Evaluating a Model to Predict Protocol Performance

DOI: 103766/jaaa.20.10.6

Robert G. Turner*
Annette Hurley*

Abstract

Background: A test protocol is created when individual tests are combined. Even with a few tests, many different protocols are possible. Often, the clinician must select a protocol without information as to the performance of the possible protocols. A model to predict protocol performance could help in this selection process.

Purpose: To evaluate the validity and accuracy of a mathematical model for predicting protocol performance.

Research Design: Predictions of the model are compared to actual data on protocol performance.

Results: With complete information, there was almost perfect agreement between predicted and actual data. With partial information, the model still made very accurate estimates of protocol performance.

Conclusions: Even with incomplete information, which is frequently the case in a clinical situation, the model can eliminate many protocols from consideration and aid in the selection of an appropriate protocol.

Key Words: Audiology, central auditory processing disorder, False Alarm Rate, hearing, Hit Rate, Intermediate criterion, Loose criterion, Strict criterion, test correlation, test protocol

Abbreviations: CANS = central auditory nervous system; (C)APD = (central) auditory processing disorder; CBA = cost-benefit analysis; CF = correlation factor; CF_{12} = correlation between Test 1 and Test 2; CS = Competing Sentences test; DD = Dichotic Digits test; FA = False Alarm Rate; FP = Frequency Patterns test; FW = Filtered Words test; HT = Hit Rate; HT_{12} = actual Hit Rate of protocol consisting of Test 1 and Test 2; HT_{p} = predicted Hit Rate or protocol consisting of Test 1 and Test 2 assuming zero correlation; HT_{p} = predicted Hit Rate or protocol consisting of Test 1 and Test 2 assuming maximum positive correlation; HT_{n} = predicted Hit Rate or protocol consisting of Test 1 and Test 2 assuming maximum negative correlation; INT_{2} = Intermediate criterion for the 4-Test Protocol that requires at least two tests to be positive for the protocol to be positive; INT_{3} = Intermediate criterion for the 4-Test Protocol that requires at least three tests to be positive for the protocol to be positive

Frequently, individual tests are combined to form a test protocol. The traditional audiologic test battery is one historical example. Today, test protocols are used, for example, in newborn hearing screening and assessment programs and the assessment of (central) auditory processing disorders ([C]APDs). In the future, audiologists may find other applications for test protocols.

Several individual tests can produce many protocols. For example, four tests can be combined into 11 different protocols consisting of two, three, or four tests. If we consider criterion, then the number of possibilities is greater. Criterion is the minimum number of individual tests that must be positive for the protocol to be positive. Thus, if we consider the four tests plus all protocols with all possible criteria, the clinician is faced with 32 choices. The challenge is to determine which individual test or test protocol is best for clinical use.

Ideally, a protocol should be selected on the basis of a cost-benefit analysis (CBA [Turner, 1991b; Turner et al, 1999]). This analysis attempts to evaluate both objective factors (e.g., financial costs) and subjective factors (e.g., morbidity) so that “benefit” and “cost” can be determined for each possible protocol. A quantitative calculation of cost and benefit for each protocol is best but is frequently impossible because of the difficulty of assigning a

*Department of Communication Disorders, Louisiana State University Health Sciences Center

Robert G. Turner, PhD, 1900 Gravier Street, New Orleans, LA 70112; Phone: 504-568-4339; Fax: 504-568-4352; E-mail: rturne@lsuhsc.edu
quantitative cost and/or benefit to some factors, for example, social isolation. Frequently, a simpler, more qualitative CBA is performed. Various factors are considered when selecting a protocol, but there is no attempt to quantify cost and benefit.

An essential and, typically, the most important part of a CBA is a reasonable estimate of protocol performance. Protocol performance alone may not be sufficient to select the protocol because the protocol with the “best” performance may not be the optimal protocol due to other factors such as financial cost. However, it is these measures of protocol performance that drive the calculations of cost and benefit. In addition, estimates of protocol performance can usually eliminate many protocols from consideration even without a detailed CBA.

There are several measures of protocol performance, but the most basic are Hit Rate (HT) and False Alarm Rate (FA), which are also needed to calculate most other measures. Hit Rate is the probability that the test or protocol will be positive for an individual with the condition of interest, for example, hearing loss. False Alarm Rate is the probability that the test or protocol will be positive for an individual who does not have the condition of interest.

There are several strategies for determining protocol HT/FA. One strategy is a comprehensive study utilizing research and/or clinical data. There is a test group consisting of individuals with the condition of interest. There is also a control group consisting of individuals who do not have the condition of interest. All subjects in both groups must be evaluated with all of the individual tests under consideration. From the results, HT/FA can be calculated for the individual tests and all possible test protocols. This is the best and most direct strategy, but it can be very difficult or impractical for the clinician. Unfortunately, there is seldom any actual data available on protocol performance. The reality is that clinicians will design a protocol for clinical use based primarily on intuition, with little information as to the performance of the selected or any alternate protocols. The current extensive use of different test protocols to diagnose (C)APD is a classic example (American Speech-Language-Hearing Association, 2005).

An alternate strategy is to use a mathematical model to predict protocol performance. Techniques exist to calculate protocol performance prior to clinical use; however, this requires knowledge of protocol criterion, individual test performance, and test correlation (Turner, 1988, 1991a). Criterion, as described above, is the minimum number of individual tests that must be positive for the protocol to be positive. Test performance is the HT/FA of the individual tests. Test correlation is the tendency of two tests to identify the same individuals as positive or negative. Protocol criterion is either explicitly defined by the clinician or implicit in the design of the protocol. In either case, criterion is known. While there may be some information on individual test performance from previous studies, frequently there is little information on the correlation between tests. In most situations, complete information will not be available. A model can still be useful with partial information. For example, if test correlation is not known, protocol performance can be predicted over a range of correlations. With this information, it may be possible to eliminate many protocols from consideration, as will be demonstrated.

A model to predict protocol performance has another use. While we have some information on the relationship of criterion, individual tests performance, and test correlation to protocol performance, we do not have a comprehensive knowledge. A model can be used to study the impact of these three factors on protocol performance. With a better understanding of these factors, the available information on criterion, test performance, and/or test correlation may be better utilized for designing a protocol, estimating performance, setting limits on possible performance, and deciding if a protocol is appropriate. In addition, such knowledge would be useful in modifying a protocol once it has been implemented and clinical data become available.

Using a model to predict protocol performance has obvious advantages. It is certainly faster and more cost-effective than a comprehensive study. It can evaluate many tests and can incorporate new tests as they are developed. Even with limited information, it can help in the selection of a protocol for clinical use. It can also be used to study the factors that determine protocol performance. Obviously, the success of these applications depends on the validity and accuracy of the model. One author (RGT) has developed a model for predicting protocol performance. This model has been previously tested against older data and was reasonably successful predicting protocol performance without test correlation data (Turner, 1988). These were behavioral audiologic tests designed to distinguish cochlear from retrocochlear site of lesion. These tests are no longer in use, so an evaluation using current audiologic tests is appropriate.

**ACTUAL PROTOCOL DATA**

The objective is to evaluate the validity and accuracy of the current model using recent data provided by Musiek and colleagues (forthcoming). They evaluated four commonly used behavioral central auditory processing tests: Dichotic Digits (DD), Frequency Patterns (FP), Filtered Words (FW), and Competing Sentences (CS). Their test group consisted of 20 individuals with known lesions of the central...
The accuracy of the model can be evaluated by comparing predicted values to actual values provided by Musiek and colleagues. The model assumes that all individual tests have the same correlation with all other tests. This correlation can be specified. In theory, test correlation can vary from maximum positive to maximum negative. Maximum positive would have a correlation factor (CF) equal to +1.0, and maximum negative would have CF = −1.0. Maximum positive correlation means that two tests identify, to the degree possible, the same subjects as positive or negative. Maximum negative correlation means that two tests identify, to the degree possible, different subjects as positive or negative. Uncorrelated tests would have zero correlation (CF = 0.0), which means that the subjects identified by one test would tell you nothing about the subjects identified by the other test. Previous results indicate that the older behavioral audiologic tests can be accurately modeled assuming a mid-positive correlation (Turner, 1988). This is a correlation (CF = 0.5) halfway between maximum positive and zero correlation.

For this first example, the model uses the actual HT/FA for the four individual tests and assumes that all tests have zero correlation. Predicted values of HT and FA are plotted against actual values for all test protocols (Figure 1). The line in the figure represents perfect agreement between predicted and actual values. It may not be evident because of the overlap of data points, but there are 48 data points corresponding to all combinations of protocols and criteria evaluated by Musiek et al. (forthcoming). The solid line represents predicted value equal to actual value.

Table 1. Tests and Protocols Evaluated by Musiek and Colleagues (Forthcoming) for Different Criteria

<table>
<thead>
<tr>
<th>2-Test Protocol</th>
<th>Loose</th>
<th>Strict</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD-FP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DD-FW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DD-CS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP-FW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP-CS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FW-CS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-Test Protocol</td>
<td>Loose</td>
<td>INT2</td>
</tr>
<tr>
<td>DD-FP-FW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DD-FP-CS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DD-FW-CS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FP-FW-CS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-Test Protocol</td>
<td>Loose</td>
<td>INT2</td>
</tr>
<tr>
<td>DD-FP-FW-CS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: A check (•) indicates a protocol/criterion combination that was evaluated. There is no criterion for an individual test. INT2 means that at least two tests must be positive for the protocol to be positive. INT3 means that at least three tests must be positive. INT2 was not evaluated for the 3-Test protocols. Tests: DD = Dichotic Digits; FP = Frequency Pattern; FW = Filtered Words; CS = Competing Sentences.

Figure 1. Predicted Hit Rate (HT) and False Alarm Rate (FA) plotted against actual HT and FA for one protocol with a specified criterion, with all tests assumed to have zero correlation for calculating the predicted values. There are 48 data points corresponding to all combinations of protocols and criteria evaluated by Musiek et al. (forthcoming). The solid line represents predicted value equal to actual value.

PREDICTED VERSUS ACTUAL DATA

The auditory nervous system (CANS). Their control group consisted of 29 individuals with no known auditory problems or CANS involvement. The four tests were administered to all subjects in both groups. It was then possible to calculate the HT/FA of the four tests and all possible test protocols (Table 1). Musiek and colleagues evaluated all protocols for Loose and Strict criteria. Loose (also called Lax) means that at least one individual test must be positive for the protocol to be positive. Strict means that all tests must be positive for the protocol to be positive. They also evaluated the two Intermediate criteria for the 4-Test Protocol. An intermediate criterion is a criterion between Loose and Strict. For a 4-Test Protocol, one Intermediate criterion (INT2) requires at least two tests to be positive, while the other (INT3) requires at least three tests to be positive for the protocol to be positive. Musiek and colleagues did not calculate performance for the Intermediate criterion for the 3-Test protocols.
A comparison of actual and predicted values shows that almost 60 percent of actual values lie within the range from zero to mid-positive correlation. This suggests that the best fit to data might occur for a test correlation between CF\textsuperscript{50.0} and CF\textsuperscript{50.5}. Correlation coefficient (R) was calculated for a range of CFs between zero (CF\textsuperscript{50.0}) and maximum positive correlation (CF\textsuperscript{51.0}). As shown in Figure 3, R increases with increasing CF and reached a maximum around CF\textsuperscript{50.3} (R = 0.9973). Recalculating predicted values assuming CF\textsuperscript{50.3} provides the best agreement (Figure 4).

**TEST CORRELATION**

While very good, assuming CF = 0.3 does not provide perfect predictions of actual protocol performance. This is probably the result of several factors. First, the sample size is relatively small. With 20 test subjects, HT can only change in increments of 5 percent, that is, 75 percent, 80 percent, 85 percent, and so on. If predicted HT were 92 percent, actual data could only yield 90 or 95 percent, resulting in some disagreement. Likewise with 29 control subjects, FA has increments of about 3.5 percent. Another factor is the assumption that all tests have the same correlation; the correlation between some pairs of tests may differ from the assumed 0.3. The CF between any two tests can be calculated using the actual HT and FA data for the two individual tests and their 2-Test Protocol. For each two test, four calculations are possible corresponding to all combinations of HT or FA and the Loose or Strict criterion. The formulas are given below for the correlation between Test 1 and Test 2 (CF\textsubscript{12}) based on HT data. For FA data, the corresponding False Alarm Rates are substituted. Equation 1 should be used if CF ≥ 0; Equation 2 should be used if CF < 0:

\[
\text{CF}_{12}(\text{HT}) = \frac{\text{HT}_{12} - \text{HT}_z}{\text{HT}_p - \text{HT}_z} \quad (1)
\]

\[
\text{CF}_{12}(\text{HT}) = \frac{\text{HT}_{12} - \text{HT}_z}{\text{HT}_z - \text{HT}_n} \quad (2)
\]

In the formula, HT\textsubscript{12} is the actual HT of the 2-Test Protocol calculated from the data of Musiek and colleagues for either a Loose or Strict criterion. HT\textsubscript{z} is the predicted HT of the protocol assuming that Test 1 and Test 2 have zero correlation (Loose or Strict). HT\textsubscript{p} is the predicted HT of the protocol assuming that Test 1 and Test 2 have maximum positive correlation (Loose or Strict). HT\textsubscript{n} is the predicted HT of the protocol assuming that Test 1 and Test 2 have maximum negative correlation (Loose or Strict).

There are six test pairs corresponding to all possible pairs of the four individual tests. The calculated correlation factors for these six pairs are shown in
Figure 5. For example, for the two tests DD and FW, the correlation between these tests (CF) is about +1.0 when calculated using HT data for both a Loose criterion and a Strict criterion. For FA data, it is about +0.17 for both criteria. As a result, CF could not be calculated using FA data for any test pairs containing CS. Test correlation is accurate for the specific data in Musiek and colleagues; however, given the small sample sizes and test variability, actual test correlations may be significantly different than if calculated from the data in that study. A difference of just one subject testing positive or negative can produce very different calculated values of CF. For example, in Musiek and colleagues, 10 test subjects tested positive on both tests DD and FW, resulting in a maximum positive correlation (CF = 1.0). Had just one of those 10 tested negative, the tests would have had zero correlation (CF = 0.0).

Still, these data provide some interesting insights into test correlation. It is evident that the calculated correlations between all of the tests are not the same, explaining the errors in the predicted values when assuming all tests have the same correlation. It is interesting to note that CF is the same whether calculated using data for a Loose or Strict criterion. It is reasonable that the correlation between two tests does not depend on criterion.

Another general result is that for any two tests, the calculated CF is different for HT and FA data. Hit Rate is based on the results from the test group, whereas FA is based on the results from the control group. This suggests that the correlation between two tests may differ depending on the population being tested.

The model was modified to utilize the calculated test correlations (CFs) in Figure 5. With this additional information, the model made essentially perfect predictions of the actual data (Figure 6). It is not surprising that the 2-Test predictions were accurate, as data from these protocols were used to calculate CF. More important, the model accurately predicted the data for the 3-Test and 4-Test protocols.

The primary point of this exercise is to demonstrate the validity of the model. If the model makes near-perfect predictions given individual test performance and test correlations, then this validates the underlying mathematics of the model.

SELECTING A PROTOCOL

As discussed above, measures of protocol performance are required to perform a CBA; however, this information is rarely available. Protocol performance
can be predicted if sufficient information on individual test performance and test correlations are available. In the typical clinical situation, there may be some information available as to individual test performance but, typically, no information on test correlation, making it impossible to exactly predict protocol performance. Can a model assist in the selection of a protocol for clinical use without test correlation data? Can estimates of protocol performance for a range of correlations be used to eliminate many protocols from consideration and indicate a few candidates for clinical use?

For this example, we will use the actual HT/FA data for the four individual tests, but no test correlation data, to calculate predicted values. HT is plotted versus FA for predicted (zero and mid-positive correlation) and actual data for 11 protocols (Figures 7–9). Thus, there are three data points for each protocol corresponding to the predicted values for the two assumed test correlations (circle and triangle) and the actual value (square). Also plotted are the individual tests (diamond). The protocols will be evaluated and a test or protocol will be selected using the two predicted values corresponding to zero and mid-positive correlation. The actual value will be used to evaluate that selection.

Assume that a qualitative CBA indicates a need for a protocol with a high HT and that the four tests are comparable in financial and other “costs.” Many screening protocols will employ a high HT so as to reduce the number of individuals with the condition of interest who are missed. Even though the protocol with a higher HT often has a higher FA, this is usually acceptable because of the “cost” of not identifying for follow-up someone with a serious condition.

In Figure 7, a Loose criterion is specified. Generally, a Loose criterion is used to produce a high protocol HT. The higher FA produced by the Loose criterion is an acceptable cost, provided the FA is not unreasonably high. The ideal test or protocol would have HT = 100 percent and FA = 0 percent. This corresponds to the upper left corner in the figure. Thus, tests or protocols that are closest to that corner typically offer the best performance.

Considering just the predicted values, not the actual values, most of the protocols in Figure 7 can be eliminated. Protocol FP-CS provides a predicted HT of about 95 to 98 percent, depending on correlation, and actual data for 11 protocols (Figures 7–9). Thus, there are three data points for each protocol corresponding to the predicted values for the two assumed test correlations (circle and triangle) and the actual value (square). Also plotted are the individual tests (diamond). The protocols will be evaluated and a test or protocol will be selected using the two predicted values corresponding to zero and mid-positive correlation. The actual value will be used to evaluate that selection.

Assume that a qualitative CBA indicates a need for a protocol with a high HT and that the four tests are comparable in financial and other “costs.” Many screening protocols will employ a high HT so as to reduce the number of individuals with the condition of interest who are missed. Even though the protocol with a higher HT often has a higher FA, this is usually acceptable because of the “cost” of not identifying for follow-up someone with a serious condition.

In Figure 7, a Loose criterion is specified. Generally, a Loose criterion is used to produce a high protocol HT. The higher FA produced by the Loose criterion is an acceptable cost, provided the FA is not unreasonably high. The ideal test or protocol would have HT = 100 percent and FA = 0 percent. This corresponds to the upper left corner in the figure. Thus, tests or protocols that are closest to that corner typically offer the best performance.

Considering just the predicted values, not the actual values, most of the protocols in Figure 7 can be eliminated. Protocol FP-CS provides a predicted HT of about 95 to 98 percent, depending on correlation, with FA = 10 percent. Protocol DD-CS can be eliminated because it has the same range of HT as FP-CS but with a higher FA. The other reasonable possibility is Protocol DD-FP. Protocol DD-FP-CS cannot be eliminated because its performance is similar to DD-FP but it requires an additional test. Protocol DD-FP may provide a slightly higher HT than FP-CS, but FA would be about twice that of FP-CS. Based on the predicted values, one would probably select FP-CS unless the highest possible HT is
required. Considering the actual values of HT and FA for all of the protocols, the logical choices would again be FP-CS (actual HT/FA = 95%/10%) and DD-FP (actual HT/FA = 100%/21%). FP-CS would probably be the choice because of the significantly lower FA unless the highest possible HT was required. The important point is that predicted values for protocol HT/FA resulted in essentially the same choice as using the actual HT/FA for the protocols. Even if one preferred DD-FP based on actual values, the selection of FP-CS based on predicted values would not be a major error.

Next, assume that a qualitative CBA indicates a need for a protocol with a low FA. Many newborn hearing screening protocols are designed to produce a lower FA even though the HT may be reduced. The concern is the financial cost of over-referral and the emotional “cost” of identifying a newborn with normal hearing as hearing impaired.

In Figure 8, a Strict criterion is assumed. Generally, a Strict criterion will produce a lower protocol FA and a lower protocol HT. Thus, we want a protocol with a low FA but a HT that is not unreasonably low. Most of the protocols in Figure 8 can be eliminated on the basis of the predicted values of HT and FA. In this case, Test CS offers a HT greater than any protocol with FA = 0 percent. The other possibility is Protocol DD-FP, which, based on predicted values, has a slightly higher HT but a FA that ranges from about 2 to 6 percent. Since the objective is a low FA, Test CS would be a better choice than any protocol. If we consider the actual HT/FA for Protocol DD-FP, Test CS (HT/FA = 75%/0%) would still be the better choice over DD-FP (HT/FA = 80%/7%) because of the much lower FA with Test CS. Again, using the predicted values resulted in the same choice one would make with the actual values. Sometimes, an individual test is a better choice than a protocol.

The protocols selected for Loose and Strict criteria can be compared to the two Intermediate criteria for the 4-Test Protocol, DD-FP-FW-CS (Figure 9). The INT3 criterion requires at least three tests to be positive for the protocol to be positive; INT2 = at least two of four tests must be positive for the protocol to be positive; FP = Frequency Patterns; CS = Competing Sentences; DD = Dichotic Digits; FW = Filtered Words.

Figure 8. Hit Rate vs False Alarm Rate for all protocols assuming a Strict criterion. Actual values (square), predicted values assuming zero correlation (circle), predicted values assuming mid-positive (MDP) correlation (triangle), and the tests (diamond) are shown. Ovals enclose the protocols or test that offer the best performance. DD = Dichotic Digits; FP = Frequency Patterns; CS = Competing Sentences; FW = Filtered Words.

Figure 9. Comparison of the best protocols, assuming a Loose or Strict criterion, to the two 4-Test protocols, assuming an Intermediate criterion. Actual values (square), predicted values assuming zero correlation (circle), predicted values assuming mid-positive (MDP) correlation (triangle), and the Competing Sentences test (diamond) are shown. Ovals enclose the three protocols being considered. INT2 = at least two of four tests must be positive for the protocol to be positive; INT3 = at least three of four tests must be positive for the protocol to be positive; FP = Frequency Patterns; CS = Competing Sentences; DD = Dichotic Digits; FW = Filtered Words.
For the 4-Test Protocol, the INT2 criterion requires at least two tests to be positive and functions somewhat like a Loose criterion. It can be compared to Protocol FP-CS, which was selected for a Loose criterion. The predicted HT of both protocols is similar for the same correlation. The FA of the 4-Test Protocol varies from less than to greater than that predicted for FP-CS, depending on correlation. Protocol FP-CS would probably be selected because it requires two fewer tests. Actual values indicate that FP-CS is a better choice than DD-FP-FW-CS because it has a slightly higher HT (95% vs 90%) and a slightly lower FA (11% vs 14%). Even if the 4-Test Protocol were selected based on predicted values, the cost of that error would be relatively small compared to the cost of using one of the rejected protocols.

A comprehensive CBA for diagnosing (C)APD is beyond the scope of this analysis. However, the model is capable of providing good estimates of protocol performance that could be the basis for a quantitative CBA. In lieu of a detailed CBA, the performance estimates could be evaluated qualitatively using known factors. For example, Musiek and colleagues (forthcoming) express concern over the scoring of Test CS. This could be a “cost” when using CS and a reason to remove CS from consideration. The remaining protocols that do not use CS could be evaluated using the same strategy demonstrated above along with any additional cost/benefit factors. For example, if one test was significantly more expensive than the others but offered little or no improvement in performance, protocols containing that test could be eliminated. Regardless of the CBA strategy, reasonable estimates of protocol performance are required to make an appropriate, defensible selection of a protocol for clinical use.

**SUMMARY**

The objective of this analysis was to use actual performance data for some current audiologic tests to evaluate the validity and accuracy of a mathematical model for predicting test protocol performance. When provided complete information, that is, individual test performance and test correlation, the model predicted, with virtually no error, the performance of all protocols evaluated by Musiek and colleagues. Even when test correlation was unknown and had to be assumed, the model provided very good estimates of protocol performance. Obviously, the model was evaluated using four tests designed to diagnose a particular auditory disorder. The results could be different with other tests; however, the model was also reasonably successful with older audiologic tests even without test correlation data. These results demonstrate the validity and accuracy of the model.

A valid and accurate model for predicting protocol performance is important because it demonstrates that we have at least a basic understanding of the dynamics of the diagnostic testing process. In addition, it has two potential uses. First, it can help select a protocol for clinical use. This model should have general application to any testing situation using multiple tests such as newborn hearing assessment or vestibular evaluation. Generally, performance data for test protocols are not available; the data provided by Musiek and colleagues (forthcoming) are a rare exception. If individual test performance and correlation data are available, the model can make accurate predictions of protocol performance that can be the basis for protocol selection. Unfortunately, test correlation data are rarely available. However, with just individual test HT/FA, the model can make accurate estimates that can eliminate many protocols from consideration and help select a reasonable, if not the optimum, protocol. Even if the “best” protocol is not selected using predicted data, the cost of this error is probably small. In general, the selected protocol would probably be superior to the protocol constructed by the clinician without the benefit of the model.

The model can also be used to study in detail the impact of criterion, individual test performance, and test correlation on protocol performance. A better understanding of these three factors may provide qualitative information that can assist in selecting test protocols when minimum information is available.

**Acknowledgment.** We would like to thank Dr. Frank Musiek for providing the manuscript containing the data that were used to evaluate the model.

**REFERENCES**