Statistics
in Medicine

Research Article

Received 7 October 2010, Accepted 10 August 2011 Published online 15 November 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4399

A latent class method for the selection of prototypes using expert ratings

William E. Miller[*†]

Latent class analysis can be applied to the outcomes of expert ratings to select objects or subjects that are regarded as prototypical of a category in an ordinal classification system. During a pilot study, Monte Carlo simulations demonstrated that the probability of correct selection is larger when using latent class analysis than when using methods that rely on agreement statistics. Further improvements in the latent class results can also be achieved by applying affine transformations to latent class estimates of sensitivity and specificity. An application is presented that involves the selection of prototypical radiographs. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: Monte Carlo simulation; methodology; affine transformation; ordinal classification; random block design

1. Introduction

Ordinal scales are often used in medical research because of the lack of more precise methods of measurement. For example, a survey of articles in one volume of the *New England Journal of Medicine* in the 1980s showed that about 20% of the articles contained ordinal data [1]. There are also a number of ordinal classification systems that are used in medical diagnosis and research. The motivating example for the present paper is the International Classification of Radiographs of Pneumoconioses [2], which was adopted by the International Labour Office (ILO) to classify radiographic abnormalities that may indicate the presence of silicosis, asbestosis, and other pneumoconioses.

The ILO classification system includes ordinal 'profusion' and size categories where increasing categories correspond to increasingly numerous shadows (for 'small' opacities) or larger shadows (for 'large' opacities) observed in the lungs. Physicians who classify chest radiographs using the ILO system are guided by a set of prototypical standard radiographs. Even when these physicians are specially trained and highly experienced, and even when they have been certified as demonstrating proficiency in using the ILO system, there can be substantial variation in classifications between and within raters using the system, particularly when rating abnormal images [3, 4]. In recent years, the use of film-based radiography imaging has been increasingly superseded in clinical practice by the use of digital radiography imaging [5]. There is a corresponding need to select new ILO standard images and other new digital images for training, proficiency testing, and quality control in a digitally-based ILO classification system for which images are both acquired and displayed digitally. In the past, film-based images were chosen for the ILO system by consultation with and the consensus of a select group of experts. This paper presents a method for providing an independent assessment for selecting images that is less subjective and potentially more accurate.

In the following, the target category is defined as the category for which we are seeking prototypical objects or subjects, and the target probability is defined as the probability of being classified into the target category. For example, the ILO system has four ordinal categories relating to large opacities – 'O', 'A', 'B', and 'C', where the latter three categories are associated with the observation of opacities of

Division of Respiratory Disease Studies, National Institute for Occupational Safety and Health (NIOSH),Centers for Disease Control and Prevention (CDC), 1095 Willowdale Rd., Morgantown, WV 26505-2888, USA

**Correspondence to: William E. Miller, Division of Respiratory Disease Studies, National Institute for Occupational Safety and Health (NIOSH),Centers for Disease Control and Prevention (CDC), 1095 Willowdale Rd., Morgantown, WV 26505-2888, USA.*

†E-mail: wem0@cdc.gov

increasing combined diameter. Suppose that nine raters independently classify each of 200 images and we wish to find prototypes for the second or 'A' category using the resulting ratings. We would like to find a method of selection that has a high likelihood of choosing the images that are most representative of that category and, given the variability between experts, we cannot assume that the contribution of each rater has the same accuracy or precision. Therefore, our search for an optimal method for selecting prototypical 'A' images can, in theory, be separated into the following two goals: (1) identifying a good score or metric (i.e., a measure of how close a sample of classifications for an image is to the target category) and (2) finding a good method of weighting the contribution of each rater to that score. Other relevant issues can include the coding of the data, the choice of the target category (e.g., whether we are selecting prototypes for middle or boundary categories of a classification), the number of categories in the classification, and the scaling of the categories (e.g., whether we assume a uniform distance between adjacent categories).

Latent class methods for the ranking and selection of prototypes will be examined in this paper. Ranking and selection procedures have also been developed in the area of engineering [6], but they are normally applied to objective measurements to select a best process or a best subgroup. Zhou and Lange [7] developed a method that considers the variability among raters, but their approach has been designed for large public or website surveys with 100,000 or more ratings and thousands of raters and objects to be rated. Such data typically include many missing rater–object combinations. Their binomial model indicates whether a rater reaches consensus or nonconsensus decisions and could be applied to a variety of rating studies. On the other hand, it is also important for researchers to be aware that customizing their methods to the problem at hand can sometimes lead to greater improvements in their results than a reliance on machine-learning and other algorithms. The solution for the particular problem presented in this paper takes advantage of the fact that we have a large amount of information for each expert rater and very few, if any, missing rater–object combinations. This solution also recognizes that measures of sensitivity and specificity are more conclusive than measures of agreement. Measures of sensitivity and specificity are essential to determining the validity of ratings, especially because inferences can differ when we use measures of agreement [8]. The uniqueness of the latent class approach is due to the primacy that it puts on the sensitivity and specificity estimates for raters in the ranking and selection of objects.

Three latent class approaches will be assessed using Monte Carlo simulations. Details will be shown for classifications with five categories, which allow us to investigate patterns that could result from more complicated classifications. The main results for classifications with two and three categories will also be discussed. The design of the simulations are complicated by the fact that they will need to incorporate differences among both the rated objects and the raters, and be able to characterize to some extent the dependence structure for raters who are classifying the same objects. While it is impossible to examine all possible conditions, the simulations are designed to reflect covariance structures which were observed in practice for experts using the ILO classification system.

2. Latent class selection

The latent class model assumes that the association between some observed categorical variables can be explained by the existence of mutually-exclusive latent classes. In other words, the model assumes that there is a conditional or local independence between the observed variables, given the assignment of objects or subjects to their latent classes. Latent class analysis (LCA) is often viewed as a type of factor analysis for categorical data, where the parameters include the latent class probabilities and some conditional probabilities that indicate how strongly related the observed variables are to the latent classes. However, as shown by Clogg [9] and Rindskopf and Rindskopf [10], the latent class model also has a natural and direct interpretation with respect to rater agreement and medical diagnosis, where each observed variable now represents the results from an individual rater.

Recent radiographic trials at the National Institute for Occupational Safety and Health (NIOSH) involved nine experienced raters. When there are two latent classes, the conditional probabilities are interpreted as estimates of each rater's sensitivity, specificity, and their complements (i.e., the false negative and false positive rates). For two latent classes and nine raters, one form of the latent class model can be written as

$$\pi_{r_1 \ldots r_9} = \sum_{c=1}^{2} \pi_c \pi_{r_1 \ldots r_9 | c} \tag{1}$$

where

$$\pi_{r_1 \ldots r_9}|c = \prod_{i=1}^{9} \pi_{r_i}|c \tag{2}$$

and where $\pi_{r_1 \ldots r_9}$ is the unconditional joint probability for a pattern of responses from the nine rates, π_c is the latent class probability, $\pi_{r_1 \ldots r_9}|c$ is the conditional joint probability for a pattern of responses, and $\pi_{r_1}|c$ is the conditional probability of the response from rater i, given that an object belongs to latent class c. More details about the estimation and the formal model can be found in [9, 11].

A distinguishing feature of the latent class model is that it can provide estimates of sensitivity and specificity in the absence of a gold standard. One requirement is that there should be at least three raters to avoid any identification problems in the model [9]. Collins and Lanza [12] and Chung *et al.* [13] address other practical considerations, such as the model fit, the cross-validation of results, and possible inconsistencies in the labeling of latent classes.

All the latent class models in this paper were estimated using the freeware procedure PROC LCA [14], which was written to work with SAS software [15]. Before fitting the latent class model, the ratings for each object were recoded into a single record of binary outcomes, where the assignment to one of the two outcomes depended on whether or not the rating was equal to the target category. This coding also identified the target category with one of the two latent classes. Using Bayes' theorem, the model estimates were then used to calculate the posterior class probability, which estimates the probability that an assignment of an object to a latent class is correct. For example, the posterior class probability that an object belongs to the first latent class can be written as

$$\pi_{c=1|r_1 \ldots r_9} = \frac{\pi_{c=1}\pi_{r_1 \ldots r_9|c=1}}{\sum_{c=1}^{2} \pi_c \pi_{r_1 \ldots r_9|c}} \tag{3}$$

where $\pi_{c=1|r_1 \ldots r_9}$ is the conditional probability of membership in the first latent class, given an object's pattern of responses from the nine raters. Because there are only two latent classes, the posterior class probability that an object belongs to the second latent class is then equal to the complement of $\pi_{c=1|r_1 \ldots r_9}$. The use of the posterior class probability provides the assignment procedure (also referred to as the modal latent class assignment) which, according to Goodman [16], minimizes the number of incorrect assignments for the latent class model.

For the simulations, we compare a standard LCA approach with two other approaches. Therefore, results will be shown for the following:

- Standard LCA. After fitting the standard latent class model to the ratings, a selection of prototypes is made based on the ranking of the posterior class probabilities.
- Affine Transformation. The standard latent class model is used. However, before calculating the posterior class probabilities, the estimates of the conditional probabilities are subjected to an affine transformation [17], that is, a transformation in location and scale that preserves the ratios of distances between values so that, for instance, the mean of the original values is transformed to the mean of the new values. The approach here utilizes PROC LCA together with SAS/IML code [18] to apply an affine transformation that converts a given range of values to a new range of values. As shown in [19], this affine transformation can be defined as follows:

$$NewValue = \left[(OldValue - OldMinimum) \times \frac{NewRange}{OldRange} \right] + NewMinimum \tag{4}$$

- Fuzzy LCA. An ordinal classification can be said to differ from a nominal one with respect to the relevance of classifications into those categories that are adjacent to the target category. Some of the results of the pilot study suggested that, although this was not always true, the results were often improved by considering the classifications into categories that are adjacent to the target category. Therefore, we give some consideration for the ordinal scale of the data by incorporating this information into a nonstandard 'fuzzy' latent class model with an additional nine binary variables, where the binary assignment now depends on whether or not a classification is equal to a category that is adjacent to the target category.

The application of affine transformations will allow us to tune the results, so that readers with large false positive or false negative estimates will have more influence on the calculation of the posterior class

probabilities. Although this introduces some bias into the estimates, it can also lead to increased discrimination and lower misclassification. After deriving the estimates of the sensitivities, the specificities, and their complements using the LCA model and before the application of Equation (3) to estimate the posterior class probabilities, the affine transformation of Equation (4) is applied independently to either the set of false negative estimates (and their complements) or to the set of false positive estimates (and their complements). In the first case, the false positive values are left untransformed, but four different transformations are applied to the false negative estimates:

(a) Transforming the false negative values to the single point value of 0.5 (i.e., this is the special degenerate case where the scale value is zero and all the values are collapsed to a single point);
(b) Transforming the values to range between 0.5 and 0.6;
(c) Transforming the values to range between 0.5 and 0.7; and
(d) Transforming the values to range between 0.5 and 0.9.

In the second case, the false negative values are left untransformed, but the false positive values take on one of the four ranges just described. Note that in transformation (a) above, when all the values are collapsed to the single point value of 0.5, we are essentially nullifying or suppressing the effect of the false negative or false positive estimates on the calculation of the posterior class probabilities.

3. Monte Carlo simulation

The search for a method for selecting prototypes includes the two goals of identifying a good score or metric and finding a good method of weighting the contribution of each rater to that score. If we focus on only the first goal of determining a good score or metric for selecting prototypes for a classification with five categories, then there is a relatively simple approach to simulating ratings. Following the notation of Penny and Jolliffe [20], we can denote our data as

$$50\%A(p_1 - p_5) + 50\%B(q_1 - q_5) \tag{5}$$

which indicates that 50% of the data is to be generated by the first multinomial distribution $A(p_1 - p_5)$ where p_i is the probability of being classified into the i^{th} category, and where $A(p_1 - p_5)$ is being used to generate the ratings for the prototypes. The other 50% of the data is to be generated by the second multinomial distribution $B(q_1 - q_5)$, which is used to generate the ratings for the nonprototypes. The target probability will always be larger for the first multinomial distribution. We can then generate observations for 100 objects for each of the two multinomial distributions, use the score given for each of our methods to rank the objects with respect to how close they are to the target category, and then select the 100 highest-ranked objects using each method. The outcome for each method is then defined as $P(Correct\ Selection)$, the proportion of the objects generated by the first multinomial distribution which are among the 100 highest-ranked objects.

Misclassification will likely increase when many of the rated objects are close to the cutpoint or diagnostic threshold between two categories. We can use this fact to choose multinomial distributions that challenge the various methods to be compared. Pilot studies showed that there were trivial differences between the methods with respect to $P(Correct\ Selection)$ when the second multinomial distribution had its highest probability at a category other than the target category. It is also advisable, when we generate the ordinal data, to avoid complex modeling assumptions, such as the assumption that the underlying continuum for the latent trait follows a certain probability distribution [21]. At the same time, we want to approximate multinomial patterns which are observed in practice. These requirements led to deriving the following equation, which defines the probability of a classification into the i^{th} category as

$$prob_{ith\ category} = \frac{target}{ratio^k} \Big/ \sum_{i=1}^{n} \frac{target}{ratio^k} \tag{6}$$

where n = the number of categories, *target* = the target category, $k = |i - \text{target}|$, and where the value of the *ratio* parameter determines how peaked the distribution will be at the target category. For example, when there are five categories with the third category being chosen as the target category, we have the following multinomial probabilities for various values of the ratio parameter:

	Category				
Ratio parameter	1	2	3	4	5
1.0	0.20	0.20	0.20	0.20	0.20
2.0	0.10	0.20	0.40	0.20	0.10
4.0	0.04	0.15	0.62	0.15	0.04

Equation (6) provides a concise and systematic way to specify a set of multinomial probabilities for a target category, and it also allows us to challenge the methods by examining scenarios where subtle differences exist between the prototypes and the nonprototypes. By using the ratio parameters to specify the multinomial probabilities together with the RANTBL function found in SAS software [22], we can generate classifications for Monte Carlo simulations which compare various scores. This was the objective of the earliest pilot studies.

The above approach will lead to generating independent classifications. To address the second goal of finding a good method of weighting the contribution of each rater, we also need to incorporate rater differences and a covariance structure into our simulations. One way to do this is to generate individual samples for each rater by first specifying the ratio parameter for five sets of multinomial probabilities and then using a mixture of these five sets, so that the sample can be denoted as

$$\sum_{i=1}^{5} a_i \% M(p_{i1} - p_{i5}) \tag{7}$$

where α_i is the percentage of the sample that uses the i^{th} multinomial distribution. A covariance structure is introduced into the simulations by requiring the largest of the five p_{ij} probabilities for the i^{th} multinomial distribution to be given for category $j=i$. In other words, the target variable is set at the i^{th} category for the subsample generated by the i^{th} multinomial distribution. However, it is the choice of the α_i percentages, also found using Equation (6), which determines the target probability for the full sample. The result is that the square contingency table for two replicate samples reflects varying degrees of agreement, based on the number of entries at or near the diagonal of the table.

During each trial of a simulation, Equation (7) is used to generate two different samples for each rater, one for each set of 100 prototypes and nonprototypes. If the sample is being generated for the first 100 prototypical objects, then the expected sensitivity is equal to the target probability for the sample. If the sample is being generated for the second 100 nonprototypical objects, then the expected specificity is equal to the complement of the target probability. While the overall likelihood of being selected, pooled over all the raters, is always larger for the first 100 objects, we can challenge the methods of weighting the rater contributions by having two groups with five and four raters, where the first group of five raters generally has better expected sensitivity, specificity, and within-group agreement than the second group of four raters. The expected sensitivity is defined for values between 30% and 70%, whereas the expected specificity is defined for values between 40% and 80%. (Note that an examination of patterns in ordinal classifications for actual rating data suggest that specificity tends to be higher than sensitivity).

Levels for the simulations, referred to here as factors, are defined in terms of the expected differences in sensitivity and specificity for the two groups of raters. Table I displays the 16 factor combinations for the differences that are examined in this paper. For example, the factor combination '20/20' indicates that the sensitivity and specificity are both expected to be 20% larger, on average, for the first group of raters than for the second group of raters. Table I also shows that the overall target probabilities for the two groups of 100 objects are fairly constant over the factor combinations, with the first group of 100 objects being, on average, about 10 to 14% more likely to be classified into the target category. For these factor combinations, the expected within-group kappa statistics for the raters range from about 0.10 to 0.40. The ratio parameter values are also chosen so that the expected chance agreement is constant over the categories.

Because there are 200 objects and nine raters in our scenario, then each trial of a simulation has 1800 microreplications, and there are 500 trials or macroreplications for each factor combination. The final result for a factor combination is the mean of $P(\textit{Correct Selection})$ over the 500 trials, or the mean proportion of correct selection. The results for different methods are then compared by calculating the mean difference over the trials within a factor combination, where the factor combinations function as

Table I. Design matrix with the factor combinations that are associated with the expected differences in sensitivity and specificity for the two groups of raters, together with the expected target probabilities, pooled over the raters, for the two groups of 100 objects.

	Expected sensitivities		Expected specificities		
Factor combination[a]	Group 1	Group 2	Group 1	Group 2	Target probabilities for two sets of objects
0 / 0	0.50	0.50	0.60	0.60	0.50 / 0.40
0 / 10	0.50	0.50	0.65	0.55	0.50 / 0.39
0 / 20	0.50	0.50	0.70	0.50	0.50 / 0.39
0 / 40	0.50	0.50	0.80	0.40	0.50 / 0.38
10 / 0	0.55	0.45	0.60	0.60	0.51 / 0.40
10 / 10	0.55	0.45	0.65	0.55	0.51 / 0.39
10 / 20	0.55	0.45	0.70	0.50	0.51 / 0.39
10 / 40	0.55	0.45	0.80	0.40	0.51 / 0.38
20 / 0	0.60	0.40	0.60	0.60	0.51 / 0.40
20 / 10	0.60	0.40	0.65	0.55	0.51 / 0.39
20 / 20	0.60	0.40	0.70	0.50	0.51 / 0.39
20 / 40	0.60	0.40	0.80	0.40	0.51 / 0.38
40 / 0	0.70	0.30	0.60	0.60	0.52 / 0.40
40 / 10	0.70	0.30	0.65	0.55	0.52 / 0.39
40 / 20	0.70	0.30	0.70	0.50	0.52 / 0.39
40 / 40	0.70	0.30	0.80	0.40	0.52 / 0.38

[a]For example, the factor combination of '10/20' indicates that the sensitivity is expected to be 10% larger for the first group of raters than for the second group of raters, and the specificity is expected to be 20% larger for the first group of raters than for the second group of raters.

the blocks in a randomized block design. There are also five possible target categories for each factor combination. However, because of the symmetry of the multinomial distributions, which are defined using Equation (6), the results for the first and second categories will be practically identical to those for the fifth and fourth categories, respectively. Therefore, we only examine the results for the first three categories. The Monte Carlo simulations are all performed using SAS software (simulation code available from the author).

The key assumption of latent class analysis is that of conditional or local independence. According to Brenner [21], this assumption pervades much of the work in diagnostic measurement, and Uebersax [23] has argued that conclusions about the performance of the kappa statistic also involve this assumption. Raters using the ILO system have similar specialized medical training, and the simulations here were developed to mimic the levels of agreement found among raters in past radiographic trials. These substantial levels of agreement will introduce a dependence structure that violates the assumption of conditional independence. Because the simulations were designed so that higher levels of agreement are also associated with higher levels of sensitivity and specificity, this dependence will generally be stronger in the first group of raters than in the second group. Torrance-Rynard and Walter [24] discuss some situations where the violation of this assumption leads to biased estimation for LCA models. We will briefly investigate the potential of this inter-rater correlation for influencing the misclassification by using Equation (6) together with the RANTBL function to generate classifications for raters who are, for all practical purposes, independent (i.e., with kappas close to zero). These raters will nevertheless still retain the same levels of sensitivity and specificity for a factor combination.

4. Simulation results

During each trial of the simulations to assess each of the three latent class approaches, the outcome for $P(\textit{Correct Selection})$ was found after ranking the 200 objects with respect to the posterior class probability. The Appendix A shows some selected results from a pilot study to compare the method using the standard latent class model to some other methods that each use a different score for ranking the objects. For the following results, there was no evidence of an important interaction between the target categories and the factor combinations. Therefore, only an overall mean, pooling over the target categories, is reported for each factor combination.

4.1. Results using standard LCA

The second column of Table II shows the mean proportions using standard LCA for the 16 factor combinations, stratified by whether the factor combinations are associated with equal sensitivity and specificity differences for the two groups of raters, larger sensitivity differences, or larger specificity differences. As can be seen, the mean proportion of correct selection varies from about 50%, when there are small differences between the two groups of raters, to over 80% when there are large differences with respect to both sensitivity and specificity.

Figure 1 shows the relationship between the expected and estimated false positive and false negative rates using the standard LCA model for the 16 factor combinations and the two groups of raters. In general, the true rates are underestimated (i.e., biased downward) by the latent class model, resulting in an overestimation of the sensitivity and specificity. Some preliminary results showed that inverse regression or calibration [25] could be used to correct this bias, resulting in a general improvement in the mean proportion of correct selection. The benefits of such an approach would be unclear in practice, because it would assume that we know the relationship between the true and estimated rates in general, not just for our selection of factor combinations. Subsequent investigations showed that similar improvements could be produced by applying the affine transformations discussed in the next section.

4.2. Results using affine transformations

An examination of the average results in Figure 2 for the four affine transformations indicates that, when the expected differences in the sensitivities are as large as, or larger than, the expected differences in the specificities, the maximum result occurs between the second and third affine transformations of the false positive estimates, that is, when those estimates are constrained to a range between 0.5 and 0.6 or between 0.5 and 0.7. The best affine transformation of the false negative values under the same conditions would be to use the fourth transformation, that is, to constrain those estimates to range between 0.5 and 0.9. When the expected differences in the specificities are larger than those for the sensitivities, it appears that any of the four transformations of the false negatives achieves a result that is close to maximum. These results suggest the following rule-of-thumb (RT):

Table II. Mean proportions of correct selection based on standard LCA and RT[a] methods for specified factor combinations, stratified by whether there were equal sensitivity and specificity differences for the two groups of raters, larger sensitivity differences, or larger specificity differences.

	Mean *P*(*Correct Selection*)		
Factor combination	Standard LCA	RT method	Mean RT – LCA difference 95% CI
Equal sensitivity/specificity differences			
0 / 0	0.580	0.572	(–0.009, –0.007)
10 / 10	0.611	0.622	(0.010, 0.012)
20 / 20	0.679	0.700	(0.021, 0.023)
40 / 40	0.811	0.826	(0.014, 0.016)
Larger sensitivity differences			
10 / 0	0.498	0.517	(0.017, 0.020)
20 / 0	0.597	0.634	(0.035, 0.038)
40 / 0	0.649	0.737	(0.085, 0.089)
20 / 10	0.634	0.667	(0.032, 0.034)
40 / 10	0.694	0.761	(0.065, 0.068)
40 / 20	0.737	0.782	(0.044, 0.046)
Larger specificity differences			
0 / 10	0.598	0.584	(–0.015, –0.013)
0 / 20	0.624	0.605	(–0.021, –0.017)
0 / 40	0.692	0.716	(0.022, 0.025)
10 / 20	0.649	0.659	(0.009, 0.012)
10 / 40	0.728	0.754	(0.025, 0.027)
20 / 40	0.755	0.778	(0.021, 0.024)

CI, confidence interval.
[a]RT procedure described in Section 4.2

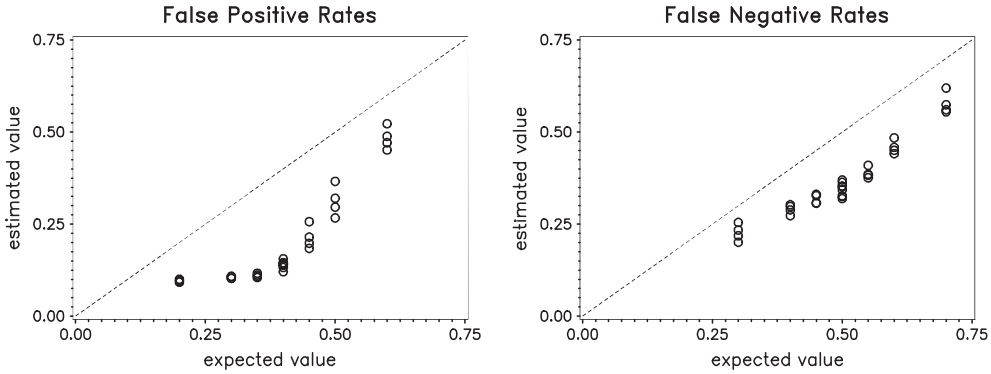

Figure 1. The estimated false positive and false negative rates are plotted against their expected values for the 16 factor combinations and two groups of raters.

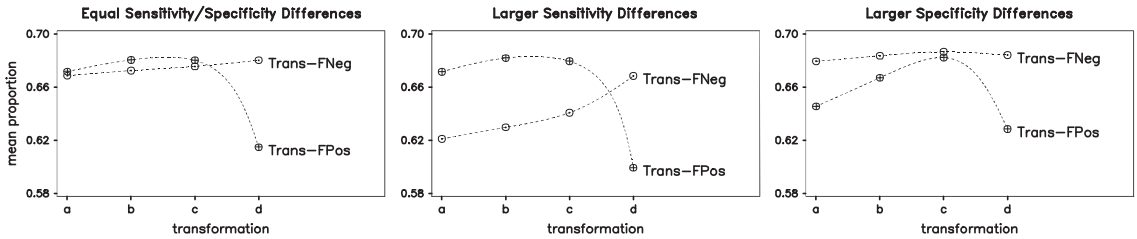

Figure 2. The mean proportions of correct selection using various affine transformations are plotted for the factor combinations that are associated with equal sensitivity and specificity differences for the two groups of raters, larger sensitivity differences, or larger specificity differences. Results are shown using the four affine-transformations of either the false negative rates ('Trans-FNeg') or the false positive rates ('Trans-FPos'), as described in Section 2.

Rule-of-thumb procedure:

a) If the estimated range for the false negatives is as large as, or larger than, the estimated range for the false positives, we then apply an affine transformation to constrain the false positive estimates to range between 0.5 and 0.65.
b) Else, if the estimated range for the false positives is larger than the estimated range for the false negatives, we then apply an affine transformation to constrain the false negative estimates to range between 0.5 and 0.9.

Note that this approach is relatively robust, insofar as it does not assume any knowledge about the exact relationship between the estimated and true false positives and false negatives rates. Although other results suggested that the estimates of the ranges were relatively unbiased, this approach also does not require any rigorous assumptions about the relationship between the estimated and true ranges.

The mean proportions of correction selection using the RT procedure and the 95% confidence interval for the mean difference (after matching on trials) between the RT and standard LCA methods are also shown in Table II. As can be seen, the RT method performs well except when the expected range in specificities is close to zero and the expected range in sensitivities is less than or equal to about 20%. Although the details are not shown here, there was no evidence that applying linear transformations to both the false positive and false negative estimates would further improve the results.

4.3. Results using fuzzy LCA

When the fuzzy latent class model with 18 variables was applied, we generally found a small improvement over the standard LCA results, but the results were still inferior to those that used affine transformations. For example, Table III shows the evolution in the results for the factor combination which defines a 40% difference in sensitivity and a 10% difference in specificity. The mean proportions of correct selection are, respectively, 0.694 and 0.761 for the standard LCA result and for the result of applying an

Table III. Mean proportions of correct selection using various methods for the factor combination where the expected sensitivity and specificity for the first group of raters are 40% and 10% larger, respectively, than for the second group of raters.

Method used	Mean P(*Correct Selection*)
Standard LCA	0.694
Transforming false positives using the RT procedure ('Trans-Fpos')	0.761
Fuzzy LCA	0.699
Combining Trans-FPos and Fuzzy LCA	0.707
Transforming Fuzzy LCA to suppress adjacent-category information	0.761

affine transformation to the false positive estimates using the RT procedure. The use of fuzzy LCA does show a small improvement over the standard LCA result, from 0.694 to 0.699, and a further improvement to 0.707 is achieved by using fuzzy LCA together with the affine transformation for the original nine variables but with no constraints on the estimates associated with the new fuzzy variables. However, this is still smaller than the result of 0.761, and the additional application of affine transformations to the fuzzy LCA variables achieves a maximum of about 0.761 by setting the estimates associated with the fuzzy variables all to 0.5. In other words, this is equivalent to nullifying the new part of the model.

4.4. Results assuming various inter-rater dependence structures

Figure 3 shows the simulation results for four inter-rater dependence structures: (i) where the dependence structure is the one assumed by the previous results; (ii) where the raters in the first group are dependent, but the raters in the second group are independent; (iii) where the raters in the second group are dependent, but the raters in the first group are independent; and (iv) where all the raters are independent. Results are shown for a few of the more extreme factor combinations: '0/0', '40/0', '0/40', and '40/40'. Factor combinations that are associated with moderate differences, such as the '20/10', '10/20', and '20/20' factor combinations, had patterns that were similar to those found for the '40/0'

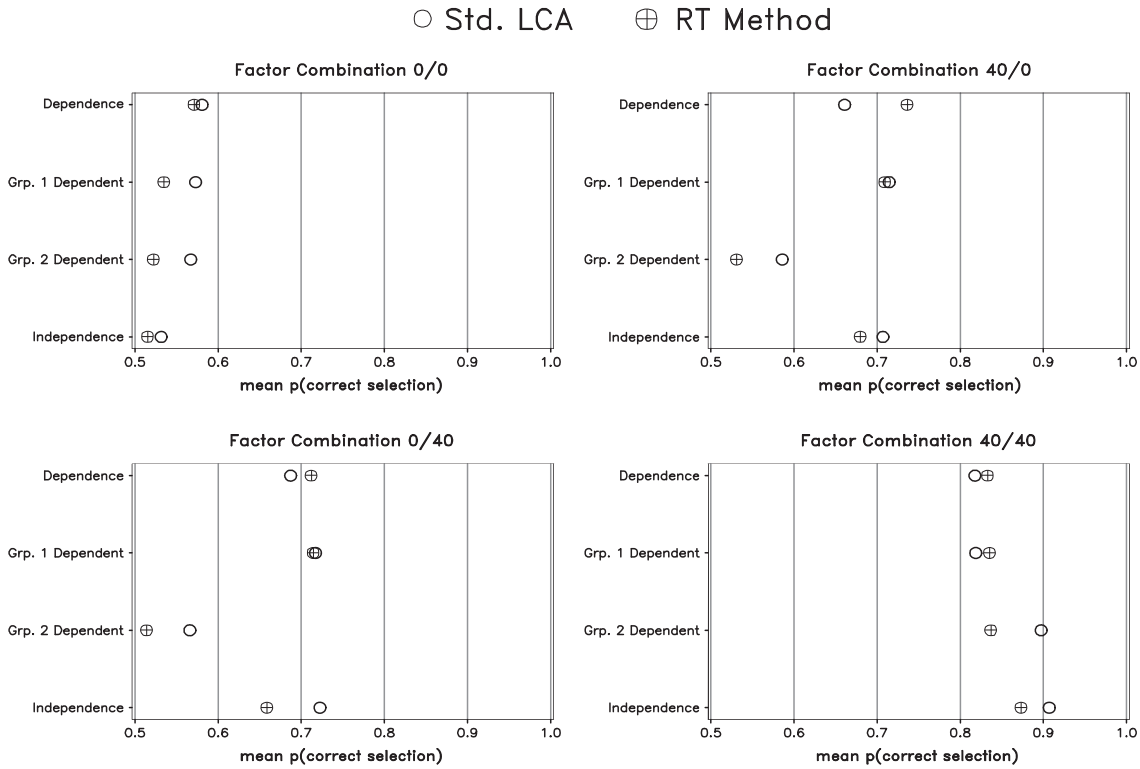

Figure 3. The standard LCA and RT mean proportions of correct selection for the second target category are plotted for four of the factor combinations and four inter-rater dependence structures.

or '0/40' factor combinations. Results are only shown here for the second target category, but the pattern of results was similar for the other target categories. For the '0/0' factor combination (i.e., where there are no differences between the two groups with respect to sensitivity, specificity or agreement), the misclassification generally increases when there are independent raters. However, for the '40/40' factor combination, when there are both large sensitivity and specificity differences, the misclassification decreases for independent readers. For the other two factor combinations, the misclassification is much worse when all the independent raters belong to the first group. In general, the results tend to be better using the standard LCA method when all the raters are independent.

5. Example

One important application is in the area of radiology. Recently, NIOSH conducted a trial to choose prototypical radiographs for training purposes. The aim of the trial was to select the prototypes from a pool of hundreds of radiographic films using classifications from nine experienced users of the ILO system. When the LCA results for the NIOSH data were examined for some important categories of the ILO classification system, it was found that the range of the estimated sensitivities was always larger, and sometimes much larger, than the range of the estimated specificities. In addition, the estimated range of the sensitivities was never less than 0.20. Therefore, for this particular application, the RT procedure should perform better than the standard LCA method in its ability to find candidate images.

As mentioned, the large opacity profusion classification consists of the four ordered size categories of 'O', 'A', 'B', and 'C'. The RT method was applied to the results of classifications of 153 radiographic films by the nine raters to determine the best category 'A' candidates. For these classifications, the false positive rate estimates ranged from 0 to about 0.03, whereas the false negative rate estimates ranged from about 0.32 to 0.77. Therefore, the false negative estimates (or, equivalently, the sensitivity estimates) will have much more influence on the calculation of the posterior class probabilities. The estimated false negative rates for the nine raters are shown in Table IV, along with their classifications for Film #132 and Film #64. For the results using the RT method, Film #132 was the highest ranked candidate (i.e., ranking the films by the resulting posterior class probabilities), whereas Film #64 was ranked 11^{th} highest, even though Film #64 was classified into the 'A' category more often than Film #132. However, two of the six 'A' classifications for Film #64 belong to the raters with the estimated highest false negative rates (i.e., lowest sensitivities), whereas all five of the 'A' classifications for Film #132 are from the raters with the five lowest false negative rates.

The standard LCA approach uses the same false positive and false negative estimates as the RT method, but without the affine transformation, to calculate the posterior class probabilities. When this approach was applied to the data, Film #64 and Film #132 were tied with other films for the highest rank, so it might be necessary in the selection process to use some other measure as a tie-breaker. There were even more films tied for the highest rank when the fuzzy LCA method was applied to the data. In other words, Film #64 and Film #132 are equally likely candidate images with respect to these two methods, but the results using the RT method suggested that Film #132 was much more likely to be a prototypical 'A' film than Film #64. In this instance, the RT method provided more discrimination.

Table IV. Selected results from a trial to determine the most representative 'A' category films among 153 radiographs classified for large opacity profusion by nine raters.

	Rater number								
	1	2	3	4	5	6	7	8	9
False negative rate	0.37	0.40	0.37	0.54	0.32	0.47	0.37	0.59	0.77
	Large opacity profusion classifications								
Film #132	A	A	A	B	A	B	A	B	B
Film #64	A	O	A	O	A	O	A	A	A

The estimated false negative rates are shown for the nine raters, along with their classifications for Film #132 and Film #64. The results using the standard LCA and the fuzzy LCA methods indicated that these two films were indistinguishable as candidates, but the RT method ranked Film #132 as a better candidate image.

6. Discussion

Details for the Monte Carlo simulations for classifications into two and three categories are not shown here, but the three-category results were similar to those for five categories. As with the three-category and five-category simulations, the results for classifications into only two categories still showed that the standard LCA method was superior to the other methods that were examined during the pilot study, but the differences between the methods were smaller. However, there was little or no improvement in the two-category results when affine transformations were applied to the LCA estimates. Therefore, for classifications with only two categories, the standard LCA method would be recommended. It is possible that this result is related to a principal feature of ordinal ratings. When there are three or more categories for a classification, it is reasonable to assume that it would normally be easier for a rater to say an object does not belong to the target category than to say that it does. For example, if the rating process was random for a classification with three categories, then the expected sensitivity and specificity would be about 33% and 67%, respectively, and for five categories, the same expected values would be 20% and 80%. In other words, with three or more categories, it would generally be easier for all the raters to achieve similar levels of high specificity, resulting in a smaller range of specificities than of sensitivities. However, if a classification has only two categories and the ratings are randomly distributed, the expected sensitivity and specificity would both be 50%, so the estimates of sensitivity and specificity should generally be less distinct and have similar ranges.

The results in Section 4.4 demonstrate that the latent class method can be useful, even when there is conditional dependence. They also indicate that the nature of the dependence will influence the results. We would expect less experienced raters to have smaller or more variable agreement and, therefore, be more independent, but increased independence might also occur if the training of the raters was quite variable, so that their ratings reflect different attributes of the rated objects. If the raters are generally independent, then the standard LCA approach should be used. Another possible approach is to model the conditional dependence. Dendukuri *et al.* [26] did this by adding random effects to adjust for conditional dependence, but they also reported small changes in the sensitivity and specificity estimates using their model. Further work needs to be carried out to better understand how various methods perform when this assumption is violated.

Although the coding of an ordinal outcome into a binary one for a latent class model represents a loss of information, it also provides a simple interpretation for the LCA model with two latent classes. The attempt to incorporate the adjacent-category information into the fuzzy LCA model did not improve the results beyond that found by using affine transformations. Campbell *et al.* [27] found no advantage for ordinal models when the purpose was classification. This suggests that, although the additional information contained in ordinal outcomes might improve the ability to discriminate between objects in other circumstances, it does not necessarily highlight the differences that improve our ability to rank the objects with respect to the target category. However, there are other latent class models for ordinal data that could be further investigated [28, 29]. Another possible approach is the use of Rasch models [30].

Just as in laboratory experiments, it is often necessary for computer simulations to simplify the conditions of the problems that are being studied. The initial scenarios for the simulations, which compared methods for ranking two homogeneous sets of 100 objects, made it relatively simple to evaluate the different methods and to later incorporate rater variability into the simulations. Future work could use a more complicated approach, such as having different ratio values for each of the objects and then evaluating the methods using Kendall's Tau, while at the same time examining whether some procedures are better able to determine the correct number of prototypes that should be selected. Future work will also need to examine the influence on the results of the number of raters that are available for a study, and to investigate results for more complicated patterns of inter-rater variability and the possible advantages of having repeated classifications of objects from the same group of raters. There may also be an advantage for first subsetting the data to focus the selection on the objects that are the most likely candidates by, for example, excluding objects that have few classifications into the target category. One could argue that the variation in such a subset is more relevant, especially if the rater variation depends on the category of a classification. On the other hand, by reducing the size of the data set, we could be providing much less information about the overall sensitivities and specificities of the raters.

The RT procedure is simple and it works well except for a few factor combinations. Further work, not shown here, demonstrated that changes to the RT procedure could be found, which improved the results for these few factor combinations, but only at the cost of diminishing the results for the other factor combinations. However, future work might be able to improve this situation by, for example, using

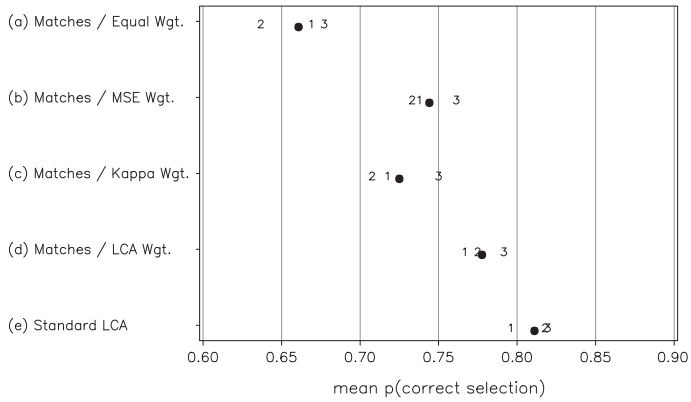

Figure A.1. Pilot study results are shown for the mean proportions of correct selection for the three target categories and their overall mean (plotting symbol = •) for the factor combination in Table I which is associated with the largest differences in sensitivity and specificity for the two groups of raters. Results are shown using the fuzzy-matches method together with four different rater weights and for the standard LCA method.

discriminant analysis to identify clustering statistics, which could in turn improve the ability to better discriminate between the various conditions represented by the factor combinations. Further refinements in the methods could also include contributions from the area of classification and from strategies that use multinomial distributions [31].

The overall aim of this work was to find a method that performed well over the range of conditions represented by the factor combinations. Therefore, any conclusion for a single factor combination applies only to a narrow inference space with respect to this aim. In general, the application of the LCA method to a sample of ratings provides an independent assessment for the selection of prototypical objects, and a further refinement using affine transformations generally improves the selection. In addition, although the simulations were designed with ordinal classifications in mind, the methods might also be applicable to classifications on a nominal scale.

APPENDIX A. Selected results for pilot study

A pilot study was conducted which first looked at 10 different metrics or scores used to rank objects with respect to their suitability as prototypes. It then examined a subset of these metrics in combination with six different methods of weighting the raters. The metrics included the absolute sample bias, the mean-squared error, the mean-absolute error, and an approach that used correspondence analysis. The choice of rater weights included a function of the weighted kappa statistic and a function of the mean-squared error, which has properties similar to those of the concordance statistic given by Lin [32].

A measure of similarity that worked well in a variety of situations was fuzzy matching. Simple matching counts the number of times in a sample when a rating coincides with the target category. Fuzzy matching simply extends this by adding to this count a value of 1/2 whenever a classification coincides with a category that is adjacent to the target category. A larger sum for an object would indicate that it is a better candidate, and so objects are then ranked from the highest to lowest values of this sum. Note that the use of fuzzy matching takes some account of the ordinal scale for a classification.

Figure A.1 presents some selected Monte Carlo results for the last factor combination shown in Table I of the text. Results are shown for fuzzy matching with equal weighting and using other weighting. They indicate that the method of weighting that worked best with fuzzy matching was a function of the false positive and false negative estimates from LCA. However, the best results, overall, came from using standard LCA. In general, the pilot study suggested that the latent class results for various factor combinations were superior to those using the other scores and weights that were considered.

Acknowledgements

The author thanks Lee Petsonk for the many discussions related to this paper, and is grateful to Michael E. Andrew and Kathleen Fedan of NIOSH, Gerry Hobbs, and Robert M. Castellan for their comments and suggestions. The author is indebted to James T. Wassell of NIOSH for his valuable suggestions on the presentation of

the methods. The author also thanks the referees for their comments and, in particular, for their suggestion to include an example in the paper.

Disclaimer: The findings and conclusions in this report are those of the author and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

References

1. Moses LE, Emerson JD, Hosseini H. Analyzing data from ordered categories. *New England Journal of Medicine* 1984; **311**:442–448.
2. International Labour Office. *Guidelines for the Use of the ILO International Classification of Radiographs of Pneumoconioses, Revised Edition 2000*, Occupational Safety and Health Series, Vol. 22. International Labour Office: Geneva, 2002.
3. Ducatman AM. Variability in interpretation of radiographs for asbestosis abnormalities: problems and solutions. *Annals of the New York Academy of Sciences* 1991; **643**:108–120.
4. Welch LS, Hunting KL, Balmes J, Bresnitz EA, Guidotti TL, Lockey JE, Myo-Lwin T. Variability in the classification of radiographs using the 1980 International Labor Organization Classification for Pneumoconioses. *Chest* 1998; **114**:1740–1748.
5. McAdams HP, Samei E, Dobbins J, Tourassi GD, Ravin CE. Recent advances in radiology. *Radiology* 2006; **241**:663–683. DOI: 10.1148/radiol.2413051535.
6. Gibbons JD, Olkin I, Sobel M. An introduction to ranking and selection. *The American Statistician* 1979; **33**:185–195.
7. Zhou H, Lange K. Rating movies and rating the raters who rate them. *The American Statistician* 2009; **63**:297–307. DOI: 10.1198/tast.2009.08278.
8. Nam J. Comparison of validity of assessment methods using indices of adjusted agreement. *Statistics in Medicine* 2007; **26**:620–632. DOI: 10.1002/sim.2562.
9. Clogg CC. Latent class models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Arminger G, Clogg CC, Sobel ME (eds). Plenum Press: New York, 1995; 311–359.
10. Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Statistics in Medicine* 1986; **5**:21–27.
11. McCutcheon AL. *Latent Class Analysis*. Sage Publications: Newbury Park, CA, 1987.
12. Collins LM, Lanza ST. *Latent Class and Latent Transition Analysis*. Wiley: Hoboken, NJ, 2010.
13. Chung H, Loken E, Schafer JL. Difficulties in drawing inferences with finite-mixture models: a simple example with a simple solution. *The American Statistician* 2004; **58**:152–158. DOI: 10.1198/0003130043286.
14. Lanza ST, Lemmon DR, Schafer JL, Collins LM. *PROC LCA & PROC LTA User's Guide, Version 1.1.5 beta*. The Methodology Center: Pennsylvania State University, 2008.
15. SAS Institute. *SAS/STAT® Users Guide, Release 9.1 Edition*. SAS Institute: Cary, NC, 2004.
16. Goodman LA. On the assignment of individuals to latent classes. *Sociological Methodology* 2007; **37**:1–22. DOI: 10.1111/j.1467-9531.2007.00184.x.
17. Hauser MH. *A Vector Space Approach to Geometry*. Dover: New York, 1998.
18. SAS Institute. *SAS IML® Users Guide, Release 9.1 Edition*. SAS Institute: Cary, NC, 2004.
19. Greenacre M. *Correspondence Analysis in Practice*, 2nd ed. Chapman and Hall: New York, 2007.
20. Penny KI, Jolliffe IT. A comparison of multivariate outlier detection methods for clinical laboratory safety data. *The Statistician* 2001; **50**:295–308.
21. Brenner H. How independent are multiple 'independent' diagnostic classifications? *Statistics in Medicine* 1996; **15**:1377–1386.
22. SAS Institute. *SAS/STAT 9.1 Language Reference: Dictionary*, Vol. 2. SAS Institute: Cary, NC, 2004.
23. Uebersax JS. Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin* 1987; **101**:140–146.
24. Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine* 1997; **16**:2157–2175.
25. Neter J, Wasserman W, Kutner MH. *Applied Linear Regression Models*. Irwin: Homewood, IL, 1983.
26. Dendukuri N, Hadgu A, Wang L. Modeling conditional dependence between diagnostic tests: a multiple latent variable model. *Statistics in Medicine* 2009; **28**:441–461. DOI: 10.1002/sim.3470.
27. Campbell MK, Donner A, Webster KM. Are ordinal models useful for classification? *Statistics in Medicine* 1991; **10**:383–394.
28. Uebersax JS. Modeling approaches for the analysis of observer agreement. *Investigative Radiology* 1992; **27**:738–743.
29. Uebersax JS. Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association* 1993; **88**:421–427.
30. Andrich D, deJong JHAL, Sheridan BE. Diagnostic opportunities with the Rasch model for ordered response categories. In *Applications of Latent Trait and Latent Class Models in the Social Sciences*, Rost J, Langeheine R (eds). Waxmann: New York, 1997; 59–70.
31. Miller JO, Nelson BL, Reilly CH. Efficient multinomial selection in simulation. *Naval Research Logistics* 1998; **45**:459–482.
32. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; **45**:255–268.