Signal detection theory (SDT) has been widely applied in situations where observers attempt to detect or discriminate between two or more events. The usefulness of SDT with latent classes was illustrated in the context of an educational situation that can be readily conceptualized as a signal detection task: grading term papers. The approach assumes that the graders attempt to discriminate between latent classes of papers by using a decision criteria in combination with their perceptions of the quality of each paper. Three graders (a professor and two graduate assistants) graded 85 term papers from a graduate course on measurement. A fit of the latent class signal detection model indicates that the graders discriminate equally between two latent classes, but their response criteria differ. These are similar to results typically found in signal detection experiments with observed events. The findings show that SDT offers a simple summary of the graders' performance in terms of their ability to discriminate between the latent classes and their arbitrary use of grade categories. (Contains 1 figure, 2 tables, and 11 references.) (SLD)
Signal detection with latent classes: A perspective on paper grading

Lawrence T. DeCarlo
Teachers College, Columbia University

Poster presented at the 2000 annual meeting of the American Educational Research Association, New Orleans, LA.
Signal detection theory (SDT) has been widely applied in situations where observers attempt to detect or discriminate between two or more events (see Macmillan & Creelman, 1991). It has played an important role in memory research in psychology, for example, in part because it provides a measure of memory that is separate from arbitrary response effects. In this type of application, it is known whether or not an event actually occurred (e.g., whether or not a word was previously presented during a study period). In other situations, however, the task is again one of signal detection, but the event is not observed. An example is attempting to determine whether or not a person has a psychological or physical condition, such as depression or disease, where the true state of the person is not known. In this case, the psychological theory is the same (i.e., SDT), with the only difference being that the events of interest are latent.

Signal detection theory can readily be applied to this type of situation by incorporating it into a latent class analysis (Dayton, 1998; McCutcheon, 1987). As shown below, latent class signal detection models are simply generalized linear models with latent categorical predictors (one or more signals versus noise; see Figure 1); they are closely related to located latent class models (e.g., Formann, 1985; Uebersax, 1993) and to discretized latent trait models (Clogg, 1988; Heinen, 1996), but they differ with respect to parameterization and perspective. For example, the latent classes are viewed in signal detection as being qualitative, and not as arising from the discretization of a continuous latent variable.

The utility of SDT with latent classes is illustrated in the context of an educational situation that can readily be conceptualized as a signal detection task: grading term papers. The approach assumes that the graders attempt to discriminate between latent classes of papers by using a decision criteria in combination with their perception of the quality of each paper. It is shown that SDT offers a simple summary of the graders performance in terms of their ability to
discriminate between the latent classes and their arbitrary use of grade categories. The approach also provides measures of the reliability of the graders individually and as a set. Some evidence as to the validity of the latent classes, namely their relation to students' average grade on two course exams, is also presented.

Consider the situation where \( j \) independent observers examine stimuli and make decisions as to which of \( C \) events are present; the discussion here focuses on the basic situation with two events (signal and noise), but the extension to three or more events is straightforward. A general signal detection model for binary or rating responses and two events is

\[
p(Y_{ij} \leq k | X) = F(c_{jk} - d_{j}X),
\]

where \( K \) is the number of response categories, \( 1 \leq k \leq K-1 \), \( X \) is a dummy coded variable that indicates the two events, \( p(Y_{ij} \leq k | X) \) is the cumulative probability of response \( k \) by observer \( j \) conditional on \( X \), \( c_{jk} \) is the distance of the \( k \)th response criterion from the mode of the reference distribution for the \( j \)th observer, \( d_{j} \) are the distances between the two underlying distributions for the \( j \)th observer, and \( F \) is a cumulative distribution function (CDF) for the underlying distributions. The inverse of \( F \) corresponds to a link function \( g \), with common choices being the logit, inverse normal, and complementary log log links, which give signal detection models based on logistic, normal, and extreme value distributions, respectively (DeCarlo, 1998).

To extend the model to the situation where the events are latent, the observed categorical variable \( X \) is replaced by a latent categorical variable, say \( X_c \), with \( c = 1,2 \). The model can be incorporated into a restricted latent class model by using differences between the cumulative probabilities,
for the conditional probabilities of a latent class model, which for three observers and two latent
classes can be written as

\[ p(Y_j = k | X_c) = F(c_{jk} - d_j X_c) \quad k = 1 \]

\[ p(Y_j = k | X_c) = F(c_{jk} - d_j X_c) - F(c_{jk-1} - d_j X_c) \quad 1 < k < K \]

\[ p(Y_j = k | X_c) = 1 - F(c_{jk-1} - d_j X_c) \quad k = K, \]

for \( k = 1 \), \( 1 < k < K \), and \( k = K \).

Equations 1 and 2 offer a general class of signal detection models with latent classes that
can be used in situations that can be conceptualized in terms of SDT, such as when observers
attempt to detect or discriminate latent categorical events. The model can be fit using software
for latent class analysis that allows one to restrict the conditional probabilities using different
cumulative link functions, such as LEM (Vermunt, 1997).

**Methods**

Three graders (professor and two graduate assistants) graded 85 term papers from a
graduate course on measurement. The papers were graded on a scale from 1-4, with the graders
instructed to consider a below average paper as 1, an average paper as 2, an above average paper
as 3, and an excellent paper as 4. Graders were instructed to first read five or six papers, chosen
at random, before grading any of the papers, to obtain an idea of what the average paper might be
like.
Results

Table 1 shows, for latent class logistic signal detection models with from one to four latent classes, information based goodness of fit indices, namely the Bayesian information criterion (BIC) and Akaike’s information criterion (AIC) (see Agresti, 1990). The criteria can be used to compare nested and non-nested models, with smaller values indicating a better model. The eigenvalues of the information matrix did not indicate identification problems for the two or three class models, but there were near zero values for four or more classes. Different runs with different starting values resulted in recovery of the parameter estimates for the two and three class models.

The values of both the BIC and AIC are smallest for the model with two latent classes. Thus, the results suggest that the graders can discriminate between two latent classes (e.g., grades of A and B). Goodness of fit statistics for the two class model are $X^2 = 25.97$, $df=50$, $p=.998$ for the chi-square statistic and $L^2=30.12$, $df=50$, $p=.988$ for the likelihood ratio statistic, both of which suggest acceptable fit.

The top part of Table 2 shows the parameter estimates and standard errors for the model with two latent classes. The estimated sizes of the latent classes are .46 and .54 for classes 1 and 2, respectively. Inspection of the estimated conditional probabilities (not shown) shows that latent Class 1 represents a lower latent class and Class 2 a higher latent class. The detection parameters are close in magnitude (that for observer 1 is higher, but the standard error is large), indicating that the graders discriminate equally. A likelihood ratio test of a restricted model with detection parameters equal across the three observers gives $LR =1.22$, $df=2$, $p=.54$, so the restricted model is not rejected; the values of BIC and AIC are also both smaller than those for the unrestricted model. The lower half of Table 2 shows the parameter estimates for the restricted
model. The estimate of $d$ is 2.36, so the odds of a higher response are $\exp(2.36)=10.6$ times higher for class 2 than for class 1, which is comparable to detection found in memory and psychophysics experiments. The table also shows that the standard errors for the restricted model tend to be considerably smaller. A correlation-like conditional measure of reliability, Yule's $Q$, can be obtained from $d$ as $\frac{\exp(d)-1}{\exp(d)+1}$, which in this case gives .83. Lambda, the relative reduction in prediction error, provides a measure of the reliability of the observers as a set (see Clogg & Manning, 1996), and in this case its estimate is .71.

The estimates of the response criteria suggest that the three graders differ, and a likelihood ratio test of the restriction of equal criteria across the graders leads to rejection of the restriction. The main difference, as can be seen in Table 2, is that grader B had a higher criteria for a grade of 2 than the other two graders. Since the graders were instructed to consider 2 as average, this suggests that grader B had a stricter view as to what average is.

Each paper can be classified into one of the latent classes using the modal posterior probability, that is, $p(X_0|Y_1,Y_2,Y_3)$. Evidence as to the validity of the classification is given by a comparison of the average score on two course exams across the latent classes; the mean was 76.4 for Class 1 (the lower class) and 81.5 for Class 2, with the difference being significant ($t=2.6$, $df=83$, $p=.012$). Thus, students in the higher latent class had an average score on two course exams that was about five points higher. Note that if one wishes to assign finer ordinal grades to individuals (e.g., A, A-, B+, B), this can be done using the modal posterior probabilities by grouping the probabilities into categories. This is consistent with Clogg's (1988; also see Uebersax, 1993) suggestion to use the product of the posterior probabilities and values assigned to the latent classes in order to assign scores to individuals. The difference in this case is that the latent classes are treated as purely categorical, so the values assigned to the latent classes...
are simply zero and one (in which case Clogg’s suggested scoring system simply uses the posterior probabilities as scores).

In sum, a fit of the latent class signal detection model indicates that the graders discriminate equally between two latent classes, but their response criteria differ; these are similar to results typically found in signal detection experiments with observed events. The magnitude of $d$ and the measures of conditional reliability indicate good discrimination; the latent classes also differed with respect to average exam grade, which provides evidence as to validity.

Conclusion

Paper and essay grading has been studied from several perspectives, such as that offered by the Rasch model and by item response theory. The approach via SDT provides a somewhat different perspective. For one, the latent classes are viewed as being categorical, and not as arising from a discretization of a latent trait. The result is that measurement in this case is qualitative. Second, the discrimination parameter in SDT is viewed as a fixed characteristic of the observer, whereas the response criteria are not; in item response theory the discrimination and item difficulty (rater severity) parameters are both considered fixed. The view via SDT also suggests that a large body of research and theory in experimental psychology is relevant to paper and essay grading, and it suggests new research, such as attempting to manipulate the graders’ response criteria across sessions to see if their discrimination remains constant, as found in classic experiments in SDT with observable events. This would provide an important experimental validation of the model and theory.
References


Table 1

Information Criteria for Latent Class Signal Detection Models

<table>
<thead>
<tr>
<th># of Classes</th>
<th>BIC</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>675.81</td>
<td>653.83</td>
</tr>
<tr>
<td>2</td>
<td>671.20</td>
<td>639.45</td>
</tr>
<tr>
<td>3</td>
<td>687.11</td>
<td>645.59</td>
</tr>
<tr>
<td>4</td>
<td>697.32</td>
<td>646.02</td>
</tr>
</tbody>
</table>

Notes: BIC = Bayesian information criterion, AIC = Akaike's information criterion.
Table 2

**Parameter Estimates and Standard Errors for Latent Class Signal Detection Model with Two Classes**

<table>
<thead>
<tr>
<th></th>
<th>$d_j$</th>
<th>$g_{j1}$</th>
<th>$g_{j2}$</th>
<th>$g_{j3}$</th>
<th>$p(X_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observer A</strong></td>
<td>3.59 (1.60)</td>
<td>-0.88 (0.45)</td>
<td>2.93 (1.48)</td>
<td>4.55 (1.59)</td>
<td>.46</td>
</tr>
<tr>
<td><strong>Observer B</strong></td>
<td>2.09 (0.64)</td>
<td>0.46 (0.44)</td>
<td>2.20 (0.59)</td>
<td>4.30 (0.74)</td>
<td></td>
</tr>
<tr>
<td><strong>Observer C</strong></td>
<td>2.04 (0.68)</td>
<td>-1.41 (0.44)</td>
<td>0.96 (0.51)</td>
<td>3.23 (0.68)</td>
<td></td>
</tr>
</tbody>
</table>

**Equal Detection:**

<table>
<thead>
<tr>
<th></th>
<th>$d_j$</th>
<th>$g_{j1}$</th>
<th>$g_{j2}$</th>
<th>$g_{j3}$</th>
<th>$p(X_1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observer A</strong></td>
<td>2.36 (0.37)</td>
<td>-1.04 (0.44)</td>
<td>1.96 (0.55)</td>
<td>3.42 (0.56)</td>
<td>.47</td>
</tr>
<tr>
<td><strong>Observer B</strong></td>
<td>2.36 (0.37)</td>
<td>0.55 (0.51)</td>
<td>2.39 (0.55)</td>
<td>4.52 (0.64)</td>
<td></td>
</tr>
<tr>
<td><strong>Observer C</strong></td>
<td>2.36 (0.37)</td>
<td>-1.36 (0.47)</td>
<td>1.11 (0.54)</td>
<td>3.49 (0.56)</td>
<td></td>
</tr>
</tbody>
</table>
Observed signal  Latent signal
Signal detection with latent classes: A perspective on grading.

Lawrence T. DeCarlo

Teachers College, Columbia University

April 25, 2000
May 8, 2000

Dear AERA Presenter,

Hopefully, the convention was a productive and rewarding event. As stated in the AERA program, presenters have a responsibility to make their papers readily available. If you haven’t done so already, please submit copies of your papers for consideration for inclusion in the ERIC database. We are interested in papers from this year’s AERA conference and last year’s conference. If you have submitted your paper, you can track its progress at http://ericae.net.

Abstracts of papers accepted by ERIC appear in Resources in Education (RIE) and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of RIE. Abstracts of your contribution will be accessible through the printed and electronic versions of RIE. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the 2000 and 1999 AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC’s criteria for inclusion in RIE: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form enclosed with this letter and send two copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can mail your paper to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 2000/ERIC Acquisitions
        University of Maryland
        1129 Shriver Laboratory
        College Park, MD 20742

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

ERIC is a project of the Department of Measurement, Statistics & Evaluation