Center for Quality and Applied Statistics
Kate Gleason College of Engineering
Rochester Institute of Technology

Combining Successive-Categories (Rating) Data And Dichotomous Incidence Data: A Forced Classification Procedure for Dual Scaling

Adam E. Wyse
Michigan State University

Daniel R. Lawrence
Center for Quality and Applied Statistics
Rochester Institute of Technology
drleas@rit.edu

Technical Report 2006–4

September 2006

Combining Successive-Categories (Rating) Data And Dichotomous Incidence Data: A Forced Classification Procedure for Dual Scaling [1]

Adam E. Wyse

Michigan State University

Daniel R. Lawrence

Rochester Institute of Technology

Abstract

Demographic information collected on surveys is often of interest to those who use these surveys to study behavioral or psychological phenomenon. This article presents a method for imposing a dichotomous incidence variable, possibly a demographic variable, on the weighting of items in a set of successive-categories (rating) data in a dual scaling analysis. The original matrix of rating data is augmented with the "criterion" item and "centered" between two numbers that represent the two criterion groups, so this item determines one of the initial solutions of the dual scaling analysis. The procedure is discussed with both practical guidelines for its use and interpretation of results. Examples of application involve both fabricated and real data.

Keywords: Dual Scaling, Forced Classification, Successive-Categories Data, Rating Data

Introduction

It is often of particular interest to those who develop surveys to solicit not only respondents' preferences and opinions, but also to study commonalities or differences in the respondents that are being surveyed. Many of these differences or commonalities are captured in the form of demographic information. For example, it might be of particular interest to those who are conducting the survey to study the effect that gender has on how respondents answered a specific item or set of items. Further, it might be of importance to identify the questions whose responses are most consistent with gender, since products and services are often designed to target segments of the population that have these characteristics.

In psychological measurement, this information might be a classification of whether an individual has high or low motivation. It could also be a classification as to whether someone does or does not have a certain characteristic, such as a learning disability. The items on which such information is collected are often categorical in nature and can usually be classified as incidence data, since membership in a specific category is generally absolute and does not indicate a degree of "preference" of one category over another. If the data of the survey are also in the form of incidence data, then this particular characteristic

---

[1] This article is adapted from a Master's Thesis written by the first author while a graduate student at Rochester Institute of Technology.

can be included with the incidence preference data of the survey and submitted to dual scaling for a forced classification analysis (see Nishisato (1984)). In this way a demographically defined grouping variable[2] can be made to determine the solution. The survey items or questions whose responses are most influenced by the particular demographic will have the highest absolute weights (e.g., see Day (1989) and Nishisato (1994)). Unfortunately, though, preference data are not always collected in the form of incidence data. A different strategy is needed when these data have been collected in the form of dominance data (i.e., paired-comparison, rank-order, or successive-categories data) and the user still wants to determine the effect of a demographically-defined grouping variable whose information has been collected as incidence data.

Lawrence (2000) discussed this problem as it pertains to rank-order data. He described a method for imposing a dichotomous (male vs. female, user vs. non-user, etc.) criterion item on the weighting of items in a dual scaling analysis of a set of rank-order data. Lawrence (2006) also formulated a procedure for handling paired-comparison data that is similar to his method for analyzing rank-order data. In both cases, the criterion variable is "transformed" to dominance data and included as part of the data matrix, which is then subjected to a dual scaling analysis. Presently, there is no such method for handling successive-categories (rating) data.

This article proposes a method for the case of rating data that capitalizes on both the definition of the format and the structure of the data. This procedure is applied to contrived data with practical guidelines for use of the method and interpretation of the solutions. Finally, a subset of "real" data obtained from a health survey given at the Rochester Institute of Technology is analyzed, followed by a discussion of the results.

Dual Scaling and Forced Classification

Dual scaling (Nishisato, 1980 & 1994) is a multivariate method that can be used to analyze a wide range of categorical data, including both incidence data (contingency tables, multiple-choice responses, and sorting formats) and dominance data (paired comparisons, ranking data, and successive-categories (or rating) data). The objective of dual scaling, of course, is to determine optimal weights for stimuli/items (columns) and scores for subjects/respondents (rows) given some initial constraints or conditions—that is, some criterion that needs to be maximized and constraints that must be satisfied based on the formulation (normalization, sum-of-squares-equal-to-a-prescribed-total, sum-of-weighted-responses-equal-to-zero, and so on).

Dual scaling also allows for a unique option known as forced classification (Nishisato, 1984). This procedure is based the *Principle of Internal Consistency* (Guttman, 1950) which states that if a given response pattern is repeated in the data matrix, it will become the primary factor in determining the initial solution. Basically, forced classification augments an original matrix by repeating a given response pattern a certain number of times k. As the value of k is increased, theoretically to infinity, the particular item that has been repeated k times, referred to as the "criterion" item, becomes the principal factor in determining the first solution of the dual scaling analysis. In fact, the repeated item and the new dimension have a correlation that approaches one, since the repeated item effectively defines the dimension (Nishisato & Gaul, 1990). An obvious benefit of forced classification is that other items that correlate highly with this dimension can be easily identified. Nishisato (1986) generalized his forced classification procedure to handle not only multiple-choice data but also sorting data and, in a specific way, rank-order data and paired-comparison data. He further demonstrated that the value of k can be chosen not only to get a particular solution to *dominate* the analysis but also to cause a particular solution to be *suppressed*. Forced classification has also been adapted for use on contingency tables, and this is commonly referred to as *conditional forced classification* (Nishisato & Baba, 1999). The mathematical

[2] In this paper, we use the expressions "grouping variable" and "criterion item" interchangeably. Since they can refer to *any* nominal classification variable, the procedure outlined in this paper can be used to force any relevant dichotomous nominal characteristic on a dual scaling analysis of a set of rating data.

aspects of the procedure for each of these particular data types, excluding conditional forced classification, are discussed by Nishisato (1988 & 1994).

Several different applications of the forced classification method have recently been considered, but in virtually all of them the preference data were either incidence data or treated as incidence data. Especially in the context of forced classification, the study of dominance data (i.e., paired-comparison, rank-order, and successive-categories data) has been much more limited. Nishisato (1984, 1986 & 1994) did study forced classification of dominance data, but he did it in the framework of choosing a *pair* of items to be used as the criteria (Nishisato, 1994). He stated that *two* items were needed to "drive" the solution and that each of these arbitrarily chosen criterion items must be multiplied by *k* within a "dominance matrix" in order to force the solution. He described how this could be done in the cases of paired-comparison and rank-order data but did not extend this to successive-categories data.

Lawrence (2000) proposed a method for imposing a dichotomous incidence variable, the criterion item, on the weighting of items in a dual scaling analysis of rank-order (dominance) data. Essentially, Lawrence's method combined the incidence item with dominance data by, in effect, converting the incidence data to rankings and combining them with the original body of rank data. The subsequent matrix of rankings was then subjected to a standard dual scaling analysis and was shown to "force" the dimension of interest to be the first solution of the analysis—that is, the criterion item was made to determine the first solution. More recently, Lawrence (2006) has applied a similar strategy in the forced classification analysis of paired-comparison data, here again subject to a dichotomous incidence variable. The case involving successive-categories (rating) data is the focus of this article.

Successive-Categories (Rating) Data

The study and implementation of methods for the analysis of successive-categories (rating) data have been somewhat scarce with results that are sometimes debatable. Successive-categories data, also know as ordered-choice or Likert data, are based on a set of strictly ordered ratings (e.g., *Poor*, *Fair*, *Good, Excellent*) that are assigned to a given set of stimuli by each respondent. Many analysts, in their multivariate analyses, treat this type of data as if it were continuous, often subjecting the data to principal components analysis or factor analysis. Neither of these techniques is really appropriate for analyzing such data, though, especially when the response categories are few in number. Nishisato's dual scaling approach best analyzes this type of data for what it is, namely categorical, and generally should be used for analyses of this sort.

The simplest of three methods presented by Nishisato and Sheu (1984), the approach used by dual scaling in the analysis of successive-categories data assigns ranks to both the stimuli and the (implicit) category boundaries that exist in the data. *Table 1* is an example of a typical successive-categories response pattern with four subjects rating five items using any one of four possible ratings (or categories), *Poor*, *Fair*, *Good* and *Excellent*, where 1 is made to represent *Poor*, 2 is made to represent *Fair*, 3 is made to represent *Good*, and 4 is made to represent *Excellent*. There would be three category boundaries, one between each pair of successive categories—that is, between *Poor* and *Fair*, *Fair* and *Good*, and *Good* and *Excellent*. Ranks are then assigned to both the items that have been rated by the respondents along with the category boundaries between these ratings.

Table 1:

Subject	Item 1	Item 2	Item 3	Item 4	Item 5
1	3	4	2	3	1
2	4	4	2	1	3
3	1	2	4	3	2
4	2	2	3	4	2

Nishisato (1994) describes a systematic procedure for assigning these ranks: Starting with the first category, the average rank $(k_0 + 1)/2$, where k_0 is the total number of stimuli classified into Category 1, is given to all the stimuli in that category. The category boundary τ_1, between Categories 1 and 2, is given rank $k_0 + 1$. The next set of stimuli, k_1, where k_1 is the total number of stimuli classified into Category 2, are assigned the rank $(k_1 + 1)/2 + (k_0 + 1)$. The second category boundary τ_2, between Categories 2 and 3, is given the rank $(k_0 + 1) + (k_1 + 1)$. This process continues until all the categories and category boundaries have received the proper rankings. The data in *Table 1* would have the rankings shown in *Table 2*. These rankings are then converted to dominance numbers e_{ij} by the formula

$$e_{ij} = 2K_{ij} - (n + m + 1), \tag{1}$$

where K_{ij} is the rank assigned by subject i to item j, n is the number of items, and m is the number of category boundaries (Nishisato, 1994). The resulting matrix of dominance numbers is then subjected to a dual scaling analysis.

Table 2:

Subject	τ_1	τ_2	τ_3	Item 1	Item 2	Item 3	Item 4	Item 5
1	2	4	7	5.5	8	3	5.5	1
2	2	4	6	7.5	7.5	3	1	5
3	3	5	7	1.5	4	8	6	1.5
4	1	5	7	3	3	6	8	3

A major source of debate and a reason for limited analysis of successive-categories data (as categorical data) revolves around the aspect of multidimensionality. Nishisato (1994) pointed out that due to the empirical nature of the analysis, solutions beyond the first one commonly do not have numerically ordered category boundaries. This often makes it rather difficult to interpret any solution other than the first. Extensive work has been done to try to remedy this (e.g., Odondi (1997)), but the suggested methods are not very user-friendly and often require more than one pass through the data.

Extracting only the first solution, which has ordered category boundaries, appears to be the most popular analysis strategy since this solution would seem to be the only one that can be reasonably easily interpreted. However, this issue takes on greater importance and generates more debate when respondents from different subpopulations are represented in the same data set. It makes sense, at least intuitively, that these people would answer questions in a different manner. Males and females, for instance, might be subconsciously using a different continuum or set of boundaries when answering questions. In this sense, the category boundaries of a dimension so determined might not be ordered due to these differences, and this problem would only be compounded as the number subgroups responding to the survey increases.

In fact, though, the analyst is at least as, if not more, concerned with the weights of items of a survey as he is those of the category boundaries—especially given that the category boundaries have been arbitrarily introduced into the analysis anyway. With this in mind, the procedure of this article is based on extracting more than one solution in the dual scaling analysis and then identifying the items whose weights most reflect the different demographic characteristics of the subgroups, ignoring the category boundaries and their order. The first solution is still of certain importance, since it has the ordered category boundaries and explains most of the variation in the data, but the focus is placed almost entirely on subsequent solutions (primarily the second) and the weights assigned to the items in those solutions.

Finally, there are two additional attributes or characteristics of successive-categories data that are worth mentioning. *Table 1* illustrates a very common but notable situation that occurs when people are asked to respond using rating scales. Subject 4 of *Table 1* has chosen not to give any of the five items a

Poor rating. It is quite possible that a subject would not use all of the categories in assigning ratings to a set of items, or that a given item would receive the same rating from all subjects or none of a particular rating from all subjects. This will be important in dual scaling analyses of successive-categories data to follow.

A second important attribute of successive-categories data is that each response is generally independent of any other. The rating that a respondent gives to each item is based on how he defines the categories in relation to that particular item. In essence, his responses are based on his definition of the continuum of choices he has to choose from based on that item. There is not a strict inter-dependency of items—as there is in the case of rank-order data, say, where once one particular item is given the top ranking, no other item can receive the same rank (assuming no ties). With these concepts in mind, the method of this article is formulated.

Matrix Notation

Let $\mathbf{F}_{N\times n}$ be a $N \times n$ matrix of successive-categories data, where *N* is the number of subjects and *n* is the number of items. Also, let $\mathbf{x}_{N\times 1}$ be a column vector representing some grouping or demographic characteristic. In most cases, membership in a demographic (demo) group is generally coded as "1", "2", "3", …, where "1" represents membership in the first group, "2" represents membership in the second group, and so on.

Suppose a survey is given to two groups of respondents—males and females, or some other dichotomous grouping—and we are interested in the effect of the male-female dichotomy, say, in the way these people respond to the items on a survey. Membership in the two groups represented in this variable is recorded on the survey. A column vector $\mathbf{x}_{N\times 1}$ for the grouping variable, indicating the group membership of each respondent, is constructed with attribute A (e.g., male) coded as a "1" and attribute B (e.g., female) coded as a "2". The goal is to include the information of this incidence demographic variable with the dominance successive-categories data, collected as ratings on the survey, so that in a subsequent dual scaling analysis the weighting of the items in the data set will reflect the effect of the grouping variable in one of the initial solutions of the analysis.

We will begin by introducing the column vector $\mathbf{x}^*_{N\times 1}$, a sort of redefinition of $\mathbf{x}_{N\times 1}$. Let all of those that belong to the first group in $\mathbf{x}_{N\times 1}$ be given a value of 1 in $\mathbf{x}^*_{N\times 1}$. Then, let all those that belong to the second group in $\mathbf{x}_{N\times 1}$ be assigned the value $c+2+2c_1$ in $\mathbf{x}^*_{N\times 1}$, where c is the number of categories in the original data set and c_1 is the number of new categories to be included between the original data and the two extreme values in $\mathbf{x}^*_{N\times 1}$. (Note that the variables c_1 and $c+2+2c_1$ are set simultaneously.) More simply, for the $i^{\underline{th}}$ respondent

$$\mathbf{x}^*_{N\times 1} = \begin{cases} 1, & \text{if respondent is from Group 1} \\ c+2+2c_1, & \text{if respondent is from Group 2} \end{cases}. \tag{2}$$

By way of example, suppose $N = 5$ respondents, belonging to either one of two groups, rate $n = 3$ items on a scale of 1 to $c = 3$ (i.e., into any one of three categories). The grouping vector $\mathbf{x}$ and the data matrix $\mathbf{F}$ might, respectively, look like

$$\mathbf{x}_{5\times 1} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{F}_{5\times 3} = \begin{bmatrix} 1 & 3 & 2 \\ 3 & 2 & 1 \\ 3 & 1 & 2 \\ 1 & 2 & 2 \\ 1 & 3 & 2 \end{bmatrix}.$$

If $c_1 = 5$ new categories were to be included on either side of the categories of the original data matrix **F**, then we would have $c + 2 + 2c_1 = 3 + 2 + 10 = 15$ and

$$\mathbf{x}^*_{5\times 1} = \begin{bmatrix} 1 \\ 15 \\ 15 \\ 1 \\ 1 \end{bmatrix}.$$

The original data (consisting of ratings) should also be modified so that 1 and $c + 2 + 2c_1$ are the extreme choices of the successive categories, and the ratings in the original data are "centered" between 1 and $c + 2 + 2c_1$. (The purpose of "centering" the data matrix $\mathbf{F}_{N\times n}$ and altering the vector $\mathbf{x}_{N\times 1}$ will be explained in the following section, as will the manner in which the value of c_1 is determined.) The modification of **F** is done in the following manner: Let $\mathbf{G}_{N\times n}$ be a "centering" matrix to be added. Define **G** such that

$$\mathbf{G} = (c_1 + 1) \cdot \mathbf{1}_{N\times n}, \tag{3}$$

where c_1 is again the number of new categories included and $\mathbf{1}_{N\times n}$ is the unit matrix. We now modify the original data matrix **F** by adding **G** and augmenting **x*** to that sum. Expressed symbolically,

$$\mathbf{F}^*_{N\times(n+1)} = \left[(\mathbf{F} + \mathbf{G})_{N\times n} \middle| \mathbf{x}^* \right]. \tag{4}$$

The new $N \times (n+1)$ matrix **F*** is then submitted to dual scaling for analysis.

Again, by way of example, if **F**, c_1 and **x*** are defined as above, the centering matrix would be

$$\mathbf{G}_{5\times 3} = (5+1) \cdot \mathbf{1}_{5\times 3} = \begin{bmatrix} 6 & 6 & 6 \\ 6 & 6 & 6 \\ 6 & 6 & 6 \\ 6 & 6 & 6 \\ 6 & 6 & 6 \end{bmatrix},$$

so that

$$\mathbf{F}^*_{5\times 4} = \left[\left[\begin{pmatrix} 1 & 3 & 2 \\ 3 & 2 & 1 \\ 3 & 1 & 2 \\ 1 & 2 & 2 \\ 1 & 3 & 2 \end{pmatrix} + \begin{pmatrix} 6 & 6 & 6 \\ 6 & 6 & 6 \\ 6 & 6 & 6 \\ 6 & 6 & 6 \\ 6 & 6 & 6 \end{pmatrix} \right] \middle| \begin{bmatrix} 1 \\ 15 \\ 15 \\ 1 \\ 1 \end{bmatrix} \right] = \left[\begin{array}{ccc|c} 7 & 9 & 8 & 1 \\ 9 & 8 & 7 & 15 \\ 9 & 7 & 8 & 15 \\ 7 & 8 & 8 & 1 \\ 7 & 9 & 8 & 1 \end{array} \right].$$

Note that the categories of the original data have been "centered" in **F*** between 1 and 15. There are ten new categories in **F***, five (2, 3, 4, 5 and 6) between 1 (the category representing Group 1 in **x***) and 7 (the "adjusted" lowest category of **F**) and five (10, 11, 12, 13 and 14) between 9 (the "adjusted" highest category of **F**) and 15 (the category representing Group 2 in **x***).

Principles Behind the Procedure

The procedure described in this article is based on several principles that allow for its formulation and the interpretation of its results. First of all, the values in the "group" vector **x*** must be assigned in such a way that they are at the extreme categories of the "new" data matrix **F*** and away from the categories of the original data matrix **F**. The way that dual scaling handles the data makes the reason for the definition of **F*** apparent. If the dichotomous elements that comprise **x*** were simply the extreme categories of original the data, then the groupings of the respondent scores would be dependent on how many other items had been classified by respondents in those extreme categories. This would make it difficult to identify the two groups by way of the respondent scores, since each score within a subgroup would not necessarily have the same, or even approximately the same, magnitude. This necessitates the assignment of values in **x*** to be such that the two groups are represented by the extreme categories of **F***, away from, or beyond, the categories of the original data.

The definition of **x*** also depends on the number of new categories c_1 that are included between the original data and the two extreme categories representing the two subgroups. (Including additional categories between the extreme categories of the data and the categories represented in the grouping vector **x*** doesn't violate the "integrity" of the original data; the added categories could simply be thought of as categories that are unused when a respondent rates the items.) In most cases, adding ten new categories, denoted by c_1, to each side of the data matrix should be sufficient to impose the effect of the demographic group on the weighting of the items. (If the number of respondents increases significantly, it might be necessary to increase the value of c_1.)

An examination of the proposed procedure points to differences between this method and other methods for doing forced classification. One major difference is that in this procedure the effect of the grouping variable appears in the second solution of the dual scaling output instead of the first solution, as it does in the forced classification procedures that have been developed to handle other types of data. In the case involving successive-categories data, the first solution will have the category boundaries ordered and represent the original data as free of the effect of the grouping variable as possible, while the second solution will not necessarily have ordered category boundaries but will reflect the effect of the grouping variable. The reason for this is no doubt related to the additional categories introduced into the matrix **F***. These new categories are directly related to the values (group number) in the vector **x***, and the corresponding category boundaries are linear combinations of each other in the dominance matrix. The dominance numbers for the category boundaries between the category corresponding to Group 1 and the categories for the original data are "complementary" to those between the categories for the original data and the category corresponding to Group 2. The difference between the dominance numbers for Group 1 and Group 2 in each of these newly introduced category boundaries is only two, which follows directly from the way the dominance matrix is constructed.

If an infinite number of new categories (and category boundaries) were introduced, the difference in the dominance numbers would be so minimal that it would be impossible to distinguish between the two groups based on the dominance numbers for the category boundaries in the dominance matrix. Under these circumstances, the first solution of the dual scaling analysis would account for virtually 100% of the variation in the data with the categories boundaries appearing to determine that first solution. Meanwhile, the respondent scores for the two groups approach one and the weights for the items in **F*** will approach those for the items in **F**.[3] It is important to note that **F*** contains the group vector as an additional item; hence, there will never be a perfect correlation between weights for the items in **F** and **F*** unless the grouping variable had no effect on the original responses comprising the data.

---

[3] An exception to this occurs when item responses in the original data are similar—in the same or opposite direction—to the groupings in the grouping vector. In this case, the correlation between the weights for the items in **F** and the weights for items in **F*** will not be as high as it would be if none of the items of where similar to the groupings in the demo vector.

Since the first solution of the dual scaling analysis accounts for most of the variation in the original data and category boundaries, the second solution will account for the grouping variable. (Since the elements of the vector in the dominance matrix corresponding to the grouping variable are the most positive and most negative in value, it is not surprising that this would constitute the next most variation to be explained.) Accordingly, the assignment of weights to the items will be influenced by the grouping variable. Item weights that are directionally the same as the weight for the criterion item (positively influenced) will have the highest signed value in the same direction. Moreover, the respondent scores will fall into two groups (consistent with the grouping in the criterion item) that are opposite in sign but similar in magnitude; within the two groups, the respondent scores are very nearly the same. These are desired results of a forced classification analysis.

Another major difference between this forced classification procedure and the procedures for handling other types of dominance data is the need for only a one-column formulation of the grouping variable to force the solution. The apparent reason that only one column is needed in this case is linked to the nature of successive-categories data; the items in successive-categories are not inter-dependent. In fact, as was previously stated, the response to each item is based on the respondent's definition of the rating categories relative to each individual item. Therefore, it makes sense that a single-column formulation of the criterion item could be used.

Expositional Application

Suppose ten respondents are asked six questions that involve ratings based on a rating scale where 1 represents "poor", 2 represents "fair", 3 represents "good", and 4 represents "excellent", leading to a 10×6 data matrix **F**. Also, suppose that information on a dichotomous grouping variable for the ten respondents is recorded along with the rating data. The group information comprises a one-column vector **x**, where among its elements a 1 represents membership in Group 1 and a 2 represents membership in Group 2. The proposed modification to **F** is shown below with ten new categories included between the centered data and the two "adjusted" values of the grouping variable, which become the two extreme categories of **F***. The centering matrix **G** is added to **F** and the new group vector **x*** is then augmented to form the 10×7 matrix **F***. An analysis of both the original data matrix **F** and the new matrix **F*** is carried out using dual scaling. The results of the two analyses are shown, including the item weights, category boundary weights, respondent scores, and percent of total variation explained by each solution. The two sets of results are then interpreted and compared.

$$\text{Let } \mathbf{F} = \begin{bmatrix} 4 & 1 & 3 & 4 & 2 & 3 \\ 3 & 3 & 2 & 1 & 2 & 3 \\ 3 & 2 & 3 & 2 & 4 & 2 \\ 4 & 1 & 4 & 3 & 2 & 4 \\ 3 & 2 & 3 & 4 & 3 & 1 \\ 4 & 2 & 3 & 4 & 2 & 2 \\ 2 & 4 & 2 & 4 & 3 & 1 \\ 4 & 2 & 2 & 3 & 2 & 4 \\ 1 & 3 & 3 & 2 & 4 & 3 \\ 3 & 1 & 1 & 1 & 3 & 3 \end{bmatrix} \text{ and } \mathbf{x} = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 1 \\ 2 \\ 1 \\ 2 \end{bmatrix}.$$

$$\text{Then, } \mathbf{G} = \begin{bmatrix} 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \end{bmatrix} \text{ and } \mathbf{x}^* = \begin{bmatrix} 26 \\ 1 \\ 1 \\ 26 \\ 26 \\ 26 \\ 1 \\ 26 \\ 1 \\ 26 \end{bmatrix}, \text{ so that}$$

$$\mathbf{F}+\mathbf{G} = \begin{bmatrix} 4 & 1 & 3 & 4 & 2 & 3 \\ 3 & 3 & 2 & 1 & 2 & 3 \\ 3 & 2 & 3 & 2 & 4 & 2 \\ 4 & 1 & 4 & 3 & 2 & 4 \\ 3 & 2 & 3 & 4 & 3 & 1 \\ 4 & 2 & 3 & 4 & 2 & 2 \\ 2 & 4 & 2 & 4 & 3 & 1 \\ 4 & 2 & 2 & 3 & 2 & 4 \\ 1 & 3 & 3 & 2 & 4 & 3 \\ 3 & 1 & 1 & 1 & 3 & 3 \end{bmatrix} + \begin{bmatrix} 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \\ 11 & 11 & 11 & 11 & 11 & 11 \end{bmatrix} = \begin{bmatrix} 15 & 12 & 14 & 15 & 13 & 14 \\ 14 & 14 & 13 & 12 & 13 & 14 \\ 14 & 13 & 14 & 13 & 15 & 13 \\ 15 & 12 & 15 & 14 & 13 & 15 \\ 14 & 13 & 14 & 15 & 14 & 12 \\ 15 & 13 & 14 & 15 & 13 & 13 \\ 13 & 15 & 13 & 15 & 14 & 12 \\ 15 & 13 & 13 & 14 & 13 & 15 \\ 12 & 14 & 14 & 13 & 15 & 14 \\ 14 & 12 & 12 & 12 & 14 & 14 \end{bmatrix}$$

$$\text{and } \mathbf{F}^* = \left[\begin{array}{cccccc|c} 15 & 12 & 14 & 15 & 13 & 14 & 26 \\ 14 & 14 & 13 & 12 & 13 & 14 & 1 \\ 14 & 13 & 14 & 13 & 15 & 13 & 1 \\ 15 & 12 & 15 & 14 & 13 & 15 & 26 \\ 14 & 13 & 14 & 15 & 14 & 12 & 26 \\ 15 & 13 & 14 & 15 & 13 & 13 & 26 \\ 13 & 15 & 13 & 15 & 14 & 12 & 1 \\ 15 & 13 & 13 & 14 & 13 & 15 & 26 \\ 12 & 14 & 14 & 13 & 15 & 14 & 1 \\ 14 & 12 & 12 & 12 & 14 & 14 & 26 \end{array}\right].$$

The output of the dual analysis of **F** follows:

Table 3: Variation by Solution for DSA of **F** (Original Data)

Variation	1	2	3	4	5
Pct Total	40.313	22.032	20.704	8.560	4.558
Cum Pct	40.313	62.400	83.104	91.664	96.222

Table 4: Item Weights by Solution for DSA of **F** (Original Data)

Item	1	2	3	4	5
1	1.418	0.377	-1.009	0.357	-1.010
2	-1.260	-0.357	0.798	1.849	1.003
3	0.126	-0.438	-0.232	-1.728	1.753
4	0.553	-2.108	-0.916	0.307	0.108
5	-0.312	-0.068	1.746	-1.328	-0.828
6	0.157	1.986	-0.694	0.132	1.078

Table 5: Category Boundaries by Solution for DSA of **F** (Original Data)

Cat Bound	1	2	3	4	5
1	-1.775	0.254	-1.000	-0.445	-1.313
2	-0.231	0.074	-0.076	0.313	-0.452
3	1.325	0.281	1.384	0.544	-0.338

Table 6: Respondent Scores by Solution for DSA of **F** (Original Data)

Resp	1	2	3	4	5
1	1.443	-0.373	-0.775	-0.166	-0.079
2	0.643	1.267	0.782	1.730	0.598
3	0.914	0.039	1.436	-1.279	-0.480
4	1.218	0.633	-0.774	-1.033	1.470
5	1.026	-1.474	0.506	-0.456	-0.533
6	1.353	-0.941	-0.376	0.323	0.137
7	0.155	-1.564	1.131	1.524	0.184
8	1.291	0.694	-0.456	1.094	0.219
9	0.036	0.339	1.922	-0.666	1.587
10	0.783	1.323	0.739	-0.051	-2.084

The output of the dual scaling analysis of **F*** follows:

Table 7: Variation by Solution for DSA of **F*** (with Criterion Item)

Variation	1	2	3	4	5
Pct Total	89.942	9.069	0.467	0.218	0.161
Cum Pct	89.942	99.011	99.478	99.696	99.857

Table 8: Item Weights by Solution for DSA of **F*** (with Criterion Item)

Item	1	2	3	4	5
1	0.157	0.396	-0.817	-2.404	-1.389
2	-0.161	-0.866	0.759	-1.360	2.891
3	-0.012	-0.134	0.807	0.567	-4.306
4	0.042	0.287	3.914	-1.465	0.031
5	0.012	-0.667	0.211	4.346	0.187
6	-0.011	0.048	-3.774	-1.098	-0.296
7	0.395	5.434	0.008	0.625	0.580

Table 9: Category Boundaries by Solution for DSA of **F*** (with Criterion Item)

Cat Bound	1	2	3	4	5
1	-1.726	-0.047	-0.002	-0.019	-0.047
2	-1.612	-0.055	-0.002	-0.019	-0.045
3	-1.498	-0.064	-0.002	-0.020	-0.043
4	-1.383	-0.073	-0.002	-0.020	-0.041
5	-1.269	-0.081	-0.001	-0.020	-0.040
6	-1.155	-0.090	-0.001	-0.020	-0.038
7	-1.041	-0.098	-0.001	-0.020	-0.036
8	-0.926	-0.107	-0.001	-0.020	-0.034
9	-0.812	-0.115	-0.001	-0.020	-0.032
10	-0.698	-0.124	-0.001	-0.020	-0.030
11	-0.584	-0.132	-0.001	-0.020	-0.028
12	-0.367	-0.109	-0.441	0.844	1.122
13	-0.059	-0.216	-0.133	-0.361	0.509
14	0.283	-0.319	-0.529	0.750	1.083
15	0.558	-0.218	0.000	-0.020	-0.009
16	0.673	-0.227	0.000	-0.020	-0.007
17	0.787	-0.235	0.000	-0.021	-0.005
18	0.901	-0.244	0.001	-0.021	-0.004
19	1.015	-0.252	0.001	-0.021	-0.002
20	1.129	-0.261	0.001	-0.021	0.000
21	1.244	-0.269	0.001	-0.021	0.002
22	1.358	-0.278	0.001	-0.021	0.004
23	1.472	-0.287	0.001	-0.021	0.006
24	1.586	-0.295	0.001	-0.021	0.008
25	1.700	-0.304	0.001	-0.021	0.010

Table 10: Respondent Scores by Solution for DSA of **F*** (with Criterion Item)

Resp	Group	1	2	3	4	5
1	2	1.020	0.824	0.330	-0.569	-0.576
2	1	0.972	-1.230	-1.264	-1.515	0.400
3	1	0.974	-1.230	-0.033	0.815	-1.600
4	2	1.018	0.812	-0.669	0.261	-1.588
5	2	1.019	0.766	1.470	0.995	0.274
6	2	1.020	0.801	0.908	-0.634	-0.260
7	1	0.968	-1.251	1.589	-0.692	1.201
8	2	1.020	0.801	-0.725	-0.930	0.686
9	1	0.968	-1.280	-0.294	1.305	-0.066
10	2	1.017	0.751	-1.313	1.486	1.537

A look at the first five solutions of the dual scaling analysis of **F** indicates that the demographic/grouping characteristic does not define any of these dimensions. This does not mean that the ratings of the respondents were not influenced by the grouping variable in the analysis, only that the latent effect of this variable does not show up in the first five solutions. With the grouping variable included as the criterion item, however, the dual scaling analysis (of **F***) clearly shows that this item, not unexpectedly, determines the second solution of the analysis. In effect, the grouping variable fixes the axis of the second solution. It appears that Items 1 and 4 "load" most positively on this group-defined dimension, while Item 2 and 5 "load" most negatively. An examination of the original data and the criterion item reveals that the ratings for Items 1 and 4 seem to align with the 1's and the 2's of the grouping variable, and the ratings for Item 2 and 5 seem to align in the opposite direction. The weights of the first six items in the first solution of the DSA of **F*** are similar to the weights of the same items in the DSA of the original data matrix **F**. It is important to note that the item weights would increasingly differ in the DSAs of **F** and **F*** as the ratings for one or more items "lined up" with the elements of the criterion item.

The category boundaries of the first solution of both analyses are ordered, as one would expect. The category boundaries of the second solution of the DSA of **F*** are not ordered, but the category boundaries for the items of the original data (Category Boundaries 12, 13, and 14) are ordered. This is not always the case using this procedure, but as was previously mentioned, in a forced classification analysis the analyst would generally be more concerned with the item weights than the category boundaries. Category Boundary 14, having the highest magnitude, appears to be most affected by the criterion item, albeit negatively. As for the respondent scores in the second solution, they settle into two groups determined by the pattern in the group vector, with Group 1 scores taking on values around -1.2 and Group 2 scores assuming values around 0.8. We now consider the application of this procedure in the analysis of some "real" data.

Student Stress Survey Example

A health survey was administered by the Student Health Center at Rochester Institute of Technology (RIT) asking students about their stress level (see Gadzella (1991)). The rating scaling on the survey had "1" representing "never", "2" representing "seldom", "3" representing "occasionally", "4" representing "often", and "5" representing "most of the time". Students were also asked several demographic questions on the survey (*Gender*, *Employment Status*, *Health Appointment vs. No Health Appointment*, etc.). A convenient subset of twenty respondents who answered the first section of the survey (concerning frustrations) was selected for inclusion in the analyses to follow, which in the case of the forced classification analyses, included the demographic item "*Gender*" (*male* vs. *female*). The output from the DSA of the original data is compared to the output from the forced classification procedure that imposed the effect of the *Gender* variable. The output for each analysis includes the percent of variation accounted for by solution, item weights, and corresponding respondent scores.

From the survey responses, we have

$$\mathbf{F} = \begin{bmatrix} 3 & 3 & 2 & 3 & 2 & 4 & 2 \\ 5 & 3 & 2 & 2 & 2 & 1 & 1 \\ 3 & 3 & 2 & 2 & 2 & 2 & 3 \\ 3 & 3 & 4 & 3 & 2 & 3 & 3 \\ 4 & 3 & 3 & 3 & 3 & 1 & 1 \\ 4 & 3 & 4 & 5 & 2 & 3 & 3 \\ 3 & 3 & 4 & 3 & 2 & 2 & 4 \\ 3 & 3 & 4 & 3 & 2 & 2 & 2 \\ 4 & 3 & 2 & 2 & 1 & 1 & 2 \\ 3 & 3 & 3 & 4 & 3 & 2 & 3 \\ 2 & 2 & 3 & 2 & 2 & 3 & 3 \\ 3 & 3 & 2 & 2 & 3 & 1 & 2 \\ 4 & 4 & 3 & 2 & 2 & 2 & 2 \\ 3 & 3 & 5 & 2 & 3 & 3 & 2 \\ 3 & 3 & 4 & 3 & 1 & 1 & 2 \\ 3 & 3 & 4 & 3 & 4 & 5 & 2 \\ 4 & 3 & 3 & 3 & 4 & 3 & 3 \\ 5 & 4 & 5 & 3 & 4 & 5 & 4 \\ 3 & 4 & 5 & 4 & 2 & 1 & 3 \\ 4 & 4 & 3 & 2 & 2 & 5 & 2 \end{bmatrix} \text{ and } \mathbf{x} = \begin{bmatrix} 2 \\ 2 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 1 \\ 2 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 1 \\ 1 \\ 1 \\ 2 \\ 1 \end{bmatrix},$$

where $\mathbf{F}$ is the matrix of original rating data for the twenty respondents, and $\mathbf{x}$ the *Gender* vector with "1" representing "*male*" and "2" representing "*female*".

"Centered" and augmented by the modified vector $\mathbf{x}^*$, $\mathbf{F}$ becomes

$$
\mathbf{F}^* = \left[\begin{array}{ccccccc|c}
14 & 14 & 13 & 14 & 13 & 15 & 13 & 27 \\
16 & 14 & 13 & 13 & 13 & 12 & 12 & 27 \\
14 & 14 & 13 & 13 & 13 & 13 & 14 & 1 \\
14 & 14 & 15 & 14 & 13 & 14 & 14 & 1 \\
15 & 14 & 14 & 14 & 14 & 12 & 12 & 27 \\
15 & 14 & 15 & 16 & 13 & 14 & 14 & 27 \\
14 & 14 & 15 & 14 & 13 & 13 & 15 & 27 \\
14 & 14 & 15 & 14 & 13 & 13 & 13 & 1 \\
15 & 14 & 13 & 13 & 12 & 12 & 13 & 27 \\
14 & 14 & 14 & 15 & 14 & 13 & 14 & 1 \\
13 & 13 & 14 & 13 & 13 & 14 & 14 & 1 \\
14 & 14 & 13 & 13 & 14 & 12 & 13 & 1 \\
15 & 15 & 14 & 13 & 13 & 13 & 13 & 27 \\
14 & 14 & 16 & 13 & 14 & 14 & 13 & 27 \\
14 & 14 & 15 & 14 & 12 & 12 & 13 & 27 \\
14 & 14 & 15 & 14 & 15 & 16 & 13 & 1 \\
15 & 14 & 14 & 14 & 15 & 15 & 14 & 1 \\
16 & 15 & 16 & 14 & 15 & 16 & 15 & 1 \\
14 & 15 & 16 & 15 & 13 & 12 & 14 & 27 \\
15 & 15 & 14 & 13 & 13 & 16 & 13 & 1
\end{array}\right],
$$

where $\mathbf{F}^*$ includes ten new categories between the "centered" data and extreme values of the *Gender* criterion item. According to procedure, the augmentation of the modified *Gender* vector followed the addition of the centering matrix to the original data matrix $\mathbf{F}$.

The output of the dual scaling analysis of $\mathbf{F}$ follows:

Table 11: Variation by Solution for DSA of $\mathbf{F}$ (Original Data)

Variation	1	2	3	4	5
Pct Total	57.443	14.782	9.733	5.544	4.464
Cum Pct	57.443	72.225	81.958	87.502	91.966

Table 12: Item Weights by Solution for DSA of $\mathbf{F}$ (Original Data)

Item	1	2	3	4	5
1	0.681	-0.132	-1.356	-0.634	0.793
2	0.350	-0.521	-0.678	0.109	1.111
3	0.595	0.274	2.091	-1.199	1.688
4	-0.132	-0.857	1.111	-1.024	-1.737
5	-0.768	0.650	-0.985	-1.568	-1.127
6	-0.609	2.799	0.029	0.666	0.029
7	-0.608	-0.344	1.204	1.872	-0.550

Table 13: Respondent Scores by Solution for DSA of **F** (Original Data)

Resp	1	2	3	4	5
1	0.914	0.966	-0.759	0.999	-1.155
2	0.999	-0.883	-1.677	-0.657	0.585
3	1.021	-0.430	-0.815	2.049	-0.409
4	1.140	0.365	1.258	0.460	0.174
5	1.079	-0.613	-0.961	-1.638	0.402
6	1.025	-0.120	1.199	-0.703	-0.904
7	1.051	-0.492	1.469	0.999	-0.107
8	1.241	-0.384	0.600	-0.630	0.459
9	1.070	-1.127	-0.845	0.967	0.658
10	1.003	-0.638	0.508	-0.567	-2.374
11	0.838	0.851	0.945	2.044	-0.526
12	0.976	-0.724	-1.559	-0.145	-0.759
13	1.188	-0.254	-0.855	0.198	1.345
14	1.039	0.940	0.132	-0.849	1.149
15	1.151	-0.936	0.813	-0.295	0.738
16	0.699	1.993	0.092	-1.133	-0.605
17	0.780	1.525	-0.945	-0.556	-1.536
18	0.647	1.746	0.359	-0.621	1.398
19	1.021	-0.957	1.459	-0.793	0.450
20	0.886	1.477	-0.787	0.825	1.205

The output of the dual scaling analysis of **F*** follows:

Table 14: Variation by Solution for DSA of **F*** (with *Male* vs. *Female*)

Variation	1	2	3	4	5
Pct Total	90.009	8.727	0.391	0.315	0.174
Cum Pct	90.009	98.736	99.127	99.442	99.616

Table 15: Item Weights by Solution for DSA of **F*** (with *Male* vs. *Female*)

Item	1	2	3	4	5
1	0.172	0.002	-0.059	2.353	0.610
2	0.078	-0.004	-0.833	1.169	-0.587
3	0.145	-0.070	0.679	-3.705	1.571
4	-0.046	0.050	-1.173	-1.918	2.058
5	-0.175	-0.481	0.971	1.781	3.477
6	-0.114	-0.714	5.006	-0.018	-1.718
7	-0.153	-0.334	-1.071	-2.092	-2.627
8	0.002	5.696	0.594	0.022	-0.112

<u>Table 16:</u> Respondent Scores by Solution for DSA of **F*** (with *Male* vs. *Female*)

Resp	Group	1	2	3	4	5
1	2	0.999	0.960	1.409	0.762	-1.263
2	2	1.000	1.015	-0.463	1.684	0.448
3	1	1.000	-0.983	-1.304	0.762	-1.663
4	1	1.003	-0.988	-0.286	-1.313	-0.401
5	2	1.002	1.006	-0.236	0.961	1.610
6	2	1.001	0.997	0.362	-1.189	0.513
7	2	1.001	0.991	-0.317	-1.476	-0.907
8	1	1.005	-0.961	-1.012	-0.658	0.665
9	2	1.001	1.018	-0.834	0.840	-1.159
10	1	1.000	-0.974	-1.415	-0.546	1.232
11	1	0.997	-1.019	0.115	-0.983	-1.718
12	1	0.999	-0.974	-1.527	1.517	0.684
13	2	1.004	0.999	0.080	0.837	-0.565
14	2	1.002	0.964	1.377	-0.138	0.537
15	2	1.003	1.020	-0.577	-0.820	0.071
16	1	0.995	-1.038	1.637	-0.104	1.196
17	1	0.997	-1.033	0.981	0.923	0.873
18	1	0.994	-1.026	1.500	-0.357	0.371
19	2	1.000	1.016	-0.566	-1.450	0.690
20	1	0.999	-1.014	1.104	0.756	-1.209

A look at the weights assigned to the items in the dual scaling analysis of the original data matrix **F** suggests that students tend to have experienced more frustration in the areas addressed by the first three questions than in the areas addressed by the last four questions. It appears that students seem to be experiencing higher frustrations in professionally and economically-related areas than they are in socially-related areas. In this "standard" analysis, there is no clear evidence of the effect of *Gender*.

The dual scaling analysis of **F*** forces the second solution to reflect the effect of *Gender* in the analysis (with the first solution paralleling the first solution in the analysis of **F**). All respondents in each gender group are assigned the same score in the second solution, approximately 1.0 for males and approximately -1.0 for females. Among the items, there are none that strongly "load" on this *Gender* dimension in a positive way. Item 4 loads most positively, indicating that females might be slightly more inclined than males to experience frustration at not meeting their goals. Item 6 loads in a highly negative way, indicating that at least among the twenty students selected, males experience a very high degree of frustration in dating. This finding is not surprising since the student population at RIT disproportionately favors males.

Discussion

The proposed procedure allows the user to determine the second solution of a dual scaling analysis so that the effect of a particular demographic (grouping) characteristic constituting the criterion item defines that solution. (As noted earlier, it is the fact that the "forced" dimension shows up in the second solution in this procedure that sets it apart from the other forced classification methods that have the forced dimension showing up in the first solution.) Robust results were obtained when at least ten "new" categories were included. These results are easily interpretable and allow the analyst to identify in the scores the two groups into which the respondents were classified and, more importantly, the specific items that strongly load on the dimension defined by the grouping variable. In this forced dimension, item weights that have the same sign as the weight assigned to the criterion item are positively influenced, while item weights that have the opposite sign are negatively influenced. Except in situations where there are several items whose ratings strongly reflect the influence of the criterion item, the first solution using the proposed forced classification procedure "parallels" (i.e., is very similar to) the first solution of a standard DSA of the original data.

A reasonable "next generation" of this procedure would be one that can enable handle criterion items that are comprised of three or more groupings. (This extension might also require only a single-column group vector **x** to force the solution, but this requires some investigation.) Another generalization of this procedure to consider involves the use of the "centering" matrix **G**. Although **G** was specifically defined in this paper, it is quite possible that a matrix other than **G**, as defined, could be used (e.g., one that is simply a multiple of the unit matrix) rather than one that "centers" the ratings in the data matrix **F** between the two numbers that make up the modified group vector **x***.[4] Adding different multiples of the unit matrix to the original data would almost certainly variously affect the item weighting, but more work is needed to see just what that effect would be.

References

Day, D. A. (1989). Investigating the validity of the English and French versions of the Myers-Briggs Type Indicator. Unpublished doctoral dissertation; University of Toronto, Canada.

Gadzella, B. (1991). *Student-Life Stress Inventory*. Commerce, Texas: Author.

Guttman, L. (1950). The principal components of scale analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfield, S. A. Star, and J. A. Clausen. *Measurement and Prediction*. Princeton: Princeton University Press.

Lawrence, D. R. (2000). Imposing a dichotomous incidence variable on the weighting of rank-ordered items in a dual scaling analysis. Published in the *Proceedings of The 2000 International Conference on Measurement and Multivariate Analysis*; Banff, Alberta, Canada.

Lawrence, D. R. (2006). Getting a dichotomous incidence item to determine the first solution in a dual scaling analysis of paired-comparison data. Paper presented at the 2006 International Meeting of the Psychometric Society; Montreal, Quebec, Canada.

Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and Its Application*. Toronto: University of Toronto Press.

Nishisato, S. (1984). Forced classification: A simple application of a quantification technique. *Psychometrika*, 49, 25-36.

Nishisato, S. (1986). Generalized forced classification for quantifying categorical data. In E. Diday, Y. Escoufier, L. Lebart, J. Pages, Y. Schektman & R. Tomassone (Eds.), *Data Analysis and Informatics, IV* (pp. 351-362). Amsterdam, The Netherlands: North-Holland.

---

[4] As was discussed earlier in this article, "centering" the data produces robust results that are not dependent on the homogeneity of the items.

Nishisato, S. (1988). Forced classification procedure of dual scaling: Its mathematical properties. In H. H. Boch (Eds.), *Classification and Related Methods of Data Analysis* (pp. 523-532). North-Holland: Amsterdam, The Netherlands.

Nishisato, S. (1994). *Elements of Dual Scaling: An Introduction to Practical Data Analysis*. Hillsdale, NJ: Lawrence Erblaum.

Nishisato, S. & Baba, Y. (1999). On contingency, projection and forced classification of dual scaling. *Behaviormetrika*, 26, 207-219.

Nishisato, S. & Gaul W. (1990). An approach to marketing data analysis: The forced classification procedure of dual scaling. *Journal of Marketing Research*, 27, 354-360.

Nishisato, S. & Sheu, W. J. (1984). A note on dual scaling of successive categories data. *Psychometrika*, 49, 493-500.

Odondi, M. (1997). Multidimensional analysis of successive categories (rating) data by dual scaling. Unpublished doctoral dissertation; University of Toronto.