Computer Simulation Technique for Assessing Pediatric Auditory Test Protocols

Anne Marie Tharpe*
Daniel H. Ashmead*

Abstract

Adaptive testing procedures are widely used for evaluating the hearing of young children. A standard protocol for such testing, however, has not been recommended and, as a result, many variations of the procedure are used clinically. This study, by utilizing computer simulations, varied several test parameters and examined the resulting test outcomes. We evaluated the effects of starting level, stimulus step size, and the use of conditioning trials on test outcome, while also varying the hearing levels and false positive and negative rates of our simulated subjects. Results indicated that a low starting level, with a 20-dB down 10-dB up step size, and no conditioning trials produced the most accurate estimates of thresholds under most conditions.

Key Words: Computer simulation, conditioning, false negative rates, false positive rates, starting level step size, visual reinforcement audiometry

Adaptive testing, used frequently when little is initially known about threshold level, is an approach in which the stimulus level is altered according to the responses obtained throughout the test. Although various adaptive procedures are used, the general strategy is to decrease the stimulus level after correct responses and increase it after incorrect responses. Adaptive procedures are used because they provide an efficient way of estimating an initially unknown threshold.

Visual reinforcement audiometry (VRA) has proven to be an effective method for assessing hearing sensitivity in infants as young as 6 months of age (Wilson et al, 1976). Typically, VRA is used in the context of an adaptive test strategy. No standard protocol exists for VRA, however, and numerous variations of this procedure are used in infant research laboratories and in clinical settings (Nozza and Wilson, 1984; Thompson and Folsom, 1984; Olsho et al, 1987; Diefendorf, 1988; Gravel, 1989; Thompson et al, 1989). These variations are made in step size, starting intensity level, and the use of conditioning trials. It is important to consider the effects of these variations on the outcomes of research studies and clinical decision making.

In general, a test procedure should maximize the likelihood of an accurate threshold estimate in an efficient manner. The emphasis on efficiency is paramount when testing infants or young children. Previous research has revealed that infants below 2 years of age will usually participate readily for only about 10 to 50 trials prior to onset of fatigue and distractibility (Moore et al, 1975; Trehub et al, 1980; Primus, 1987; Thompson et al, 1989). In general, test sessions are expected to last approximately 20 minutes (Primus, 1991).

The accuracy of an audiometric test procedure refers to its ability to approximate true sensitivity threshold. Neither over- nor underestimations of true threshold are desirable from a clinical perspective. It is undesirable to alarm a family by suggesting that their infant may have a hearing impairment when one does not exist. In addition, an audiologist's time and resources are limited and are not well spent retesting normal-hearing patients. Similarly, the potential delay in intervention by misdiagnosing a hearing-impaired child may be detrimental to his or her ultimate academic and communicative success. Of course, there is a trade off. Efforts made to reduce overestimations of thresholds result in an increase in underestimations of thresholds, and vice versa. The extent and di-
rection of error one is willing to accept must be pre-determined when selecting test procedures. This study examined the effects of modifications of test procedure on auditory test outcome. We conducted computer simulations of the test procedures, so that sample sizes of sufficient magnitude could be used to allow subtle but systematic effects to be observed. Such testing on human infant subjects would be prohibitive in terms of experimenter and subject time when one considers the numbers of infants and test sessions that would be required in order to examine several test parameters and the statistical properties of the test methods. Previous reports involving computer simulations of auditory testing have emphasized the effects of starting sound level, step size, probability of task orientation, and, to a limited degree, hearing level of the subjects (Kollmeier et al, 1988; Green et al, 1989; Marshall and Hanna, 1989; Eilers et al, 1991). Our simulations incorporated the effects of these factors as well as several others. One of these additional factors was the use of initial conditioning trials, in which subjects are trained for the testing situation before the threshold estimation procedure itself is started. Another factor that we considered was response error rates, both false positives (apparently responding when a sound was not really detected) and false negatives (failing to respond when a sound was presumably detectable). These error rates are associated with factors such as observer bias and subjects' attentional and motivational states. We also considered the interaction of test procedure factors with an extensive range of hearing capabilities of the subjects. A fundamental consideration with regard to clinical testing is that test procedures that are optimal for subjects whose thresholds are in the normal hearing range may be inappropriate for estimating thresholds of subjects with hearing impairment, or vice versa. Therefore, we evaluated the effects of test procedures for subjects with hearing sensitivity ranging from normal to profound hearing loss. The following is a consideration of the rationale for each of the procedural factors that were systematically varied.

**PROCEDURAL FACTORS**

**Conditioning Trials**

Clinicians differ in the use of conditioning trials prior to initiating the test session. For example, in some test paradigms the auditory stimulus may be presented simultaneously with the visual reinforcer in order to condition or shape the infant response (Thompson and Folsom, 1984; Gravel, 1989, 1989; Thompson et al, 1989). Others may skip the conditioning trials altogether and visually reinforce the infant only after initially responding correctly (i.e., head turn) to the auditory stimulus (Thompson and Folsom, 1981; Ashmead et al, 1991; Eilers et al, 1991). There is ample evidence to suggest that conditioning is not necessary to teach infants the localization response if they are at least at a 6-month developmental level because infant orientation is spontaneous when an auditory stimulus is suprathreshold (Hoversten and Moncur, 1969; Moore et al, 1975; Thompson and Folsom, 1984; Tharpe et al, in preparation). Exceptions to this may be found with unilaterally hearing-impaired individuals and those with neurologic dysfunction. Eliminating the conditioning trials could allow for additional, and more valuable, test trials.

**Responsivity**

Although direct evidence regarding the effects of nonsensory factors on infant test performance is not available, one must consider that behavioral test procedures, regardless of the specific parameters, are not immune to the influence of infant responsiveness. Anyone having experience in the testing of infants can attest to the presence of intra- and inter-subject variability in responsiveness to auditory stimuli. At least part of that variability may be attributed to factors such as attention, motivation, cognition, and response bias (Werner and Bargones, 1992). One way that these effects are manifested is in false positive and false negative rates. A false positive is an apparent response, either on a control trial when no stimulus has been presented, or on a trial when the stimulus is below the subject's true threshold. A false negative is an apparent lack of response when the stimulus is presumably well above the subject's true threshold. Generally, false positive rates are monitored by means of inserting control trials into the test session. Such trials in experimental environments have revealed false positive rates in infants ranging from 8 to 30 percent (Nozza and Wilson, 1984; Thompson and Folsom, 1984; Eilers et al, 1991). Although less well documented, false negative rates in infants have been found to be greater than 0 percent (Moore et al, 1975, 1977; Trehub et al, 1980; Olsho et al, 1988; Ashmead et al, 1991).
This certainly suggests that variations in factors that affect false positive and false negative rates, such as infant state, should be considered when evaluating the accuracy and efficiency of test procedures.

Eilers et al (1991) tested the effects of a factor that they called probability of task orientation. This is analogous to what we call the false negative rate. For example, if the probability of task orientation is 0.80, then the false negative rate is 1.00 - 0.80, or 0.20. Thus, the effects of probability of task orientation, as reported by Eilers et al (1991), correspond to the effects of false negatives as reported in the present paper. A careful examination of the Eilers et al (1991) procedure shows that on test trials the false positive rate was always zero. In other words, their simulated subjects never responded on test trials when the stimulus level was well below threshold. The false positive and false negative rates are logically independent, although only an empirical evaluation could establish the extent to which they are correlated in practice. In the present paper, false positive and false negative rates were varied independently of one another. In summary, the present paper differs from that of Eilers et al (1991) in that false positive rates were varied, and in that the false positive and false negative rates were considered to be potentially independent of each other.

Starting Level

Several investigators have reported on the effect of starting level on the accuracy or efficiency of auditory threshold estimation (Thompson and Folsom, 1984; Thompson and Wilson, 1984; Eilers et al, 1991). These investigators, using human research subjects and computer simulations, found that both efficiency and accuracy are improved when the starting level is close to the subject's true threshold. This makes sense when one considers that starting at too high (or low) an intensity level could result in several wasted responses while trying to reach threshold. In addition, starting too many steps away from threshold may result in a biased estimate of threshold.

Step Size

The degree of intensity change in the stimulus level is known as the step size. As with starting level, the step size used in auditory testing varies across investigators and clinics (Moore and Wilson, 1978; Nozza and Wilson, 1984; Thompson and Wilson, 1984; Gravel, 1989; Primus, 1991). Precision of threshold estimation is dependent upon step size. The larger the step size, the more imprecise threshold estimation will be (Levitt 1970, 1978). When considering the testing of infants, however, we are limited in test time and generally must compromise with a larger step size in order to converge on threshold more rapidly. For clinical VRA purposes, downward step size is usually 10 or 20 dB HL while upward step size is 10 dB HL. Levitt (1978) has pointed out that it is common audiologic practice to halve the initial downward step size after the first reversal or "no" response. This is designed to enhance efficiency early during the procedure and precision later on.

METHOD

A computer simulation was designed to follow a modified one-up, one-down staircase algorithm. The procedure assumed a trial structure in which the observer determined from the subject's behavior whether a sound had been heard (i.e., a localization response had occurred). In this case, the observer would respond "yes." The rules for all sessions were as follows. Two consecutive "yes" responses at the initial value were required for the downward threshold search to begin. If no response occurred at the initial value, the value increased by 20 dB until two consecutive "yes" responses occurred. The stimulus then decreased one step following each "yes" response until a "no" response occurred, then the stimulus increased one step until the "yes" response occurred again, and so on. If the response criterion was not reached, no threshold was estimated and the subject was labeled untestable. Threshold was defined as the lowest intensity level at which two "yes" responses occurred out of three stimulus presentations. This model was selected because it is believed to be most representative of general clinical practice. Martin and Sides (1985) conducted a survey of audiologists and found that 67 percent use a two-response stopping rule when conducting routine clinical pure-tone testing. This two-response stopping rule has also been used in research VRA procedures (Primus, 1991). A review of audiology literature also revealed that a down 20 dB, up 10 dB, or down 10 dB, up 10 dB, are the two most common VRA step sizes employed (Moore and Wilson, 1978; Nozza and Wilson, 1984; Thompson and Wilson, 1984;
Gravel, 1989; Primus, 1991). For each combination of test parameters, 10,000 simulated subjects were tested at each of 13 hearing threshold levels, between 0 and 120 dB in 10-dB increments.

The psychophysical function that was simulated for each subject was as follows. There was a “true threshold” at a given sound level, with the probability of a “yes” response set at .50. All lower sound levels were assigned a probability of correct responding at whatever “floor level” was in effect, and all higher levels were assigned a probability at whatever “ceiling level” was in effect. For example, with a true threshold at 40 dB, a floor level at .20 (false positive rate of .20), and a ceiling level at .90 (false negative rate of .10), the psychophysical function was at .20 up to 30 dB, .50 at 40 dB, and .90 for 50 dB and up. Given that the stimulus levels were in 10-dB steps, we assumed that the psychophysical function went from floor to ceiling over a range of 20 dB (approximately centered on one of our sound levels). With the psychophysical function set as just described, the occurrence of a “yes” response (i.e., an apparent detection) at a given stimulus level was determined by a conventional algorithm based on a random number generator.

Numerous parameters were varied, resulting in 5,616 combinations:

1. false positive rates varied between 0 and 50 percent in intervals of 10 percent;
2. false negative rates varied between 0 and 50 percent in intervals of 10 percent;
3. starting levels of 30 or 60 dB;
4. step sizes of:
   a. up 10 dB, down 10 dB or
   b. up 10 dB, down 20 dB or
   c. 20 dB down until the first reversal, then 10 dB down, 10 dB up (referred to as the modified step size);
5. use of conditioning trials, thus simulating the initiation of the false positive responses from the start of the test, or no conditioning trials, thus simulating the initiation of the false positive responses after the first correct response; and
6. hearing levels between 0 and 120 dB in 10-dB increments.

**RESULTS AND DISCUSSION**

The accuracy of a particular test procedure was determined by examining the proportion of subjects whose estimated thresholds were either over or under true thresholds by greater than 20 dB. These calculations are clinically relevant when one considers that the data obtained from such procedures are used not only for classifying a child as normal hearing versus hearing impaired, but also for selecting appropriate hearing aid settings. These accuracy statistics were computed for all combinations of the test factors that were varied. The following discussion includes information on the total data base. For purposes of clarity and brevity, however, the figures represent only the results for subjects with hearing levels of 10, 60, and 90 dB. These hearing levels present a representative sampling of procedural effects across a wide range of hearing levels. In addition, only false positive and negative rates of 0, 10, and 30 percent are included in the figures. The 0 percent false positive and negative rates are presented in order to demonstrate the accuracy of the computer simulation under ideal conditions, while the 10 percent and 30 percent false positive and negative rates are considered more typically representative of easy to test and more difficult to test subjects.

**Comparison of 30- and 60-dB Starting Levels**

Thresholds obtained when using a 30-dB starting level were found to be either more accurate or equivalent to thresholds obtained with a 60-dB starting level for the majority of the simulations. These findings held for conditioned versus unconditioned trials, and the three step sizes, under a variety of false positive and false negative rates for the total data base. In general, a starting level of 60 dB overestimates true threshold when true threshold is in the normal or mild loss range. For example, as seen in Figures 1A and 1B, in 37 percent of normal hearing subjects tested under good conditions (10% false positive and negative rates), threshold was overestimated by more than 20 dB when starting level was 60 dB, compared to only 4 percent of subjects when starting level was 30 dB.

A review of the total data base revealed that when true threshold was in the moderate to profound hearing loss range, a 60-dB starting level yielded similar estimated thresholds as those obtained with a 30-dB starting level. In these hearing-impaired subjects, the 60-dB starting level provided an advantage over the 30-dB starting level only when there were conditioning trials at the beginning of the session and the
Figure 1A  Distribution of thresholds for trials with a 30-dB starting level, 10 dB down, 10 dB up step size, and conditioning trials.

Figure 1B  Distribution of thresholds for trials with a 60-dB starting level, 10 dB down, 10 dB up step size, and conditioning trials.
false positive rate was high. Under those conditions, as can be seen in the bottom rows of Figures 1A and 1B, the 30-dB starting level underestimated true thresholds by more than 20 dB for 90 dB hearing losses in 19 percent of subjects, compared to only 7 percent of subjects with a 60-dB starting level. Only 1 percent of these threshold underestimations, however, resulted in a misclassification of normal hearing instead of hearing impaired. When conditioning was not utilized with this hearing-impaired group, results were equivalent with the two starting levels. These findings indicate, somewhat surprisingly, that even when hearing impairment is suspected, it is reasonable to use a starting level of 30 dB. The use of a 30-dB starting level rather than a 60-dB starting level has the advantage of minimizing the number of subjects with normal hearing who are incorrectly diagnosed as having hearing impairment. For subjects who do have hearing loss, use of 30-dB starting level is unlikely to result in their being diagnosed as having normal hearing. For purposes of estimating threshold in the context of hearing aid evaluation, the estimate will not be biased, provided that the examiner chooses not to use early conditioning trials.

Our conclusion that a 30-dB starting level has advantages over a 60-dB starting level is somewhat different from the effects of starting levels reported by Eilers et al (1991). They compared starting levels of 30 and 50 dB, concluding that the higher starting level did not produce a substantial increase in average threshold estimate (although this increase was statistically significant). In contrast, our findings suggest that using a 60-dB starting level would cause an undesirable increase in the number of subjects with normal hearing who appear to have elevated thresholds. We think it is essential to consider the interaction between starting level and the false negative rate. If testing starts at a high level and the subject “never misses” (false negative rate = 0, as Eilers et al, 1991, assumed), then the true threshold will be reached quite readily. However, if the subject fails to respond to readily detectable stimuli, then the threshold estimate is likely to be misleadingly high.

Comparison of Step Sizes

Examination of the total data base revealed that a 10 dB up 10 dB down step size and the modified step size overestimated true threshold about 20 percent of the time regardless of starting level, false positive and negative rates, or conditioned versus unconditioned trials. These step size rules can overestimate (> 20 dB) as much as 88 percent of threshold under certain adverse conditions. Figures 1B, 2A, and 2B illustrate this trend to overestimate thresholds of subjects with true thresholds of 10 dB. These

![Figure 2A](image_url)  
**Figure 2A**  
Distribution of thresholds for trials with a 60-dB starting level, 20 dB down, 10 dB up step size, and conditioning trials.
findings suggest that using equivalent upward and downward step sizes introduces systematic overestimation bias in threshold estimations. Fortunately, using a 10 dB up 20 dB down step size circumvents this bias (approximately 6% of thresholds are overestimated by > 20 dB), although this underestimates true threshold < 20 dB) for moderate to profound hearing losses. This trend is more pronounced for conditioned versus unconditioned trials (underestimation = 19% and 9%, respectively) and when the false positive rate exceeds 10 percent. This underestimation of thresholds, however, rarely (2%) resulted in a misclassification of normal hearing when true threshold was in the hearing-impaired range.

Comparison of Conditioned and Unconditioned Trials

For subjects with hearing losses, thresholds obtained when conditioning trials were not used were found to be more accurate than thresholds obtained when conditioning trials were used. These findings are consistent with our assumption that utilizing conditioning trials prior to starting the test session initiates the start of false positive responding earlier than when conditioning is not utilized. As can be seen in Figures 3A and 3B, initiating false positive responses from the start of the test session (30-dB starting level) by including conditioning trials makes no difference when the subject thresholds are below the intensity level used for conditioning (30 dB). The negative impact on the accuracy of the thresholds, however, increases as the degree of hearing loss increases. At a hearing level of 90 dB, for example, with subjects exhibiting a less than ideal behavioral state (false positive rate of 30% and false negative rate of 10%), 34 percent of the thresholds were underestimated by more than 20 dB. By comparison, without conditioning trials, the same false positive and negative rates led to underestimation of only 9 percent of the thresholds. This trend also held for the test session having a 60-dB starting level but only when true threshold was greater than 60 dB. When behavioral conditions were more favorable (false positive rate of 10% and false negative rate of 10%), no differences between thresholds obtained from conditioned or nonconditioned test trials were noted for sessions using either starting level.

In most test situations, infants are not likely to make target responses before the stimulus reaches their minimum response level unless they have previously seen the reinforcer activated. For that reason, it seems reasonable to eliminate conditioning trials from the test session and to avoid activating the reinforcer within view of the infant for any other reason prior to initiation of the test session.
Finally, we compared the efficiency of what we believed to be the “best” and “worst” simulated test protocols (i.e., a 30-dB starting level with no conditioning and a 60-dB starting level with conditioning). Efficiency was assessed by determining the number of trials needed to reach threshold during test sessions. With all other factors being held constant (i.e., step size and false positive and negative rates) there was essentially no difference in the efficiency of the two protocols. For example, the mean number of trials needed to reach threshold for the 30- and 60-dB starting levels were 7.0 and 8.5 (SD = 2.0, 2.0), respectively, when true threshold was 0 dB. Similarly, with a true threshold of 100 dB, the mean number of trials for the 30- and 60-dB starting levels were 22.0 and 16.5 (SD = 3.5, 3.5), respectively. These findings are consistent with those of Thompson and Folsom (1984) who also found no difference in efficiency between these two protocols when testing human infants. One must keep in mind, however, that the conditioning trials used prior to the start of the test sessions were not included in the total number of trials.

CONCLUSIONS

One difficulty facing clinical audiologists when testing infants is uncertainty about the conditions that may be influencing test outcomes. We do not, for example, have a priori knowledge of true threshold or of false positive and false negative rates. We can consider, however, the test parameters that may affect the likelihood of an accurate test outcome regardless of those factors. In other words, we can choose test parameters that will be most likely to yield accurate results under adverse or good test conditions, whether a child’s false positive rate is high or low, and whether hearing is normal or in the profound loss range.

Attempts to evaluate such parameters in human infants are difficult, if not impossible, due to an inability to control infant behavior. Further, the sheer number of infant subjects and procedures needed for evaluation in order to be statistically valid would be prohibitive. We have effectively modeled a process by which we can evaluate a number of test parameters. This model expands on those previously used for simulated evaluations of adaptive procedures by incorporating some reflection of infant state (i.e., false positive and negative rates), their interactions upon a wide range of hearing levels, and the impact of using conditioning trials versus no conditioning trials at the beginning of a test session.

Taking all parameters into consideration (i.e., false positive and negative rates, hearing levels, starting levels, step sizes, and conditioned versus unconditioned trials), accurate
estimations of true thresholds were most often obtained utilizing a 10 dB up, 20 dB down step size, a starting level of 30 dB and no conditioning trials. Conditioned responses provided no advantages for any of the combinations of clinical parameters but did introduce some disadvantages. The disadvantages included underestimation of thresholds, which was more pronounced as hearing loss increased, and inefficiency by using up a number of valuable infant responses prior to starting the threshold search. Previous research has indicated that conditioning is usually not necessary to teach an infant to turn towards a sound source. Localizing to a sound of sufficient amplitude is spontaneous for infants by 5 to 6 months of age and should only prove to be difficult in cases of unilateral or asymmetric hearing loss, or certain neurologic disorders. Even premature infants have been shown to turn readily to a sound source as long as their motor or mental ages are at least at a 6-month level (Moore et al, 1992; Tharpe et al, in preparation). Conditioning trials, therefore, seem to add no advantage to the test procedure and use up a number of valuable infant responses prior to starting the threshold search. (Of course, some applications of VRA may require a conditioned head turn where a localization response is not expected. For example, an experiment on pitch discrimination might use a measure of head turning in response to a pitch change. The present recommendation applies only to situations in which an unconditioned tendency to turn is reasonably expected.)

A starting level of 30 dB may, under certain circumstances, be too low, for example if true threshold is in the moderate loss range or higher. As long as one is looking for a spontaneous head turn, however, it takes only a few seconds to increase stimulus level by 20 dB until a head turn is noted, and the infant has wasted no responses in the meantime. Responses are wasted when starting at too high a level and then descending before an actual bracketing of threshold can begin.

A step size of 10 dB up, 20 dB down underestimates true threshold for moderate to profound hearing losses when the false positive rate is greater than 10 percent and is hardly affected by false negative rate. Fortunately, this underestimation was rarely so great as to result in an estimation of normal hearing when true threshold was in the hearing loss range. The 10 dB up, 20 dB down step size is much less likely to overestimate true threshold than either of the other step sizes examined.

Although adaptive procedures can provide an effective approach to threshold estimation, the rules for stimulus adjustment are often selected with limited information about the impact on test outcome. The results of the present study suggest that use of the following rules...
minimizes the inherent test bias in this adaptive test procedure, while still allowing rapid estimation of thresholds.

1. Initial level of 30 dB, increasing in 20-dB steps if no response occurs.
2. After initial level is determined and replicated, step size of 20 dB down, 10 dB up should be used on the ascending and descending runs.
3. Conditioning (i.e., pairing stimulus and reinforcer) should not be utilized as a matter of routine, but rather the examiner should begin testing without conditioning.

In repeated test sessions, it is important to keep in mind that false positive responses could begin from the start of the session even without conditioning if the child remembers the previous test session. It is suggested, particularly during these conditions, that the audiologist maintain a formal method for inserting catch, or control, trials into the test procedure. Catch trials are those in which conditions are acceptable for a stimulus to be presented (i.e., the infant is in a ready state), but instead of presenting a signal, the audiologist observes whether or not a false positive response occurs. It is important to note, however, that a false alarm rate estimate is not very accurate unless a considerable number of catch trials are obtained (i.e., 15 to 20). According to our findings, if the false alarm rate is greater than about 10 percent, there is a strong possibility that the thresholds of subjects with hearing impairment will be seriously underestimated (> 20 dB).

Acknowledgment. The authors thank Dr. Allan Diefendorf for his helpful review of this manuscript and Don Riggs for his preparation of the figures.

This paper was presented, in part, at the American Academy of Audiology Annual Convention in Nashville, Tennessee, in April of 1992.

REFERENCES


