Kappa and Lambda Homework

The attached handout is taken from the draft manuscript Almond, Mislevy, Steinberg, Williamson and Yan (in preparation) *Bayesian Networks in Educational Assessment.* (under contract with Springer). The pages are the ones that go over Cohen's Kappa and Goodman and Kruskal's lambda.

The stuff about Bayesian networks may be a bit confusing, but just think of them this way: Let $\boldsymbol{S}$ be a set of proficiency variables and $\boldsymbol{X}$ a collection of observable outcome variables. The Bayes net provides a way of specificying $Pr(\boldsymbol{X}|\boldsymbol{S})$ and $Pr(\boldsymbol{S})$. Through Bayes theorem one can calcute $Pr(\boldsymbol{S}|\boldsymbol{X})$. It is also fairly easy to simulate from the data. There is really nothing in this section specific to Bayesian networks, one could easily do the same calculations with MIRT or ordered latent class models.

Your assignment is the last problem, 7.16. It is worth 25 points: 5 points each for

Kappa for Listening

Kappa for Speaking

Lambda for Listening

Lambda for Speaking

Comparing the values to the ones from the Reading and Writing given in the example.

7

Explanation and Test Construction

For Bayesian network models to be useful in educational applications, they must not only provide belief estimates for important proficiencies and claims about the learner, but they must also explain the basis of those estimates. Explanation transforms the model from a black box that pontificates an answer to a question into a glass box whose reasoning methods and assumptions can be evaluated. Contrast this to a neural network model that classifies a learner without being able to explain the rationale behind its conclusion. Usually, a preliminary model makes several unrealistic assumptions which result in unrealistic inferences. Models must be "debugged" like computer programs, to correct errors in assumption or specification (Almond, Kim, Shute, & Ventura, 2013). The mechanisms used for explanation aid in the process of model validation, criticism, and debugging.

For assessments constructed using Evidence-Centered Design (ECD; Chapter 2), it is only natural that the explanation would be in terms of evidence. Each observed outcome from an assigned task provides "evidence" for or against a claim represented by one or more proficiency variables. But how much? The *weight of evidence* quantifies the evidence provided by a single observation, a complete task or a complete test. There is a close connection between the weight of evidence and the *reliability* of an assessment.

If we have not yet seen the results from a task, we can calculate the *expected weight of evidence* for that task. This gives a guide to test construction for both adaptive and fixed form tests. Expected weight of evidence is always calculated with respect to a hypothesis, so we can use it as a spot meter to determine where an assessment has the most power. We can use expected weight of information to make cost/benefit trade-offs and focus the assessment for particular purposes, even on the fly in adaptive tests.

Section 7.1 reviews some of the literature on explanation in graphical models, describing some simple textual and coloring techniques. Section 7.2 formally defines weight of evidence and provides some of its key properties. Section 7.3 describes expected weight of evidence as a metric for Activity Selection—selecting the next task in an adaptive test. Section 7.4 expands

This can be also handled with constraints. Often what is done is to create an *enemy list*, E, of tasks that should not appear on the same form. The constraint is then $\sum_{j \in E} u_j \leq 1$. Assessment designs often require many such enemy lists. Note that enemy lists can be defined through task model variables: Any task which has a particular task model variable set to a certain value might appear in an enemy list.

- *Item sets.* One problem that has long been difficult for conventional test design is how to accommodate item sets: items that must appear together on the same form because the share common stimulus material (e.g., a reading passage). Optimization algorithms that use a greedy function to optimize the objective function (Equation 7.10) can easily get into trouble because once they pick one item from the set, the algorithm must then pick several others from that set. This can lead to bad forms if the remaining items in the set are too hard, too easy, or do not meet other constraints. Although the problem is more noticeable in adaptive testing, it makes the optimization problem more difficult in fixed form tests as well.
  The ECD model avoids many of the difficulties by assembling forms from tasks instead of items. Usually, an item set can be modeled as a single task. As the selection algorithm considers the joint evidence from the task, it is harder to get stuck by making a poor initial choice. However, there may be other considerations here: tasks may be bound together in scenarios, or some but all items from a set might be needed. Additional constraints can be written to meet these conditions.

The assembly model must contain a target rule and at least one constraint. There can be as many or as few constraints of each type as are needed to express the intent of the designers and to ensure that sufficient evidence is gathered for all of the claims. By expressing the target rule as a function of the task indicators, u_j, and the constrains as inequalities using the task indicators, one can use standard 0/1 linear programming to assemble a test form that resembles a previous form or to assemble multiple parallel test forms. Standard optimization theory and software (e.g., Nocedal & Wright, 2006) can be applied. Alternative approaches and many insights to particular challenges and kinds of constraints in test-assembly more generally are found in van der Linden (2005).

7.5 Reliability and Assessment Information

When we build the evidence model $\mathrm{P}(\mathbf{X}|\mathbf{S})$ we are acknowledging that the relationship between the proficiency variables, $\mathbf{S}$ and the observed outcomes, $\mathbf{X}$ is a probabilistic one. In other words, the outcome pattern $\mathbf{X}$ is not a pure measure of the proficiency variable $\mathbf{S}$, but contains some noise that is irrelevant to what we are trying to measure. In engineering terms, we could think about the signal to noise ratio for the assessment; in psychometrics we speak of the *reliability*.

Note that not all sources of noise are actually irrelevant to the construct we are trying to measure. Take for example an assessment of communicative competence in a given language. By restricting the setting to the academic environment, we remove one source of variability. However, other settings may also be relevant to the kinds of inferences we are trying to make. For example, if we are trying to understand how well a potential student is likely to be able to get by living in a foreign country, settings related to shopping and interacting with the local bureaucracy may be equally important. In assessment, reliability is usually taken as a measure of the irrelevant sources of variability given a specified domain of tasks and test procedures (Brennan, 2001).

Our treatment of reliability with Bayes nets differs from that of classical test theory in two important respects. First, if our proficiency model is expressed as a Bayesian network, then our scores will typically be either classifications of participants according their proficiency variables or posterior distributions over one or more proficiency variables. The majority of the literature on reliability is devoted to continuous or integer valued scores. Even when authors do talk about classifications, it is usually in the context of a cut score on a continuous variable. Second, classical test theory relies on the concept of a *true score*. Typically, the distribution of the true score in the population is unknown and must be estimated from data. In our case, the true score corresponds to the skill profile, $\mathbf{S}$. The proficiency model provides the population distribution for the skill profile, $\mathrm{P}(\mathbf{S})$.

For simplicity we start with purely discrete scores, where the student is classified as having the skill profile with the highest posterior probability (the MAP estimate). Section 7.5.1 looks at some measures of accuracy, and Section 7.5.2 looks at some measures of consistency between two test forms. However, the Bayes net score is not just a single best proficiency profile, but rather a probability distribution over possible profiles. These contain more information than the point estimates, and hence are usually better scores. Section 7.5.3 extends the discrete accuracy and consistency measures to this continuous world.

7.5.1 Accuracy Matrix

We start with a classification score. We partition the space of skill profiles into a series of disjoint hypotheses, $H_1, \ldots, H_K$, which span the space of possible skill profiles. A common case is to look at the value of one proficiency variable ignoring the others; that is $H_j : S_j = \texttt{expert}$. A more sophisticated model might look at a number of possible courses a student could be placed in and what kinds of student would benefit from which course, yielding a partitioning of all possible vectors in $\mathbf{S}$ according to placement based on proficiency profiles. When there are exactly two hypotheses, this corresponds to the setup in the weight of evidence calculation above. But when the students are to be classified into more than two categories, a new measure is needed which extends to multiple categories.

Suppose that we observe a pattern of outcomes $\mathbf{X}$ from the collection of tasks that appears on one form of the assessment. By Bayes' theorem we obtain the posterior distribution $\mathrm{P}(\mathbf{S}|\mathbf{X})$, and the implied posterior probability for each hypothesis in the partition, or $\mathrm{P}(h_k|\mathbf{X})$. We can then define a point estimate for H by $\hat{H} = \max_{h_k} \mathrm{P}(h_k|\mathbf{X})$. This is the *Maximum A Posteriori* or *MAP* estimate for H. It will be a function of $\mathbf{X}$ so we can write $\hat{H}(\mathbf{X})$.

Doing this assumes that the utility function associated with misclassification is relatively symmetric. That might not always be the case. Again in a licensure test it is more regrettable to license somebody who is not qualified than to make the opposite mistake. Similarly, it may be much more regrettable to fail to identify a student who needs remediation than the opposite. In such case, instead of choosing the value of $\hat{H}$ which maximizes the posterior probability, we would take the one which maximizes expected utility. This is called the *Bayes decision* and is covered in standard texts on decision theory (e.g., DeGroot, 1970; Berger, 1985).

We define the elements of the *accuracy matrix*[2] A as follows:

$$a_{ij} = \mathrm{P}(H = h_i, \hat{H} = h_j) = \sum_{\mathbf{x}:\hat{H}(\mathbf{x})=h_j} \mathrm{P}(\mathbf{x}|H = h_i)\mathrm{P}(H = h_i) \; . \qquad (7.11)$$

This is the probability that when h_i is the correct hypothesis, a response vector $\mathbf{x}$ will be observed for which $\hat{H} = h_j$ is the decision. The diagonal of this matrix corresponds to the cases where the decision agrees with the true classification. Perfect agreement would result in a diagonal matrix. Thus we can define the *accuracy* as the trace of the matrix, that is, $\sum_k a_{kk}$.

The accuracy matrix A may be difficult to calculate analytically, especially for a long assessment. In general, evaluating Equation 7.11 involves iteration over both the set of possible skill profiles and the set of possible outcome patterns. This becomes prohibitively expensive as the number and complexity of the tasks in the assessment increases. However, it can be easily estimated by a simple simulation experiment. First randomly select a skill profile according to the distribution of the proficiency model, $\mathrm{P}(\mathbf{S})$. We can then assign the value of H based on the selected skill profile $\mathbf{S}$. Next, we randomly select an outcome pattern $\mathbf{X}$ according to the distribution of the evidence model $\mathrm{P}(\mathbf{X}|\mathbf{S})$. We can then classify the simulated outcome with an estimated value of the hypothesis $\hat{H}(\mathbf{X})$. If we repeat this experiment many times, the observed frequencies will converge to the accuracy matrix.

This experiment contains two important assumptions: The first is that the model is correct, i.e., the model used to generate the data is the same as the one used in the classification. In practice we can never know the true data generation model. The second is that there is no accounting of the uncertainty about the probabilities (or parameters from which they are obtained) used to generate the $\mathbf{S}$s and the $\mathbf{X}$s. Part II takes up the issue of uncertainty about the parameters. Taking these problems into consideration, calculating

[2] This is sometimes called a *confusion matrix*, referring to its off-diagonal elements.

the accuracy matrix estimate in this fashion is really an upper bound on the true accuracy of the assessment.

Example 7.6 (Language Test Accuracy Matrix). *Consider once more the simplified language test from Mislevy (1995c) described in Example 7.1, (see also Appendix A.2). Suppose we perform the following experiment. First, we simulate 1,000 possible proficiency profiles from the proficiency model. Next, we generate a response vector over 63 geometric sequence tasks for each of the 1,000 simulees. Finally, we score the test for all 1,000 simulees (ignoring their actual proficiency profiles). For each simulee we should now have both their "true" (from the simulation) value of the Reading, Writing, Speaking and Listening nodes and the most likely (MAP) estimate for each node from the scored response.*

We can calculate the accuracy matrix as follows: First, set $a_{ij} = 0$ for all i and j. Next, for each simulee, if the true value of Reading is i and the MAP estimate is j, add one to the value of a_{ij}. Repeating this for all 1,000 simulees and dividing by 1,000 (the number of simulees) yields a matrix like Table 7.1.

Table 7.1. Accuracy Matrixes for *Reading* and *Writing* based on 1,000 Simulated students

	Reading			*Writing*		
	Novice	Intermediate	Advanced	Novice	Intermediate	Advanced
Novice	0.229	0.025	0.000	0.163	0.097	0.000
Intermediate	0.025	0.445	0.029	0.053	0.388	0.065
Advanced	0.000	0.040	0.207	0.000	0.051	0.183

Some authors take the accuracy defined in this way as a measure of validity rather than one of reliability. However, this contains the implicit assumption that the set of hypotheses $H_1, \ldots, H_K$ and by extension the proficiency variables $\mathbf{S}$ represent the construct on which we wish to base our decisions. We prefer to think of the accuracy as a measure of internal consistency under the model, that is, of reliability, and reserve the term "validity" for measures which take into account the use and consequences of the classification (Section 16.4).

Many other important measures of agreement can be derived from the accuracy matrix. In defining those measures, it will be helpful to have notation for the row and column sums. Let $a_{i+} = \sum_j a_{ij} = \mathrm{P}(H_i)$ be the sum of all of the elements in Row i. This is the marginal (population) probability of the hypothesis. Let $a_{+j} = \sum_i a_i j = \mathrm{P}(\hat{H}_j)$. This is the marginal probability of the classifications.

Simply normalizing the accuracy matrix by dividing by the row sums produces interesting results. The matrix normalized in this way, $a_{ij}/a_{i+} = \mathrm{P}(\hat{H} =$

$h_j|H = h_i)$, produces the operating characteristics of the assessment. If the hypothesis is binary, then a_{11}/a_{1+} is the *sensitivity* of the test, the probability of asserting that the hypothesis holds when it is in fact true. Similarly, a_{22}/a_{2+} is the *specificity* of the test, the probability of asserting that the hypothesis is false when in fact it is false. These terms are used frequently in medical testing.

Often the multiple measures of the operating characteristics are more useful than a single measure describing accuracy. This is particularly true because most of the single number summaries depend on the population distribution of H. The operating characteristics specifically condition out this distribution. They are still a useful measure of the strength of evidence in the assessment even when all of the members of the sample have the same value for the hypothesis.

Normalizing by the column sums also has another interesting interpretation. Now we are conditioning on observed decision, $a_{ij}/a_{+j} = \mathrm{P}(H = h_i|\hat{H} = h_j)$. The resulting conditional probability distributions answer the question, "If the assessment classifies a participant as $\hat{H}$, what is the probability that this is the true classification?" This is the question that end users of the assessment scores would very much like answered. As in the rare disease problem (Example 3.6), the probability of true classification depends on both the operating characteristics of the test and the population distribution of the hypothesis.

The accuracy, $\sum_i a_{ii}$, answers the question "What is the probability that the classification assigned on the basis of this assessment will agree with truth?" Note that it is possible to get a fairly large agreement by chance, even if the classification and truth are independent. Consequently, some authors recommend adjusting the accuracy for chance agreement.

One such adjusted agreement is *Cohen's* κ. Fleiss, Levin, and Paik (2003) note that adjusting for chance agreement unifies a number of different measures of agreement in a 2 by 2 table. If the true value of the hypothesis, H and the estimated hypothesis, $\hat{H}$ were two raters acting independently, the probability of agreement by chance would be $\sum_i a_{i+}a_{+i}$. The coefficient κ is expressed as a ratio of the obtained accuracy corrected for chance to the ideal accuracy:

$$\kappa = \frac{\sum_i a_{ii} - \sum_i a_{i+}a_{+i}}{1 - \sum_i a_{i+}a_{+i}} \ . \tag{7.12}$$

The chance term is based on the idea that the two classification mechanism are independent. Thus Cohen's κ answers the question "How much better is the classification given by this assessment than what we would expect if the assessment was independent of truth?" Sometimes when the categories are ordered, κ is weighted so that classifications that are one category away are worth more that classifications that are multiple categories away. In either case, κ is easier to interpret as a measure of the consistency of two classifiers (Section 7.5.2) than the accuracy of one classifier.

Goodman and Kruskal (1954) offer a different statistic, λ, that using a different baseline, $\max(p_H)$, for adjusting the agreement statistic (see also Brennan & Prediger, 1977). This is the agreement level that would be achieved by simply classifying everybody at the most likely state.

$$\lambda = \frac{\sum_n a_{n,n} - \max_n a_{n,+}}{1 - \max_n a_{n,+}}$$

The metric λ corresponds to the question, "How much better do we do with this assessment than simply classifying each person at the population mode?" This index relates directly to the decision of whether or not to use the test.

Although less well known that Cohen's κ, Goodman and Kruskal's λ is often a better choice when talking about the accuracy of an assessment (regardless of the method used to obtain the estimates). In particular, the question answered by λ is often more interesting. While κ answers how much better is the agreement (between the truth and the classifier), λ answer the question how much better is it to use the classifier than not. In fact, neither measure may be the ideal measure; Goodman and Kruskal (1954), offer a number of alternatives that could be explored.

Example 7.7 (Simplified Language Test Accuracy Matrix, Kappa and Lambda). *Using the estimated accuracy matrix for the simplified language test (Table 7.1, Example 7.6), we can calculate Cohen's κ and Goodman and Kruskal's λ. To begin, we find the diagonal of the Reading portion of Table 7.1; this is 0.881. In other words, this form of the assessment classifies slightly almost 90% of the examinees correctly on reading. Next, we sum over the rows and columns to produce marginal distributions for the true proficiency levels,* $(.254, .499, .247)$*, and the estimated proficiency levels,* $(.254, .510, .236)$.

To calculate κ, we need to calculate the probability of chance agreement. We get this by multiplying the two vectors of marginal probabilities and taking the sum, which yields 0.377. Thus, about 1/3 of the time we are likely to get the correct classification just by chance. The adjusted agreement is now $\kappa = (0.881 - 0.377)/(1 - 0.377) = 0.809$*, which means we are getting approximately a 80% improvement over the agreement we would have gotten if we just assigned everybody a label randomly.*

To calculate λ, we note that the modal category is `Intermediate`*, and that the probability that a randomly chosen simulee has* `Intermediate` *ability is .499. In other words, if we rated everybody as* `Intermediate` *we would get approximately 1/2 of the ratings correct. The adjusted agreement is now* $\lambda = (0.881 - 0.499)/(1 - 0.499) = 0.763$*, which means we are getting approximately a 75% improvement over the agreement we would have gotten if we just assigned everybody the label* `Intermediate`*.*

Turning to the `Writing` *variable, similar calculations show $\kappa = .567$ and $\lambda = .462$. These numbers are smaller as a consequence of the test design. In particular, of the 16 tasks in this assessment all but the 5 listening tasks*

involve at least some Reading and hence provide evidence for Reading. Only the three Writing tasks provide direct evidence for Writing, and because those are integrated tasks that also involve Reading, their evidence is weaker. Thus, both λ and κ are smaller.

Note that the Bayesian network does not actually assign each student to a category, rather it gives a probability distribution over the categories. We can do slightly better if we look at the probabilities rather than just the most likely category. Section 7.5.3 explores this. In fact, this is one example of a scoring rule for Bayesian networks; Chapter 10 explores other scoring rules.

7.5.2 Consistency Matrix

Suppose that we have two parallel forms of the assessment, Form X and Form Y. We could produce two accuracy matrixes A_X and A_Y, one for each form. The *Consistency Matrix* is the product of those two accuracy matrixes, $C = A_X^t A_Y$. The normalized rows and columns represent conditional probabilities which describe what we expect to happen when a person who takes Form X later takes Form Y and *visa versa*. This is of practical importance to testing program where the same assessment (with alternative forms) is given over and over again to a similar population of examinees. In this case, large shifts in the classification is likely to produce confusion among the test-takers and score-users.

The consistency matrix can be estimated with a simulation experiment as described above, it can also be estimated by giving both Form X and Form Y to a sample of examinees. If the test is long enough, it could also be used to form split-half estimates of reliability. However, this may be tricky with a diagnostic assessment. In particular, there may only be a few tasks providing direct evidence for each proficiency of interest. Hence the half-tests may be very unbalanced or have very low reliability.

The *consistency* is the sum of the diagonal elements of the consistency matrix, $\sum_i c_{ii}$. This answers the question, "What fraction of examinees who take both Form X and Form Y will get the same classification with respect to hypothesis H?" Cohen's κ is frequently used with the consistency matrix as well. It answers the question "How much better do the two forms agree than two form which are independent, that is measuring different aspects of proficiency?" Again, it may be worth exploring some of the other measures described in Goodman and Kruskal (1954) as well.

7.5.3 Expected Value Matrix

One source of error in both the accuracy matrix and classification matrix is that we are assigning a person to a class on the basis of the MAP estimate for the hypothesis. This gives equal weight to someone who we believe with high confidence is in one category and someone who is on the border between two

categories. By reporting the marginal probability of classification rather than the MAP estimate, we should better convey our uncertainty about the truth.

Suppose we simulate *proficiency profiles* and outcome vectors from N simulees. Let $\mathbf{S}_n$ be the proficiency profile for Simulee n and let $\mathbf{X}_n$ be the outcome vector. Then $\mathrm{P}(H|\mathbf{X}_n)$ is the probabilistic classification that we would assign to Simulee n on the basis of the outcome vector $\mathbf{X}_n$. We can define a *probabilistic classification matrix for Hypothesis H* by summing over these classifications.

$$z_{ij} = \sum_{n:H(\mathbf{S}_n)=h_i} \mathrm{P}(H = h_j|\mathbf{X}_n) \tag{7.13}$$

Here z_{ij} is the weighted number of individual whose true classification is h_i who are classified as h_j, where their weights are the posterior probability of classification in that class. In other words, in the simulation to estimate the accuracy matrix we place a simulee from the i^{th} category into the cell for the j^{th} decision category; to estimate the expected value matrix, we distribute that simulee across all n cells in the j^{th} column, according to its posterior probabilities for each.

The sum of the diagonal elements, $\sum_i z_{ii}$, is another measure of accuracy for the assessment. We can also look at Cohen's κ and λ for the probabilistic classification as well. In general, these should do at least as well as their non-probabilistic counterparts.

Another way to treat the probabilistic scores, $\mathrm{P}(H|\mathbf{X}_n)$ is to regard them as predictions of the true value of the hypothesis. In this case within the confines of the simulation experiment, we can use the scoring rules in Chapter 10 to evaluate the quality of the assessment for making this particular prediction.

Example 7.8 (Simplified Language Test Expected Accuracy Matrix). *The procedure is similar to the one used in Example 7.6. The initial simulation proceeds in the same way. It differs at the scoring step, instead of calculating the MAP score for Reading we calculate its marginal probability. This score will be a vector of three number over the possible classifications (*`Novice`*,* `Intermediate`*, and* `Advanced`*). The "true" value is still a single state.*

We can calculate the expected accuracy matrix as follows: First, set $z_{ij} = 0$ for all i and j. Next, for each simulee, let the true value of SolveGeometricProblems is i and the marginal estimate be $\{p_1, p_2, p_3\}$, which is a probability vector. We update the values of z_{ij} by adding the probability vector to Row i, that is, let $z_{ij} \leftarrow z_{ij} + p_j$ for $j = 1, 2, 3$. Repeating this for all 1,000 simulees yields a matrix like Table 7.2. We divide the entries in this matrix by 1,000 to put all of the numbers on a probability scale.

The agreement measures $\tilde{\kappa}$ and $\tilde{\lambda}$ are calculated in the same way. In this case, $\tilde{\kappa}_{Reading} = .73$, $\tilde{\lambda}_{Reading} = .66$, $\tilde{\kappa}_{Writing} = .43$, and $\tilde{\lambda}_{Writing} = .28$. These are lower than the agreement rates based on the modal classifications (Example 7.7), but are more honest about the uncertainty in the classifications.

Table 7.2. Expected Accuracy Matrixes based on 1,000 Simulations

Reading

	Novice	Intermediate	Advanced
Novice	0.220	0.034	0.000
Intermediate	0.037	0.413	0.050
Advanced	0.000	0.049	0.198

Writing

Novice	Intermediate	Advanced
0.162	0.092	0.007
0.091	0.331	0.084
0.003	0.076	0.154

7.5.4 Weight of Evidence as Information

The preceding discussion has introduced many different possible measures for reliability, not just one. That is because when a test user asks about the reliability of an assessment, there are a number of possible motivations for that question. She might be asking about how the results from the assessment varies when sampling tasks from a pool of possible tasks. In this case, consistency is the most appropriate answer. She might be asking about how well the assessment captures the variability in the population; in this case accuracy, perhaps as measured by Cohen's κ is a reasonable choice. She might be asking whether or not it is worthwhile to give the assessment to learn about a hypothesis, H. In this case, λ seems appropriate.

Smith (2003) presents another possible meaning for reliability, namely "Sufficiency of Information." Smith points out that a teacher may give an end of unit quiz expecting all of the students to get all of the items correct. After all, having finished the unit, the students should have mastered the material. This quiz serves several important purposes: (1) it helps the student's self-assessment of their progress on this material, and (2) it identifies for the teacher any students who have not yet mastered the material as predicted. This assessment has value, even though by many of the reliability measures posed above may have trivial values because all of the students are expected to have mastered the material.

Note that the expected weight of evidence does not depend on the population distribution of the hypothesis. The calculations for the expected weight of evidence are done with the conditional distribution given the hypothesis. Thus, if an assessment has a high expected weight of evidence for the hypothesis that the students have mastered the material of the unit it will be appropriate. (It still may be difficult to estimate the task specific parameters from the classroom population as there is little variability with respect to the proficiency variables of interest, but that is a separate problem.)

We have seen that weight of evidence provides a useful mechanism for explaining how certain patterns of evidence influence our conclusions about certain proficiency variables. Furthermore, its ability to act like a spot meter for specific hypotheses helps us to evaluate how much information is provided by a proposed assessment design for a specific purpose. If the assessment does not provide enough information, we could consider altering the assessment

type. Use a simulation experiment to calculate the accuracy matrix for the modified test.

7.16 (Kappa and Lambda for Speaking and Listening). Calculate Cohen's κ and Goodman and Kruskal's λ for *Speaking* and *Listening* proficiencies using the accuracy matrixes in Table 7.6. Compare them to the numbers from Example 7.7, and interpret what they say about the relative information in the test for the four skills.

Table 7.6. Accuracy Matrixes for *Speaking* and *Listening* based on 1,000 Simulated students

	Speaking			*Listening*		
	`Novice`	`Intermediate`	`Advanced`	`Novice`	`Intermediate`	`Advanced`
`Novice`	0.243	0.027	0.000	0.242	0.054	0.000
`Intermediate`	0.044	0.390	0.030	0.054	0.290	0.059
`Advanced`	0.000	0.033	0.233	0.000	0.087	0.214