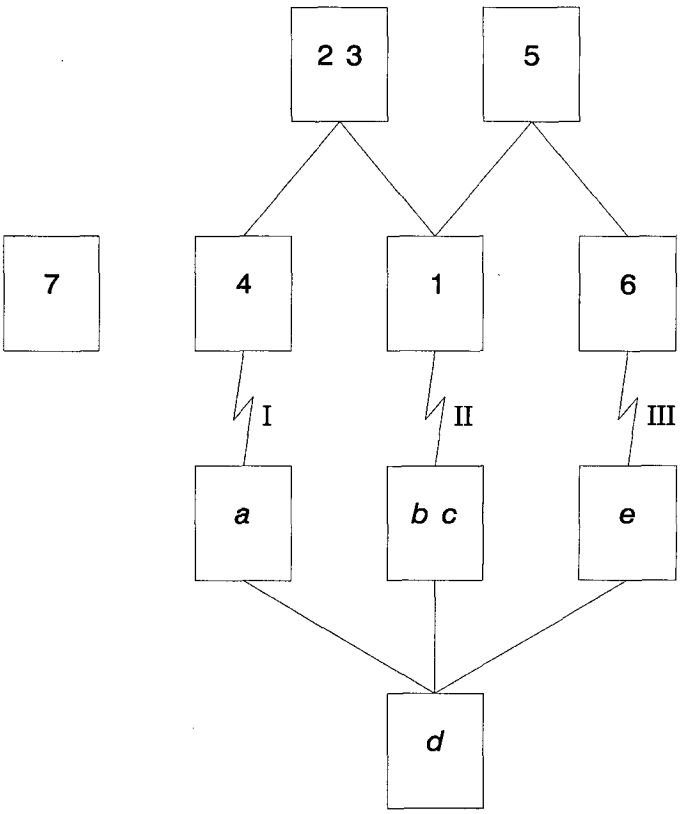



FIGURE 2.
Graphic representation of group structure for the INDCLAS model in Table 4.

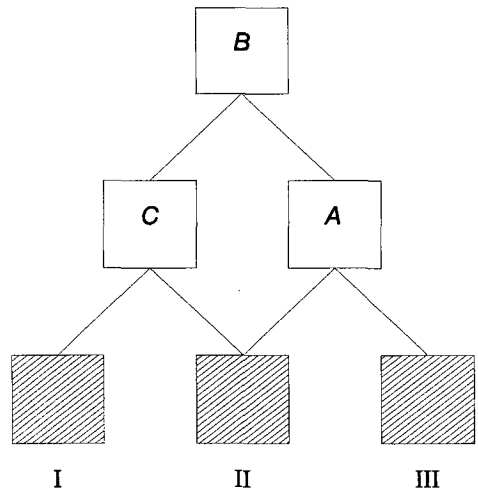



FIGURE 3.
Source structure of the INDCLAS model in Table 4.

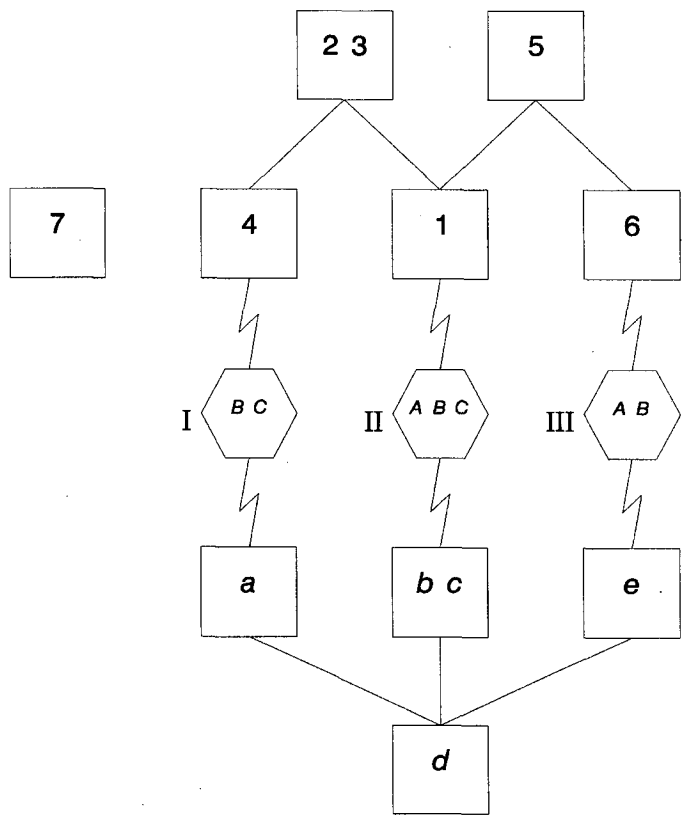



FIGURE 4.
Overall graphic representation of the INDCLAS model in Table 4.

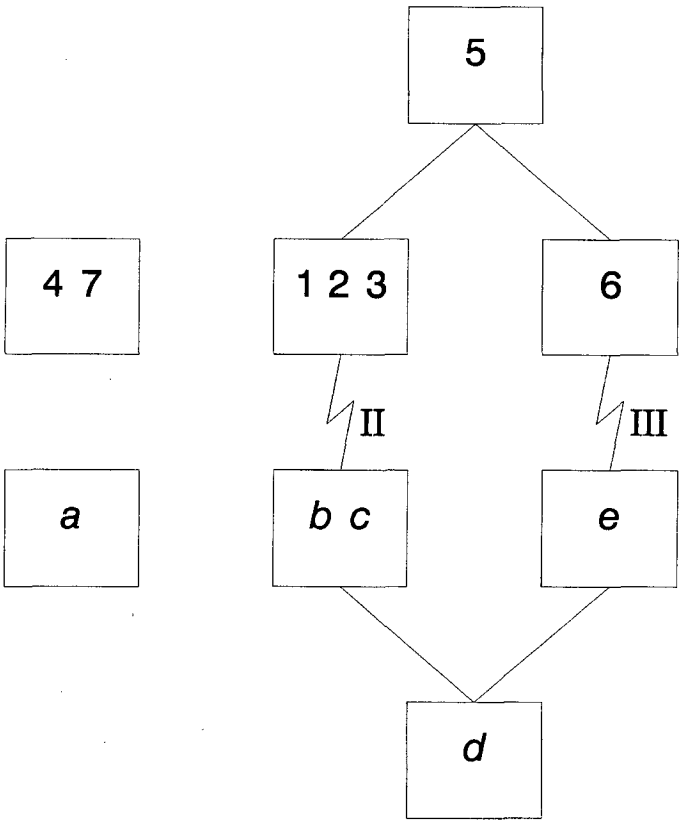



FIGURE 5.
Private structure of Source $A$ in the INDCLAS model in Table 4.

Finally, a *private structure* for each source can be derived. The private structure for source $x$ is the two-way HICLAS model that is obtained by only taking into account the bundles in the object and attribute bundle matrices Source $x$ belongs to. The private structure of Source $A$ is graphically represented in Figure 5.

*Variants*

Van Mechelen et al. (1995) proposed the conjunctive HICLAS model. Moreover, for both the disjunctive and conjunctive HICLAS models, Van Mechelen et al. distinguished four mathematically equivalent variants, each with a different psychological interpretation, though. In line with this, a conjunctive INDCLAS model can be defined and eight mathematically equivalent variants of both the disjunctive and the conjunctive INDCLAS models may be distinguished. As regards the association relation, the association rule for a conjunctive two-way HICLAS model is obtained by replacing $M$ in (1) by its Boolean complement; for the four variants, the association rules are obtained by either replacing or not replacing $S$ and $P$ in (1) by their Boolean complement. Similarly, the association rule for conjunctive INDCLAS can be obtained by replacing the array $M$ in (2) by its Boolean complement. Furthermore, by taking complements of either of the matrices $S$, $P$ and $I$, the association rules of the eight disjunctive and eight conjunctive variants of the INDCLAS model are obtained. All these variants must be considered reparametrizations, that is, different types of INDCLAS models, rather than indicating lack of identifiability for a single type of INDCLAS model. These different types of models, although being transformable into another, may imply different psychological interpretations. The choice between them should depend on substantive considerations only.

## 3. The INDCLAS Algorithm

The aim of a rank $r$ INDCLAS analysis of an $n_1 \times n_2 \times n_3$ binary data array $D$ is to look for an $n_1 \times n_2 \times n_3$ model array $M$ which is such that

$$\sum_{h=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}(d_{hjk} - m_{hjk})^2 \tag{3}$$

is minimal, subject to the constraint that $M$ can be decomposed into $n_1 \times r$, $n_2 \times r$, $n_3 \times r$ bundle matrices $S$, $P$ and $I$ according to (2), with $S$, $P$ and $I$ representing the relations of equivalence and hierarchy in $M$. In this section we propose an algorithm for INDCLAS analysis that generalizes De Boeck's (1986) two-way two-mode HICLAS algorithm.

The algorithm consists of two main routines: In the first routine, the algorithm searches for bundle matrices $S$, $P$ and $I$, which yield by means of (2) an array $M$ such that (3) is minimal. In the second routine, the bundle matrices are transformed such as to become consistent with the set-theoretical relations in $M$. We now describe each of these two routines in more detail.

The first routine can be considered an alternating *elementary binary discrete least squares procedure* (Chaturvedi & Carroll, 1994). It starts from an initial configuration for two of the three bundle matrices. This initial configuration can be obtained (a) (pseudo-)randomly, (b) rationally by a built-in heuristic in the algorithm, or (c) from the user. Assuming, without loss of generality, that an initial configuration, $S^{(0)}$, $P^{(0)}$, has been obtained for the bundle matrices $S$ and $P$, the optimal bundle matrix $I^{(0)}$, which is such that (3) is minimal, is calculated conditionally upon $S^{(0)}$ and $P^{(0)}$. Next $S^{(1)}$ is estimated conditionally upon $P^{(0)}$ and $I^{(0)}$; then, $P^{(1)}$ is estimated conditionally upon $S^{(1)}$ and $I^{(0)}$, and so on. The process of alternately estimating $S$, $P$ and $I$ is continued until no further improvement in the loss function (3) is observed.

For the estimation of a bundle matrix conditionally upon the two other bundle matrices, one can make use of a separability property (Chaturvedi & Carroll, 1994) of the loss function (3). This property means that the contribution to the loss function (3) of the bundle pattern of, for example, an object $h$ in $S$ can be separated from the contribution of the bundle patterns of the other objects in $S$:

$$\sum_{h=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\left(d_{hjk} - \bigoplus_{b=1}^{r} s_{hb}p_{jb}i_{kb}\right)^2$$

$$= \sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\left(d_{1jk} - \bigoplus_{b=1}^{r} s_{1b}p_{jb}i_{kb}\right)^2 + \cdots + \sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\left(d_{n_1jk} - \bigoplus_{b=1}^{r} s_{n_1b}p_{jb}i_{kb}\right)^2. \tag{4}$$

Hence, the overall optimal estimate of $S$ (conditional upon $P$ and $I$) can be obtained by successively optimizing row $s_h$ in each term in (4). The same holds for the conditional estimations of $P$ and $I$. The best fitting bundle pattern for an element $x$ is found by means of Boolean regression (Leenen & Van Mechelen, 1998; Van Mechelen, 1988). In general, Boolean regression searches for a subset of a set of binary predictor variables, whose Boolean sum optimally predicts a given binary criterion value. In the particular case of fitting $s_h$ in the $h$-th term of (4), $r$ predictor variables are to be considered; the cell-entries $d_{hjk}$ ($j = 1..n_2$; $k = 1..n_3$) are the values of the criterion variable and the products $p_{jb}i_{kb}$ ($j = 1..n_2$; $k = 1..n_3$) are the values of the $b$-th predictor variable. $s_{hb}$ will be eventually assigned a value of 1 if $b$ belongs to the selected subset of predictors that optimally predicts the criterion, 0 otherwise.

In the second main routine, the bundle matrices at the end of the first routine are modified such as to make them consistent with the set-theoretical relations in the array $M$, that the bundle matrices yield by (2). For this, a closure operation (Barbut & Monjardet, 1970) is successively applied to each of the bundle matrices. This operation implies that zero-entries in the bundle

matrix are turned into one if this change does not alter $M$ (and, hence, neither the value on the loss function). It may be noted that this closure operation is a sufficient, though not necessary condition for set-theoretical consistency.

Several statistics can be calculated to assess the goodness of fit of an INDCLAS model, including the proportion of discrepancies between $D$ and $M$ (the badness-of-fit statistic minimized by the algorithm) and the Jaccard goodness-of-fit statistic (Jaccard, 1908; Sneath & Sokal, 1973). In practice, when the true rank is unknown, analyses are done in ranks 1 through some integer $q$. A choice between the resulting models can be based on methods described by Van Mechelen et al. (1995).

## 4. Simulation Study

In this simulation study, goodness of fit and goodness of recovery of INDCLAS solutions were examined. In this evaluation, three different types of $n_1 \times n_2 \times n_3$ arrays are involved: a true array $M$; a data array $D$, which is $M$ perturbed with error; and a model array $M'$, yielded by the algorithm. The true array $M$ can be perfectly represented by an INDCLAS model of true rank $r$ with an $n_1 \times r$ matrix $S$, an $n_2 \times r$ matrix $P$ and an $n_3 \times r$ matrix $I$. $M'$ is the model array of an INDCLAS model in rank $r'$ with an $n_1 \times r'$ bundle matrix $S'$, an $n_2 \times r'$ bundle matrix $P'$ and an $n_3 \times r'$ bundle matrix $I'$.

Three parameters were systematically varied in a complete trifactorial design: (a) The *size* $s$ of the three-way arrays, $n_1 \times n_2 \times n_3$, with values $15 \times 15 \times 15$, $10 \times 30 \times 20$, $60 \times 30 \times 20$, (b) the *true rank* $r$ of $M$ with values 2, 3, 4 and 5 and (c) the *error level* $e$, which is the percentage of discrepancies between $M$ and $D$, with values of 0%, 5%, 10%, 20%, 30%.

For each combination of size $s$, true rank $r$ and error level $e$, 25 bundle matrices $S$, $P$ and $I$ were generated with entries that were independent realizations of a Bernoulli variable with a parameter value for the distribution chosen such that the expected proportion of ones in $M$ equals 0.5. Next, a data array $D$ was constructed from each true array $M$ by randomly altering the value of a proportion $e$ of the entries in $M$. INDCLAS-analyses were performed on each of the resulting data arrays $D$ in ranks $r'$ 1 through 7.

The goodness of fit of analyses in the true rank (i.e., $r' = r$) was assessed by the proportion of discrepancies between $M'$ en $D$. Irrespective of size, true rank or error level, the mean proportion of discrepancies between $M'$ and $D$ deviates less than .005 from the error value $e$. Only 14 out of the 1,500 analyses in the true rank yield an INDCLAS model that deviates more from the data than the true model does. For the analyses in the true rank, the badness of recovery, measured by means of the proportion of discrepancies between $M$ and $M'$, was 0 (i.e., perfect recovery) for 82.9% of the analyses; the badness of recovery was less than .01 in 92.8% and less than .05 in 99.3% of the cases. The recovery generally improves (a) for larger data sets, (b) for lower true ranks and (c) for lower error levels. Concerning the latter, it must be noted, however, that, in general, the recovery of errorfree data was unexpectedly worse than the recovery of slightly error perturbed data sets. Furthermore, goodness of recovery was found to improve with the rank $r'$ of the fitted models increasing up to the true rank $r$; for analyses beyond the true rank, goodness of recovery remains unchanged (or deteriorates a little, mainly at higher error levels).

The goodness of recovery of the bundle matrices was assessed by the proportion of discrepancies between $S'$, $P'$, $I'$ and $S$, $P$, $I$, respectively (after an optimal permutation of the bundles). The results are comparable with the results of the goodness of recovery at the level of the three-way array: perfect recovery in 82.7% of the cases; for 89.4% of the cases, the proportion of discrepancies was less than .01; and for 97.5% less than .05. The recovery is less good for small, highly error perturbed data sets of high rank: for 9 of the 25 data sets with $s = 15 \times 15 \times 15$, $r = 5$, and $e = .30$, the badness of recovery was over .05. It is interesting to note that for 99.7% of the data sets with perfect recovery at the three-way array level, also the recovery at the bundle level is perfect, which suggests that the models in question are uniquely identified.

TABLE 5.
Mean Correlated Rand Index for analyses in the true rank

| | $s = 15 \times 15 \times 15$ | | | | $s = 10 \times 30 \times 20$ | | | | $s = 60 \times 30 \times 20$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $e$ | $r = 2$ | $r = 3$ | $r = 4$ | $r = 5$ | $r = 2$ | $r = 3$ | $r = 4$ | $r = 5$ | $r = 2$ | $r = 3$ | $r = 4$ | $r = 5$ |
| .00 | .96 | 1 | .95 | .96 | 1 | .97 | 1 | .98 | 1 | 1 | 1 | 1 |
| .05 | 1 | 1 | 1 | .98 | 1 | 1 | $\approx 1$ | .98 | .98 | 1 | 1 | 1 |
| .10 | 1 | $\approx 1$ | 1 | .99 | 1 | 1 | $\approx 1$ | .99 | 1 | 1 | 1 | 1 |
| .20 | 1 | .98 | .98 | .90 | $\approx 1$ | .99 | .96 | .96 | 1 | 1 | 1 | $\approx 1$ |
| .30 | .96 | .94 | .72 | .62 | .99 | .93 | .81 | .75 | 1 | 1 | .99 | .99 |

*Note.* $s$: size of the true array $M$ and the reconstructed array $M'$; $r$: Schein rank of $M$ and $M'$; $e$: error level; mean Corrected Rand Indices in each cell are calculated across 25 analyses.

An INDCLAS model also represents the equivalence and hierarchy relations among the elements of each mode. The recovery of the equivalence classes was assessed by the corrected Rand index (Hubert & Arabie, 1985) comparing the classifications from $S'$, $P'$, $I'$ and $S$, $P$, $I$, respectively. Table 5, which displays the mean corrected Rand indices for each combination of levels aggregated across the 3 modes, shows that the equivalence classes generally are very well recovered, except for data matrices of a high error level $e$. The recovery of the hierarchical relations was measured by means of the proportion of discrepancies between the hierarchy matrices (as defined in section 2) $U'$, $V'$ and $W'$ yielded by the algorithm and the true hierarchy matrices $U$, $V$ and $W$. Table 6 shows that the proportion of discrepancies generally is very close to zero.

As an aside, it may be noted that the Jaccard goodness-of-fit statistic can be very low even if the goodness of recovery is perfect, particularly at high error levels. In particular, the Jaccard values for the data sets that were perfectly recovered ranged from .847 to .932, from .706 to .860, from .531 to .732 and from .417 to .632 for the 5%, 10%, 20% and 30% error levels, respectively. In contrast with the proportion of discrepancies, the Jaccard goodness-of-fit statistic is highly sensitive to the number of 1-entries in the data: For a fixed (strictly positive) number of discrepancies, the Jaccard coefficient decreases with the number of 1-entries.

Summarizing, the INDCLAS algorithm succeeds in finding solutions that come close to the underlying truth. Even for data sets with a high error rate, the recovery of the truth is in many cases perfect or nearly perfect, especially for large data sets. Perfect recovery further makes it very likely that the algorithm has found a global optimum for the loss function (3): Given the random construction of the true bundle matrices in this study, why else should the algorithm show preference for the truth if the truth itself is not a global optimum?

## 5. Illustrative Application

In this section we present a reanalysis with INDCLAS of a study on decision making in psychiatric diagnosis (Van Mechelen & De Boeck, 1990). Fifteen clinicians (sources) were asked

TABLE 6.
Mean badness of recovery of the hierarchies for analyses in the true rank

| | $s = 15 \times 15 \times 15$ | | | | $s = 10 \times 30 \times 20$ | | | | $s = 60 \times 30 \times 20$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $e$ | $r = 2$ | $r = 3$ | $r = 4$ | $r = 5$ | $r = 2$ | $r = 3$ | $r = 4$ | $r = 5$ | $r = 2$ | $r = 3$ | $r = 4$ | $r = 5$ |
| .00 | .01 | 0 | .01 | .01 | 0 | .01 | 0 | $\approx 0$ | 0 | 0 | 0 | 0 |
| .05 | 0 | 0 | 0 | $\approx 0$ | 0 | 0 | $\approx 0$ | $\approx 0$ | .01 | 0 | 0 | 0 |
| .10 | 0 | $\approx 0$ | 0 | $\approx 0$ | 0 | 0 | $\approx 0$ | $\approx 0$ | 0 | 0 | 0 | 0 |
| .20 | 0 | $\approx 0$ | $\approx 0$ | .02 | $\approx 0$ | $\approx 0$ | .01 | .01 | 0 | 0 | 0 | $\approx 0$ |
| .30 | .01 | .01 | .06 | .06 | $\approx 0$ | .02 | .04 | .04 | 0 | 0 | $\approx 0$ | $\approx 0$ |

*Note.* $s$: size of the true array $M$ and the reconstructed array $M'$; $r$: Schein rank of $M$ and $M'$; $e$: error level; mean badness-of-recovery statistics in each cell are calculated across 25 analyses.

to evaluate 30 case descriptions of psychiatric inpatients (objects) with respect to 23 symptoms (attributes). An entry $d_{hjk}$ of the data array equals 1 if clinician $k$ indicates that symptom $j$ was presumably present for patient $h$, and 0 otherwise.

The resulting $30 \times 23 \times 15$ data array $D$ was analyzed by means of the INDCLAS algorithm in ranks 1 through 5. The proportion of discrepancies for the resulting solutions amounts to .22, .20, .20, .19, .18, respectively; the Jaccard values of the resulting solutions were .31, .38, .41, .42, .44, respectively. Based on a scree test of the Jaccard values, the model in rank 3 was retained. Interestingly, a HICLAS analysis on the same data set with the clinicians and patients concatenated into a single mode (resulting into a $450 \times 23$ matrix) yielded a proportion of discrepancies of .16 and a Jaccard value of .47 in rank 3 (Van Mechelen & De Boeck, 1990), which is only slightly better than the INDCLAS model in rank 3 (that can be conceived as a highly restrictive two-way model for concatenated data).

If we ignore for a moment the hexagons in the middle of the figure, Figure 6 is a graphic representation of the group structure for rank 3, with the boxes in the upper half of the figure displaying the frequencies of patients and the boxes in the lower half displaying symptoms.

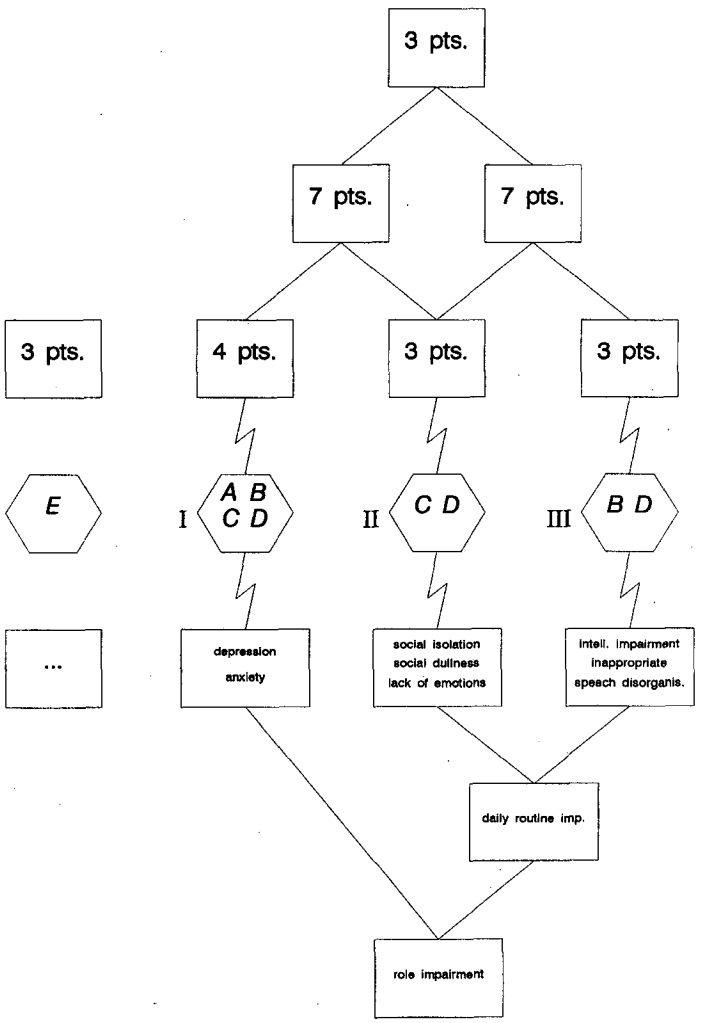



FIGURE 6.
Graphic representation of the INDCLAS model for the psychiatrist data. The classes of clinicians are represented by capital letters $A$ through $E$.

For the interpretation of the group structure, the symptom structure is most informative: Bundle I contains the depression and anxiety symptoms that are typical of affective disorders, whereas Bundles II and III contain a number of psychotic symptoms. The grouping of the symptoms in the bottom classes of Bundles II and III is in line with the classical distinction between negative and positive psychotic symptoms (Andreasen & Olsen, 1982; Stuart, Malone, Currie, Klimidis & Minas, 1995), with the bottom class of bundle II containing negative symptoms and the bottom class of bundle III containing positive symptoms. Daily routine impairment is implied by positive as well as negative psychotic symptoms, whereas role impairment is implied by all symptoms.

The validity of the group structure was supported by relating it to external information on the patients. In particular, the same clinicians had also been asked for binary judgments on the applicability of each of three diagnostic categories (Schizophrenic Disorder, Affective Disorder, and Anxiety Disorder) to each patient. Making use of the projection techniques described by Van Mechelen and De Boeck (1990), logical prediction rules can be derived to predict the diagnoses on the basis of the INDCLAS structure. For all three categories, psychologically meaningful rules were obtained: For Schizophrenic Disorder the optimal rule was to assign the diagnosis to patients that belong to Bundle II but not to Bundle I, that is, to patients that have negative psychotic symptoms but not affective symptoms. For Affective Disorder the optimal rule was to assign the diagnosis to the patients of Bundle I, that is to patients with the symptoms of anxiety and depression. Finally, Anxiety Disorder is assigned to patients that either only belong to Bundle I or to all three bundles, that is, to patients with affective symptoms and either no or both positive and negative psychotic symptoms. The goodness of fit of the three rules, as expressed by the $L_p$-statistic (Van Mechelen & De Boeck, 1990) were .348, .505, and .140 respectively (all significantly different from 0, $p < .05$). Interestingly, the $L_p$-values for the first two categories are higher than those obtained by Van Mechelen and De Boeck, who analyzed the same data with a less restrictive two-way model of a higher rank.

The hierarchy of the clinicians, grouped into five classes labeled $A$ through $E$, can be read from the source structure represented in Figure 7. The number of clinicians in each class is given

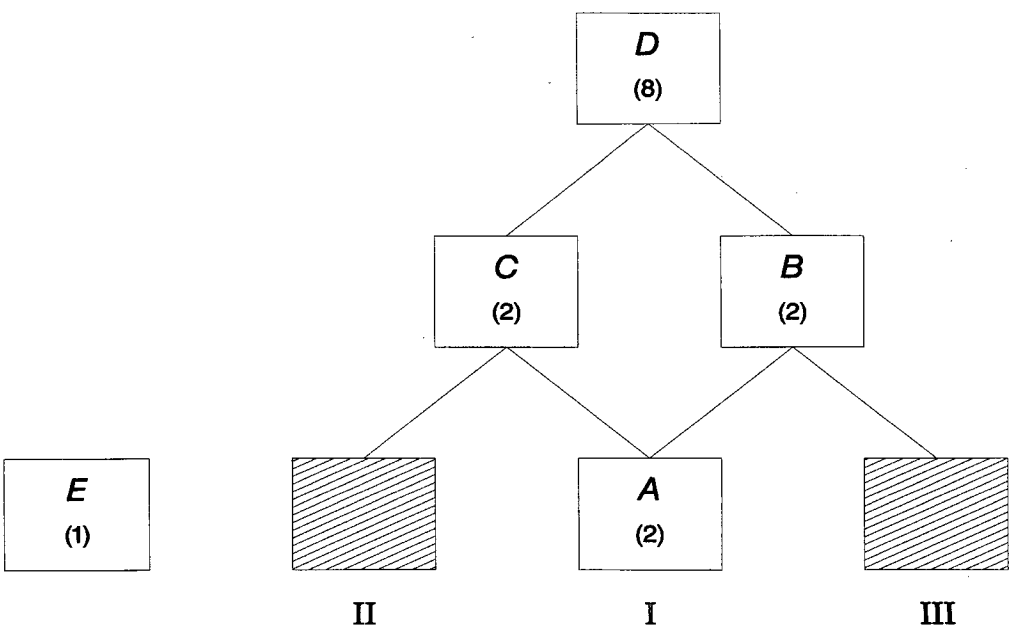



FIGURE 7.
Source/clinician structure for the psychiatrist data. The classes of clinicians are represented by capital letters $A$ through $E$.

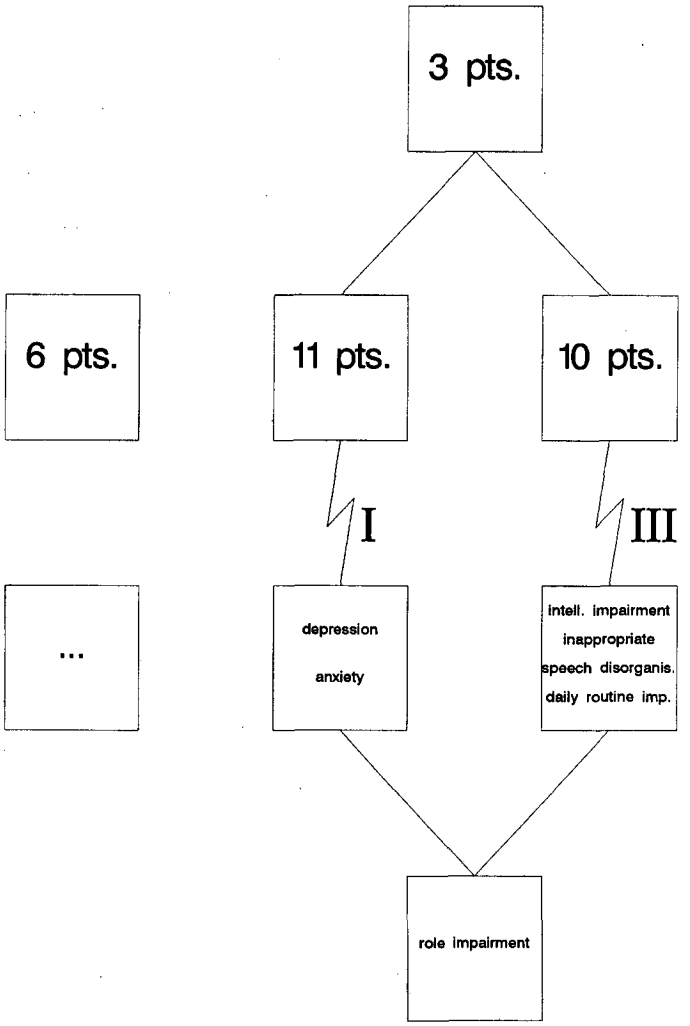



FIGURE 8.
Private structure for the clinicians in Class *B*.

between parentheses. The figure shows that the two clinicians of Class *A* only use the affective symptoms of Bundle I; diagnosticians of Class *C* disregard positive psychotic symptoms, whereas diagnosticians from Class *B* disregard negative psychotic symptoms. As an example, the private structure for the sources in Class *B* is drawn in Figure 8. The majority of diagnosticians (Class *D*) takes all types of symptoms into account. Finally, there is one clinician in the undefined class *E*, whose judgments do not fit into the group structure. Note that the partitions of the symptoms into equivalence classes may differ across source spaces, which may reveal different interpretations across sources: For example, sources in Class *B* consider daily routine impairment a positive symptom, whereas sources in Class *C* consider it a negative symptom.

Some external validation of the source structure was obtained in terms of additional information on the clinicians: The clinician in *E* was doing an internship, and was rather unexperienced and did not speak the local language (Dutch). Furthermore, the two diagnosticians in class *A* who did not take the psychotic symptoms into consideration did not have experience with psychiatric inpatients and, as a consequence, were not familiar with psychotic disorders.

Finally, in the full Figure 6 the group structure and the source structure are combined into an overall graphic representation.

## 6. Related Models

As a decomposition model for three-way data, INDCLAS is in a number of respects similar to other methods for the analysis of three-way data. Like the well-known models for three-way two-mode scaling, INDSCAL (Carroll & Chang, 1970), and three-way two-mode additive clustering, INDCLUS (Carroll & Arabie, 1983), the INDCLAS model makes use of a group structure that is differentially weighted by different sources. Moreover, the extension of HICLAS to INDCLAS is analogous to the extension of ADCLUS (Shepard & Arabie, 1979) to INDCLUS.

As a decomposition model for three-way three-mode data, INDCLAS closely resembles the PARAFAC/CANDECOMP model (Carroll & Chang, 1970; Harshman, 1970), which makes use of a decomposition rule similar to (2). More specifically, a constrained PARAFAC/CANDECOMP model, in which the component matrices are restricted to be binary, differs from INDCLAS in only two respects: First, INDCLAS involves a Boolean decomposition, whereas PARAFAC/CANDECOMP is based on an algebraic decomposition, and, second, in the INDCLAS model, unlike in the PARAFAC/CANDECOMP model, the binary bundle matrices are to represent the set-theoretical relations that exist in the predicted model array. It may be noted that both this restricted and the general PARAFAC/CANDECOMP model are instances of the generic CANDCLUS model (Carroll & Chaturvedi, 1995), which unifies many methods for $N$-way data analysis (including also INDSCAL and INDCLUS).

Finally, it may be noted that Leenen, Van Mechelen and De Boeck (in press) recently proposed a generic disjunctive/conjunctive decomposition model that subsumes the two-way and three-way hierarchical classes models as special cases. The same generic model also subsumes other disjunctive/conjunctive decomposition models for binary three-way three-mode data, including several three-way generalizations of Coombs and Kao's (1955; Coombs, 1964) nonmetric factor analysis (Leenen et al., in press).

### References


Andreasen, N.C., & Olsen, S. (1982). Negative $v$ positive schizophrenia: Definition and validation. *Archives of General Psychiatry, 39*, 789–794.

Barbut, M., & Monjardet, B. (1970). *Ordre et classification: Algèbre et combinatoire* [Order and classification: Algebra combinatorics]. Paris: Hachette.

Carroll, J.D., & Arabie, P. (1983). INDCLUS: An individual differences generalization of the ADCLUS model and MAPCLUS algorithm. *Psychometrika, 48*, 157–169.

Carroll, J.D., & Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an $N$-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35*, 283–319.

Carroll, J.D., & Chaturvedi, A. (1995). A general approach to clustering and multidimensional scaling of two-way, three-way, or higher-way data. In R.D. Luce, M. D'Zmura, D. Hoffman, G.J. Iverson & A.K. Romney (Eds.), *Geometric representations of perceptual phenomena* (pp. 295–318). Mahwah: Erlbaum.

Chaturvedi, A., & Carroll, J.D. (1994). An alternating combinatorial optimization approach to fitting the INDCLUS and generalized INDCLUS models. *Journal of Classification, 11*, 155–170.

Coombs, C.H. (1994). *A theory of data*. New York: Wiley.

Coombs, C.H., & Kao, R.C. (1955). *Nonmetric factor analysis* (Engineering Research Bulletin No. 38). Ann Arbor: University of Michigan Press.

De Boeck, P. (1986). *HICLAS computer program: Version 1.0*. Leuven: Katholieke Universiteit.

De Boeck, P., & Rosenberg, S. (1988). Hierarchical classes: Model and data analysis. *Psychometrika, 53*, 361–381.

Harshman, R.A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics, 16*, 1–84.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale [New research on floral distribution]. *Bulletin de la Société Vaudoise de Sciences Naturelles, 44*, 223–270.

Kim, K.H. (1982). *Boolean matrix theory and applications*. New York: Marcel Dekker.

Kruskal, J.B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Its Applications, 18*, 95–138.

Kruskal, J.B. (1989). Rank, decomposition, and uniqueness for 3-way and $N$-way arrays. In R. Coppi & S. Bolasco (Eds.), *Multiway data analysis* (pp. 7–18). Amsterdam: North Holland.

Leenen, I., & Van Mechelen, I. (1998). A branch-and-bound algorithm for Boolean regression. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Data highways and information flooding, a challenge for classification and data analysis* (pp. 164–171). Berlin: Springer-Verlag.

Leenen, I., Van Mechelen, I., & De Boeck, P. (in press). A generic disjunctive/conjunctive decomposition model for $n$-ary relations. *Journal of Mathematical Psychology.*

Shepard, R.N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86*, 87–123.

Sneath, P.H.A., & Sokal, R.R. (1973). *Numerical taxonomy.* San Francisco: Freeman.

Stuart, G.W., Malone, V., Currie, J., Klimidis, S., & Minas, I.H. (1995). Positive and negative symptoms in neuroleptic-free psychotic inpatients. *Schizophrenia Research, 16*, 175–188.

Van Mechelen, I. (1988). Prediction of a dichotomous criterion variable by means of a logical combination of dichotomous predictors. *Mathématiques, Informatiques et Sciences Humaines, 102*, 47–54.

Van Mechelen, I., & De Boeck, P. (1990). Projection of a binary criterion into a model of hierarchical classes. *Psychometrika, 55*, 677–694.

Van Mechelen, I., De Boeck, P., & Rosenberg, S. (1995). The conjunctive model of hierarchical classes. *Psychometrika, 60*, 505–521.

*Manuscript received 5 FEB 1997*
*Final version received 17 DEC 1997*