Observational Measures of Auditory Sensitivity in Early Infancy

Sandra E. Trehub, Bruce A. Schneider, Leigh A. Thorpe, and Peter Judge
Centre for Research in Human Development
University of Toronto, Erindale Campus
Mississauga, Ontario, Canada

A rigorous observational procedure is described for estimating the detectability of auditory signals in the first few months of life. This procedure, a modification of the observer-based psychoacoustic procedure, generates a bias-free estimate of auditory sensitivity and avoids some of the potential problems associated with alternative techniques. Data from infants 1.5, 2.5, and 3.5 months of age revealed orderly improvement in performance as a function of increasing signal intensity and age. Finally, the way in which the procedure can be used to screen infants for auditory deficits is described.

There have been considerable gains in our understanding of auditory sensitivity in infancy (Berg & Smith, 1983; Olsho, Koch, Carter, Halpin, & Spetner, 1988; Schneider, Trehub, & Bull, 1980; Sinnott, Pisoni, & Aslin, 1983; Trehub, Schneider, & Endman, 1980), but these gains have been largely restricted to infants 5 months of age and older. Few clinicians or researchers question the importance of very early auditory assessment (e.g., Downs, 1978), so limited progress on this front is not attributable to insufficient effort but to the special difficulty of identifying appropriate response measures in the first few months of life (Schneider & Trehub, 1985; Schneider, Trehub, & Bull, 1979). Indeed, researchers have sampled an unusually wide range of responses (see Schneider & Trehub, 1985; Schneider, Trehub, & Bull, 1979), none of which has proved useful with low-intensity signals. As a result, the predominant research and clinical focus has been on relatively intense signals that induce reflexive responses (Northern & Downs, 1978), leading to the prevailing clinical assumption that precise measurements of auditory sensitivity are simply not possible in the early months of life (e.g., Hodgson, 1985).

Some recent attempts to estimate behavioral thresholds at this age (Olsho et al., 1988; Olsho, Koch, Halpin, & Carter, 1987; Schneider & Judge, 1988; Trehub & Schneider, 1986) represent a notable departure from this trend. Rather than focusing on the infant and designating a target response (e.g., head turn, eye movement), these researchers have focused on the observer, who can use unspecified criteria to judge the presence or absence of a signal. The observer receives immediate feedback about such judgments and, in some cases (Olsho et al., 1987, 1988), the infant is reinforced for the observer's correct judgments.

In the first published investigation along these lines, Olsho et al. (1987) estimated the sensitivity of infants 3–5, 6–9, and 10–12 months of age to pure tones of 250, 1000, and 4000 Hz. The inclusion of older infants made it possible to compare performance on the observer-based procedure with that derived from more traditional procedures (e.g., conditioned head turning). In addition, it allowed for the possible estimation of age-related changes in sensitivity. Olsho et al. (1987) found performance of the youngest age group to be about 10 dB poorer than that of the older infants at 250 and 1000 Hz but more or less equivalent at 4000 Hz. Moreover, the 6-month-olds, for whom data were available on the observer-based procedure (Olsho et al., 1987) as well as on the head-turn procedure (Olsho, 1985), performed similarly on both tasks. Presumably, then, the observer-based psychoacoustic procedure (OPP; Olsho et al., 1987) might provide comparable information at younger ages, before the head-turn procedure can be used productively (Moore, Wilson, & Thompson, 1976).

Olsho et al. (1988) extended the previous study of pure-tone sensitivity by including a wider range of frequencies: 250, 500, 1000, 2000, 4000, and 8000 Hz. Whereas thresholds of the 6- and 12-month-olds were approximately 10–15 dB higher than those of adults, those of the 3-month-olds were about 15–30 dB higher. Moreover, the differences between adults and older infants were greatest at the lower frequencies, as in previous research (Schneider et al., 1980; Trehub et al., 1980), but the differences between adults and 3-month-olds were greatest at the higher frequencies.

These studies of Olsho and her associates represent a major methodological and empirical advance over previous research with very young infants, but they raise a number of questions. First, Olsho et al. (1987, 1988) provided reinforcement for the infant whenever the observer correctly judged the presence of a signal. The use of reinforcement in auditory detection studies has empirical justification with older infants (Trehub, Schneider, & Bull, 1981), but there is no evidence that it enhances responding to auditory (Moore et al., 1976) or visual (Mayer & Dobson, 1982) stimuli in infants younger than 5 months old. Thus, the apparent sensitivity differences between

This research was supported by grants from the Medical Research Council of Canada and the University of Toronto. We are grateful for the patience and skill of our observer, Marilyn Barraas.

Correspondence concerning this article should be addressed to Sandra E. Trehub, Centre for Research in Human Development, University of Toronto, Erindale Campus, Mississauga, Ontario, Canada L5L 1C6.
3-month-olds and 6- to 12-month-olds may be attributable, at least in part, to the greater efficacy of reinforcement with older infants (Moore et al., 1976; Trehub et al., 1981). If reinforcement with very young infants were ineffectual or differentially effective across age, this would preclude unequivocal interpretation of age-related performance differences.

A second and more central concern involves the way in which Olsho and her associates estimated signal detectability for the various age groups tested. In our view, signal detection approaches offer an alternative that minimizes some of the difficulties associated with other approaches to infant auditory psychophysics. Signal detection theory (Green & Swets, 1974; Swets, 1964, 1973) or, simply, detection theory (Macmillan & Kaplan, 1985) provides a powerful means of separating the contribution of stimulus characteristics from other nonsensory factors that can affect a listener's responsiveness during test trials. These nonsensory factors can be particularly troublesome when different age groups or procedures are to be compared. Because the theory allows for the conversion of data derived from different situations into comparable units of measurement, it is applied to infant auditory psychophysics confer numerous advantages (Thorpe, 1988). The basic tenets of signal detection theory, as they relate to the present article, are outlined in the Appendix.

Signal detection theorists have strongly advised against the practice of determining thresholds primarily from hit rates, even when a correction for guessing is used (see Egan & Clarke, 1966, pp. 236–240). This advice is especially relevant in the case of infants, where the specific consequences of training regimen, stimulus probability, and payoff are often unknown. In much research with young infants, however, false-alarm rates (i.e., responding incorrectly on no-signal trials) are used to evaluate the degree of stimulus control on the part of the infant (e.g., Berg & Smith, 1983; Sinnott et al., 1983) or the presence of adequate skill on the part of the observer (Olsho et al., 1987, 1988) and are not considered to reflect sensitivity. Moreover, subjects who exceed specified false-alarm rates are typically excluded from the data set. The underlying rationale for this practice is that a poorly trained infant or observer would merely contribute troublesome variability to the data, obscuring the true level of sensitivity. This may be reasonable in the case of highly salient stimuli, where the noise and signal-plus-noise distributions are far apart. It is considerably less reasonable, however, when the distributions have considerable overlap (see Figure A1) and when a low false-alarm rate also implies a low hit rate. Experience with infants tested on yes/no discrimination tasks has revealed that the false-alarm rate is not constant but can vary with task difficulty or the discriminability of the signal from noise (Thorpe & Trehub, 1989; Trehub & Thorpe, 1989). Indeed, it is common to see difficult discrimination tasks lead to increased false-alarm rates as well as to decreased hit rates. In fact, when an infant is unable to hear a signal, the probability of hits and false alarms should be nearly equal.

If signal detection theory provides a reasonable description of infants' performance in auditory detection tasks, then there may be serious implications of discarding data on the basis of false-alarm rate. The exclusion of infants with high false-alarm rates differentially eliminates infants with low sensitivity, because false-alarm rate is related to sensitivity. Also, the exclusion of infants with flat or nonmonotonic psychometric functions (Olsho et al., 1987) further constrains the outcome. As a result, infants who remain in the sample may exhibit greater sensitivity or lower thresholds than does the population that they purportedly represent. One can evaluate the consequences of such selection criteria by comparing the to-be-discarded data with the to-be-retained data. At the very least, one should demonstrate that those excluded are outliers. We would argue, however, for the use of subject selection criteria that are independent of sensitivity.

The dependence of the false-alarm rate on signal intensity has particular implications for adaptive procedures. Adaptive procedures are selected over alternative procedures (e.g., method of constant stimuli) because of their efficient movement to the intensity region of greatest interest, namely the near-threshold region. What is often forgotten or ignored is that the various adaptive algorithms in use, such as parameter estimation by sequential testing (PEST; Taylor & Creelman, 1967), up-down transformed response (UDTR; Levitt, 1971), maximum likelihood (Hall, 1968), and the hybrid adaptive procedure (Hall, 1981), were developed for use with tasks that involve at least two response alternatives on each trial (e.g., in which of two intervals did the signal occur?) and signal presentation on all trials, rather than yes/no or one-interval detection tasks (e.g., Is the signal present?), which involve no-signal as well as signal trials and essentially one response. Nevertheless, most investigators who have used adaptive techniques with infants (e.g., Berg & Smith, 1983; Olsho et al., 1987, 1988) have used yes/no tasks (generally go/no-go; but see Trehub, Bull, Schneider, & Morrone, 1986, for an exception) despite the fact that convergence on threshold has only been defined for two- or four-alternative, forced-choice tasks and despite the further difficulty that false-alarm rate cannot be estimated accurately in the context of variable signal levels.

Although no-signal trials are invariably included in such adaptive studies of infant threshold, how can they be linked meaningfully to signal trials at any particular intensity level? A potential problem with adaptive techniques is that, during the course of a session, observers can change the location of their criterion relative to the mean of the noise distribution. Consider, for example, two different strategies that could be adopted by observers instructed to keep their false-alarm rate under 25% in an adaptive technique. One strategy that would accomplish this is to locate their criterion .675 standard deviations to the right of the mean of the noise distribution so that the probability of a false alarm remains constant at 25% throughout the session. A second strategy could have observers locate their criterion at the intersection of the noise and the signal-plus-noise distributions. Note that an observer following this strategy would have a low false-alarm rate at the beginning of a session when signal intensity is at its maximum and the separation between noise and signal-plus-noise distributions is large. In an adaptive procedure, however, signal intensity decreases over the course of a session, bringing noise and signal-plus-noise distributions closer together. In line with this second strategy, as the separation between these two distributions decreases, the location of the criterion is moved closer to the mean of the noise distribution, with the probability of false alarms increasing over the session. This means that a high false-
alarm rate at the end of the session could be balanced against a low false-alarm rate at the beginning of a session so that the average rate remains below 25%. In the Appendix, we illustrate how these two strategies can lead to quite different psychometric functions even though the average false-alarm rate is the same for both. Furthermore, applying the conventional correction for guessing to these functions produces rather different estimates of threshold.

It should be clear, then, that the application of adaptive procedures to yes/no signal-detection tasks does not constrain the subject to use a constant criterion. Without such constraints, very different psychometric functions and threshold estimates can result, depending on the observer's strategy. This is an important argument against the use of adaptive procedures with yes/no signal-detection tasks.

By contrast, two-alternative, forced-choice techniques (e.g., Trehub et al., 1980, left or right head turn) involve signal presentation on every trial, providing information about noise and signal distributions at all times so that subject motivation, response bias, and criterion do not affect the outcome. Although it is possible to use two-alternative, forced-choice procedures with 6-month-old infants in auditory detection tasks (e.g., Schneider, Morrongiello, & Trehub, 1990; Schneider et al., 1980; Schneider, Trehub, Morrongiello, & Thorpe, 1989; Trehub et al., 1980, 1981) and with younger infants in visual preference tasks (Teller, 1979), it is not yet clear whether forced-choice auditory tasks are feasible with younger infants. Until such tasks are feasible, one can select other procedural details to maximize their compatibility with yes/no techniques, but how can we evaluate their efficacy? Conventional evaluative criteria in this domain include reasonable response-intensity functions (i.e., response probability increasing systematically with stimulus intensity) and relative freedom from bias on the part of the observer (see Schneider & Trehub, 1985).

In the present article, we report on our implementation of observer-based measures for estimating the detectability of auditory signals in the first few months of life. Our procedure was essentially a variation of the OPP of Olsho and her associates. In line with Teller's (1979, 1985) research on infant vision and Olsho et al.'s (1987, 1988) research on infant audition, our observers remained unaware of the presence of a test signal and could use any available cues to support their judgments. In contrast to the above-mentioned investigators, however, we omitted infant reinforcement to preclude interpretive problems associated with potential differences in the efficacy of reinforcement across the three age groups of the present investigation. It is important to remember, however, that the observer rather than the infant is the true subject and that the observer is reinforced by means of trial-by-trial feedback. Presumably, such reinforcement shapes the observer's behavior so that cues from the infant are used optimally. We also presented signal and no-signal trials with equal probability rather than setting the probability of signal trials at .65 (Olsho et al., 1987, 1988).

Our replacement of Olsho et al.'s (1987, 1988) adaptive procedure with trials at a uniform intensity level was motivated by two factors. As outlined earlier, we consider adaptive algorithms to be inappropriate in conjunction with yes/no procedures. One alternative is the method of constant stimuli, which involves the random presentation of a fixed number of different intensity levels. Because the uncertainty associated with stimulus variability may adversely affect infant performance, as it sometimes does for adults (Cohen, Trehub, & Thorpe, 1989; Espinoza-Varas & Watson, 1986), we opted for uniform values of stimulus frequency and intensity during any single test session. McKee, Klein, and Teller (1985) suggested the use of a single stimulus value with each subject when fewer than 100 trials can be presented. They argued that if normative performance levels can be established, then individual performance below a specific range would be indicative of a potentially important deficit.

We used pulsed octave-band signals rather than pure tones (Olsho et al., 1987, 1988) or constant, octave-band noises (Schneider et al., 1980, 1989; Trehub et al., 1980). The factors favoring such a choice included evidence of lesser infant responsiveness to pure tones compared with broad- and narrow-band noises (Hoversten & Moncur, 1969; M. Thompson & Thompson, 1972), as well as the potentially greater salience of pulsed signals. To further enhance infants' interest in the signals, they were pulsed in an irregular manner; that is, they were interrupted intermittently and unpredictably. Moreover, the pulses moved intermittently between two loudspeakers on either side of the infant.

Another departure from previous versions of the OPP was the exclusion of a formal training phase with each infant. In contrast with Olsho and her associates (Olsho et al., 1987, 1988), who used multiple observers, we used a single, highly experienced observer. In fact, our observer had used the procedure for several months before collecting the present data set and had reached a stable (i.e., asymptotic) level of performance. Olsho et al. (1987) supplemented such general training (maximizing hit rates and minimizing false-alarm rates) with training to a predetermined criterion on each infant (at least four correct out of the last five signal trials and four out of the last five no-signal trials at the initial or maximal intensity level). Some of the consequences were as follows. For 3-month-olds (the youngest infants in their sample), sessions that yielded a threshold estimate required an average of 22 training trials. Moreover, in 40% of all first sessions with these infants, the training criterion was not achieved and additional sessions were required to do so. This protracted training procedure would surely preclude the OPP's use in clinical contexts. Even for research applications, it remains to be determined whether such training achieves its goal of enhancing the observer's performance or bringing it to some asymptotic level. Optimal performance is likely to involve a trade-off between experience with the stimulus (i.e., infant's behavior) and temporal limits on infant attentiveness. A further complication is that the specific infant behaviors that maximize an observer's performance may not be identical for high- and low-intensity signals. For example, attenuation of limb movement may be the characteristic response to signals of high or moderate intensity, whereas subtle eye widening or eyebrow movement may be relevant for low-intensity signals. Thus, there may be limited transfer from training to test levels. Finally, given practical limits on the number of trials in a single test session, the presentation of training trials would necessitate a corresponding reduction of test trials.

Because the benefits of various training regimens remain unclear, we opted for their total exclusion. Instead, the observer
preceded each session by interacting with the infant (vocally and otherwise) in the waiting room and began each session (in the test booth) by watching the infant for a few minutes in the presence of auditory stimulation (i.e., vocal exchanges with the parent) and in silence. In short, general familiarization with each infant substituted for specific training.

Method

Subjects

Subjects were 237 infants, 1 to 4 months of age, who were recruited primarily by letters sent to local hospitals. All participating infants were healthy, were born at term, had no family history of hearing impairment, and were free of colds at the time of testing. Infants were excluded from the data set if they failed to complete at least 16 trials because of inappropriate state or other factors that precluded adequate observation (7 infants at 1-2 months of age, 14 at 2-3 months of age, and 4 at 3-4 months of age). Typical exclusion criteria were prolonged crying, sleeping, or hiccupping, continual state changes (i.e., crying to fussing to sleeping), postural orientation that occluded the baby's face, and the discontent of some of the older infants with the semireclining position. In addition, 6 infants were excluded because of their mother's repeated attempts to interact socially with them. The final sample of 213 infants included 65 at 1-2 months of age (mean age = 1.56 months, range = 1.1-1.9 months), 62 at 2-3 months of age (mean age = 2.53 months, range = 2.0-2.9 months), and 75 at 3-4 months of age (mean age = 3.53 months, range = 3.0-3.9 months). There were separate groups of infants at the three test intensities, 60 dB (twenty-four 1.5-month-olds, twenty-two 2.5-month-olds, and twenty-nine 3.5-month-olds), 40 dB (twenty-one 1.5-month-olds, twenty 2.5-month-olds, and twenty-four 3.5-month-olds), and 30 dB (twenty-one 1.5-month-olds, twenty 2.5-month-olds, and twenty-two 3.5-month-olds).

Apparatus and Stimuli

Test sessions took place in an Industrial Acoustics sound-attenuating chamber (double wall, 3 x 2.8 x 1.8 m) with loudspeakers (K&F 101) located 45° to the left or right, 1.8 m from the infant's position in one corner of the room. Ambient noise level in the test room was 27 dB. The infant faced the center of the room in the direction of an observer behind a video camera. The attending parent (usually mother) was seated facing the baby, who was propped in a semireclining position on a pillow or molded backrest, which afforded postural support but did not preclude freedom of movement. The observer stood directly behind the mother, where the observer had an unimpeded view of the infant.

The output of a white noise generator (General Radio, Model 1381) was filtered by a programmable bandpass filter (Bruel & Kjaer, Model 1617) set for an octave bandwidth centered at 4000 Hz. The octave bandwidth facilitated potential comparisons with older infants (Tre-hub et al., 1980) and minimized amplitude variation in the sound field (Dillon & Walker, 1982), while retaining specifiable frequency information. The rate of decrease in energy on either side of the band was 30 dB per octave. The output of the filter was routed to an impedance-matching amplifier and, then, to two programmable attenuators (one for each loudspeaker) with electronic switches that shaped the signal to have a rise and decay time of 40 ms.

The signal consisted of a series of four bursts of a 4-kHz octave-band noise with interburst time set to .349 s. The durations of the four bursts were 266, 432, 598, and 764 ms. On any trial, all four bursts were presented (total duration = 3.1 s), with the order of bursts randomly permuted on each trial. During each trial, two of the bursts appeared on the left loudspeaker, the other two on the right, with the order of locations randomly selected on each trial. Sound-pressure levels were calibrated at the approximate location of the infant's head with a Bruel and Kjaer impulse precision sound level meter (Model 2204) and a 0.5-in. microphone.

Procedure

Infants were placed in semireclining positions and observed for a few minutes, before the examiner explained some procedural details to the parent and subsequently for approximately 15-30 s of silence. The duration of silent observation was at the discretion of the observer, who could begin the session when she felt ready to do so. When the infant assumed an appropriate, stable state (quiet alert or light sleep for infants tested at 60 dB, quiet alert only for those tested at 30 and 40 dB) and posture (head centered), the observer pressed a button to initiate a trial, which involved either (a) the presentation of a, pulsed, octave-band noise or (b) a silent interval (i.e., no-signal or control trial) of comparable duration (3.1 s). The rationale for different state criteria at the different intensity levels stemmed from the greater subtlety of overt responses at lower levels and their relative obscurity during sleep states. Note, however, that infants were only tested in light sleep if repeated efforts to arouse them to a quiet alert state were unsuccessful. In the case of the 60-dB signals, extensive pilot testing revealed no indication of differential performance in quiet alert and light sleep (eyes closed, some movement, irregular respiration) states. However, the final sample had only one infant, a 1.5-month-old, who remained in light sleep throughout the test session. Four additional 1.5-month-olds and three 2.5-month-olds were in light sleep for part of the session. Visual inspection of their data revealed no differences potentially attributable to the state of testing.

The occurrence of signal and no-signal trials was randomly determined and unknown to the observer and parent, who wore headphones that delivered signals on all trials. The observer could take as much time as necessary to indicate, by pressing one of two buttons, her judgment regarding the presence or absence of a signal, and she received immediate feedback regarding the accuracy of that judgment. If the observer's view became obscured on a trial by the infant turning to one side or moving forward to a more upright posture, or by the mother interacting with the infant, that trial was not replaced but was simply excluded from the data set. (Trials were excluded infrequently for these reasons) If the infant's state changed, trials were discontinued until the infant could be calmed or aroused (as appropriate). A test session consisted of 32 trials, half signal and half no-signal trials. Three levels of signal intensity (30, 40, and 60 dB) were used, each infant receiving only a single level.

Results

The numbers of hits and false alarms were used to determine $d'$ scores for each infant, which were used to calculate mean $d'$ scores for each group of infants. These $d'$ scores were used to

---

1 Because $d'$ values are indeterminate for probabilities of hits or false alarms equal to 0 or 1, a probability of .02 was assigned when the proportion of false alarms was 0 and a probability of .98 was assigned when the proportion of hits was 1. Two other methods for dealing with proportions of 0 and 1 were also tested. Specifically, 0 proportions were changed to 1/2(N), and perfect proportions were changed to 1 - 1/2(N), where N is the number of no-signal or signal trials, respectively, as suggested by Macmillan and Kaplan (1985). Alternatively, we added 1 to both signal and no-signal trials and $1/2$ to the number of hits and false alarms (Thorpe, Trehub, Morrongiello, & Bull, 1988). Neither of these methods altered the pattern of results observed.
index sensitivity because of their freedom from judgmental biases of the observer. To examine possible changes in the observer's performance over the course of a test session, individual $d'$ scores were determined for test trials from the first and second half of the session for infants who completed all 32 trials. There was no evidence of such changes in performance for any of the age groups tested at 60 dB. At the two lower test levels, however, the observer's performance improved in the second half of the session at all ages. Analyses of variance (ANOVAs) on the first and second half of the test sessions confirmed the presence of a practice effect at 40 dB, $F(1, 51) = 34.74$, $p < .001$, and at 30 dB, $F(1, 47) = 7.67$, $p < .01$, and its absence at 60 dB $F(1, 64) = 1.68$, $p > .1$. At moderate and low test levels, then, the observer's ability to "read" an infant improves over the course of the session.

The $d'$ scores as a function of signal intensity are shown in Figure 1. It can be seen that the mean $d'$ value increased as a function of intensity for all age groups. A two-way ANOVA [Age (1.5, 2.5, and 3.5 months) × Intensity (30, 40, and 60 dB)] revealed significant effects of age, $F(2, 193) = 5.46$, $p < .005$, and intensity, $F(2, 193) = 5.21$, $p < .0001$, but no significant Age × Intensity interaction, $F(4, 193) = 2.38$, $p > .05$. Scheffé tests indicated that performance at every intensity level differed significantly from that at any other intensity level. With respect to age, however, only the 1.5-month-olds differed significantly from the 3.5-month-olds.

It is possible to define threshold for each group as that intensity corresponding to a $d'$ score of .95, because a $d'$ of .95 in a yes/no procedure corresponds roughly to a 75% correct threshold in a two-alternative, forced-choice procedure. From Figure 1, it can be seen that 40 dB is clearly above threshold for all age groups. In fact, the specific thresholds resulting from a criterion of $d' = .95$ are 38.4, 34.4, and 34.1 dB, respectively.

One can tentatively explore the implications of the data exclusion and treatment criteria used by Olsho and her associates (Olsho et al., 1987, 1988). For example, restricting the subject pool to test sessions in which the observer had a false-alarm rate equal to or less than .25 would have resulted in the decimation of our sample despite the fact that our mean false-alarm rate was approximately .25 for the three age groups tested. Nearly half of the subjects would have been eliminated from most groups, including the entire 2.5-month-old sample tested at 30 dB. The critical question, however, is whether infants excluded on this basis differ in sensitivity from those who qualify for inclusion. At the 60-dB test level, four of twenty-four 1.5-month-olds had $d'$ scores below our proposed criterion level of .95; all four had false-alarm rates greater than .25, Olsho's exclusion criterion. Only one 2.5-month-old and one 3.5-month-old had $d'$ scores less than .95, and both of these had false-alarm rates exceeding .25. At 40 dB, five of twenty 1.5-month-olds had $d'$ scores less than .95, four of these infants having false-alarm rates exceeding .25. Also, six of twenty 2.5-month-olds had $d'$ scores less than .95, four of these having high false-alarm rates (an additional one having a .25 rate). Finally, all six of the twenty-four 3.5-month-olds with low $d'$ scores had high false-alarm rates. The $d'$ scores of observer–infant pairs with acceptable and unacceptable false alarms are shown in Figure 2. It would seem, then, that Olsho et al.'s (1987, 1988) false-alarm criterion, when applied to data such as ours, would result in biased sample selection favoring infants with greater sensitivity.

Although infants with high false-alarm rates were overrepresented in the group of infants with low sensitivity, it is unreasonable to assume that all infants with high false-alarm rates had low sensitivity. On the contrary, there were a number of cases in which high false-alarm rates were coupled with even higher hit
rates, yielding d' values well above the mean for that particular group. The obvious conclusion is that it is unreasonable to consider false-alarm rates apart from the context in which they occur; in other words, they must be considered in conjunction with hit rates.

Discussion

In the present study, we demonstrated that minor modifications of a recently developed observational technique (Olsho et al., 1987, 1988) can generate a bias-free estimate of auditory sensitivity in the first few months of life. Our contention is that these modifications enhance the utility of the OPP in a number of ways. First, signal detectability is estimated from both hits and false alarms and does not depend on assumptions of constant false-alarm rates across different intensity levels. Second, it is possible to gather useful information in a single test session from infants as young as 15 months. In effect, the current procedure extends the application of detection theory techniques (Macmillan & Kaplan, 1985) to the measurement of auditory sensitivity in early infancy.

Data gathered with the current technique revealed orderly improvement in performance as a function of increases in signal intensity and of age. In essence, then, the OPP, with the present modifications, meets the criteria outlined earlier: reasonable response-intensity functions and freedom from bias. The obtained thresholds are considerably lower than those generated by other observational techniques (Hoversten & Moncur, 1969; Northern & Downs, 1978; G. Thompson & Weber, 1974), except for the earlier version of the OPP. Note, however, that considerable caution is warranted in threshold comparisons where different conditions of testing prevail. Nevertheless, the similarity of thresholds obtained from both versions of the OPP supports the utility of the OPP as a data-gathering technique and lends credence to the findings in both cases.

On the other hand, we have some reservations about the psychometric functions previously derived from the OPP (Olsho et al., 1987, 1988), because they were based on hits (with false alarms providing a lower limit), biased data-exclusion criteria, and an adaptive procedure developed for four-alternative, forced-choice tasks (see Hall, 1981). Moreover, Olsho et al.'s use of probit analysis (Finney, 1971) for threshold estimation, in the context of relatively few trials at each intensity level, could be considered problematic (O'Regan & Humbert, 1989). In principle, such procedures could result in threshold estimates that are too low or unrepresentative of the population in question.

It is tempting to venture tentative comparisons of our own threshold estimates for 3.5-month-olds with those of 6-month-olds tested with somewhat similar signals (i.e., constant octave-band noises) in the sound field (Trehub et al., 1980). Although such comparisons suggest that sensitivity increases by about 13 dB between 3.5 and 6 months of age, the availability of reinforcement for the older infants (see Trehub et al., 1981)2 and numerous other procedural differences preclude unequivocal interpretation of these findings.

The use of reinforcement and the measurement of infant as opposed to observer responses are not the only differences between the techniques used for older and younger infants. Recall that the principal reason for abandoning the focus on infant responses concerned the subtlety and variability of responses to sound in the early months of life. What was clearly evident in the present study was that infants varied considerably in their expressiveness or propensity to exhibit overt reactions to sound. Individual differences in expressiveness are necessarily problematic for any observational measure of auditory sensitivity, so that underestimates of sensitivity are likely to result from relatively inscrutable infants with intact hearing. Another factor that might account for performance differences between 3- and 6-month-olds is signal duration, which was approximately 3 s in the present study compared with its continued availability until a response occurred in Trehub et al. (1980). There is evidence that potentially unlimited signal duration enhances the performance of infants, at least in the context of a two-alternative, forced-choice procedure (Trehub et al., 1981). It is possible, moreover, that a brief signal results in even greater threshold elevation for 1- to 3-month-olds than for 6-month-olds.

The practice effect in the present investigation is both troubling and revealing. To preclude underestimates of infant sensitivity, it would be preferable to restrict the data to periods in which performance was no longer improving. Close inspection of our data revealed, in the case of 40-dB signals, continuing improvement even in the final quarter of the test session. This means that, for all practical purposes (i.e., limited test-session duration), the performance of very young infants will necessarily be underestimated, at least for some intensities. When an adaptive procedure is used (e.g., Olsho et al., 1987, 1988), however, intensity level is confounded with observer experience so that practice effects will be obscured. It is unlikely that training trials such as those implemented by Olsho et al. guarantee stable performance. If this were the case, adult psychophysical studies would have little need for the thousands of trials typically used. In fact, it is not unusual for adults to show continued improvement in threshold over the course of several extended test sessions (e.g., Zwischen, Maire, Feldman, & Rubin, 1958). There are indications, however, that improvement is greater at low than at high frequencies (Zwischen et al., 1958). In the present study, the absence of a practice effect at 60 dB and its presence at 40 and 30 dB suggest that different infant cues are operative at high and low test levels so that training trials at relatively high intensities might have little or no transfer to test trials at lower intensities.

In any case, with the judicious selection of intensity levels, it is possible to generate estimates of group thresholds, albeit conservative ones. On the other hand, our technique is impractical for estimating individual thresholds, which would require testing at three or four appropriately chosen intensity levels. Although the use of an adaptive procedure is attractive, in princi-

---

1 Trehub, Schneider, and Endman (1980) used a two-alternative, forced-choice procedure with head-turn responses and defined thresholds as the intensity corresponding to 65% correct responding. If lateralization thresholds are equivalent to detection thresholds (Schneider, Trehub, & Thorpe, in press), a d' value of .95 in our technique would correspond to 68% correct responding in the two-alternative, forced-choice technique of Trehub et al. (1980). Applying the 68% requirement to thresholds for 6-month-olds in the Trehub et al. (1980) sample would result in a threshold of 21 dB instead of the 19-dB threshold reported. The comparable value for 3-month-olds in the present study is 34.1 dB.
ple, its use is precluded until a two-alternative, forced-choice task can be applied successfully to this population or until the characteristics of single-interval adaptive procedures are better characterized. A crude but productive alternative to individual threshold determination is to test infants at a single, well-chosen intensity level, that is, one derived from normative data. This would provide a reasonable means of screening for age-appropriate auditory acuity, as others have done in the case of visual acuity (Dobson, Teller, Lee, & Wade, 1978; Fulton, Manning, & Dobson, 1978). The presumption is that the diagnostic test level (a specific stripe width, in the case of visual acuity, or intensity level, in the case of auditory acuity) would provide an efficient estimate of the presence of normal sensory functioning, an estimate that has potential clinical utility.

We can use the present data to illustrate the process of defining criterion performance at a diagnostic intensity. If we adopted the relatively stringent criterion of $d' = .95$ on a pass/fail basis, then one of twenty-two 2.5-month-olds and one of thirty 3.5-month-olds in our final sample would have failed at the 60-dB level. Naturally, this so-called failure does not necessarily imply a hearing loss, because the infant in question might simply be inattentive or uncooperative at any particular time. Thus, a second test session along the same lines would be warranted. A repetition of similar negative results would indicate the need for more intensive diagnostic evaluation by medical personnel. In the case of the 1.5-month-olds tested at 60 dB or the 2.5- and 3.5-month-olds tested at 40 dB, the numbers failing (4 of 24, 6 of 20, and 6 of 24, respectively) are highly unlikely to represent infants with hearing problems. However, the use of such test procedures with 2.5- and 3.5-month-olds who are at risk for hearing loss on the basis of family history of childhood hearing impairment, congenital perinatal infection, anatomic malformation of head or neck, birthweight less than 1,500 g, hyperbilirubinemia (at levels warranting exchange transfusion), bacterial meningitis, or severe asphyxia (Joint Committee on Infant Hearing, 1982) is likely to be cost-effective. Moreover, this could result in an earlier diagnosis of moderate hearing losses (60 dB or greater) than is currently the case.

How might the present procedure be improved? One possibility is to change from a signal of fixed duration to one of unlimited duration. This would allow the observer to continue monitoring the infant until sufficient infant cues were available to support a judgment of signal presence or absence. Trehub et al. (1981) found that in the case of 12-month-olds, the majority of long-latency responses to sound were correct responses. Similar accommodation to individual differences in infant responding or to transient attentional changes might be especially fruitful with very young infants. Perhaps the 10-s signal used by Olsho et al. (1987) was motivated by difficulties experienced with signals as short as those of the present investigation. Flexible signal duration would result in signals as short or as long as necessary for specific infant–observer pairs at any particular time.

Another possibility is to attempt a two-interval, forced-choice procedure (2IFC). This would involve two successive intervals of time (e.g., 5 s each), one with signal, one without, with the observer judging the interval (first or second) in which the signal appeared. Such a modification would obviate the need for no-signal trials while maintaining freedom from bias. Moreover, 2IFC could be used in conjunction with an adaptive procedure. It is important to establish whether 2IFC is feasible with an observer-based procedure and whether its advantages outweigh possible losses associated with fixed rather than flexible signal durations.

Finally, it is important to establish the contribution of infant reinforcement and observer training trials to performance accuracy. One possibility is to train on moderate- or relatively low-intensity signals (visual or auditory) so that infants can be excluded on the basis of low "readability" (or inscrutability) in the training phase rather than poor performance in the test phase.

The present technique incorporates 16 signal and 16 no-signal trials. We computed $d'$ values for every possible outcome and identified the outcomes producing $d'$ greater than .95. One such outcome, for example, is 5 false alarms out of 16 no-signal trials and 11 hits out of 16 signal trials for a $d'$ of .98. The probability of obtaining 11 hits and 5 false alarms by chance (i.e., without hearing) is

$$p = \frac{\binom{16}{5} p^5 (1-p)^{11}}{\binom{16}{11} p^{11} (1-p)^5},$$

where $p$ is the binomial probability of a response on a trial. Note that this equation is simply the probability of observing 5 responses on 16 Bernoulli trials (5 false alarms) times the probability of observing 11 responses on 16 Bernoulli trials (11 hits). The sum of the probabilities associated with all of the outcomes yielding a $d'$ greater than .95, therefore, yields the probability of a $d'$ greater than .95, given a deaf subject whose binomial response probability on each trial is $p$. When we performed these calculations for different values of $p$, the maximum probability of obtaining a $d'$ greater than .95 was observed when $p = 15$ or .85. This maximum value was .051. The probabilities obtained for other values of $p$ typically averaged about .02 to .03. Thus, of 100 infants who could not hear, we would expect to pass only 2 or 3 with this criterion.

References


Fulton, A., Manning, K., & Dobson, V. (1978). A behavioral method of efficient screening of visual acuity in young infants. II. Clinical ap-
OBSERVATIONAL MEASURES


(Appendix follows on next page)
Appendix

Historically, signal detection theory developed in response to apparent inconsistencies in adult threshold data such as variability in a subject’s false-alarm rate across sessions (Green & Swets, 1974). The underlying assumption of signal detection theory is that decisions regarding the presence or absence of an external stimulus (i.e., a signal) are based on the value of an internal variable, which is presumed to increase monotonically with signal intensity. The internal variable is thought to be subject to random variability from internal sources (i.e., noise within the organism) as well as external sources (i.e., environmental noise), and the degree of variability is assumed to be independent of signal intensity. Thus, the distribution of the internal variable is presumed to have the same variance whether a signal is present (the signal-plus-noise distribution) or absent (the noise distribution). Figure A1a depicts the distribution of events associated with no-signal trials (the noise distribution) along with the distribution of events associated with signal trials (signal-plus-noise distribution). Note that variation around the mean of both distributions is assumed to be normal and that the two distributions are represented as having equal variance. The distance between the means of the two distributions, expressed in standard deviation units, is a measure of the sensory magnitude of the signal and is called \( d' \). Obviously, the greater the value of \( d' \), the greater the discriminability between the noise and signal-plus-noise distributions. Because listeners must make decisions based on the magnitude of this internal variable, the axis shown in Figure A1 is often designated the decision axis.

The model is based on the notion that, on signal trials, observations are drawn from the signal-plus-noise distribution, whereas, on no-signal trials, observations are drawn from the noise distribution. On each trial, the task of the listener is to decide which distribution generated the observation. Listeners accomplish this by selecting a criterion magnitude (labeled CR for criterion in Figure A1a), which divides the axis into two decision regions. Observations whose magnitudes are greater than the criterion magnitude (to the right of the criterion shown in Figure A1a) are identified as arising from the signal-plus-noise distribution, whereas those with magnitudes less than the criterion magnitude are identified as arising from the noise distribution. According to the theory, the location of the criterion along the decision axis is under the control of the observer and can be manipulated by changing instructional variables, the probability of a signal, and the payoff matrix; the effects of these manipulations may vary from listener to listener. For adult listeners, moreover, it is assumed that the location of the criterion is unlikely to change so long as there are no changes in any of the controlling variables.

The criterion in Figure A1 divides the distributions into four areas corresponding to the probabilities of a hit (i.e., yes on a signal trial), a miss (i.e., no on a signal trial), a false alarm (i.e., yes on a no-signal trial), and a correct rejection (i.e., no on a no-signal trial). It is clear from Figure A1 that in order to estimate \( d' \), one must know not only the probability of a hit (which specifies the location of the criterion with respect to the mean of the signal-plus-noise distribution) but also the probability of a false alarm (which determines the location of the criterion with respect to the mean of the noise distribution). For a given hit rate, the sensitivity or degree of separation between the two distributions can vary widely. Figure A1b presents a case in which the hit rate is the same as in Figure A1a but in which the separation between the distributions is considerably greater. Conversely, for any given sensitivity, the hit rate can vary between 0 and 1. Figure A1c, for example, shows a case in which the separation between the two distributions is the same as in Figure A1a but in which the hit rate is considerably different.

In signal detection paradigms, factors affecting the location of the criterion are typically held constant during a session to ensure a stable criterion, thereby permitting a bias-free estimate of sensitivity, namely \( d' \). In an adaptive procedure, however, several different intensity values are presented over the course of a session, with the high-intensity signals occurring at the beginning and the low-intensity signals at the end of the session. At different times during the session, different signal

Figure A1. Probability density functions for no-signal trials (noise distribution, N) and signal trials (signal-plus-noise distribution, SN).

Figure A2. Top panel: Three signal-plus-noise distributions are shown along with the noise distribution (Model 1). Middle panel: Three signal-plus-noise distributions are shown along with the noise distribution (Model 2). Bottom panel: Psychometric functions for Models 1 and 2; proportions of hits as a function of dB SPL for Models 1 and 2.
intensities could lead the subject to adjust the location of the criterion to reflect the degree of separation between noise and signal-plus-noise distributions in effect at that time. In particular, let us assume that the subject locates the criterion midway between noise and signal-plus-noise distributions (Model 1). Figure A2, top panel, shows how the criterion location would change depending on the signal intensity that is in effect at that time. We show later that this model can result in an average false-alarm rate that is under 25% (given equal numbers of trials at each intensity). Alternatively, a subject could set the criterion so that the probability of a false alarm remains constant at 0.25 throughout the entire experiment (Model 2). This would require that the location of the criterion remain fixed throughout the experiment, as shown in Figure A2, middle panel. Table A1 shows how the proportion of hits and false alarms varies as a function of signal intensity in these two models for the case in which the separation between noise and signal-plus-noise distributions in $d'$ units is given by the equation

$$d' = .05 \text{dB} - 1.$$  \hspace{1cm} (A1)

Equation A1 was chosen to represent the relationship between $d'$ and dB SPL because it is a reasonable first approximation to the relationship between $d'$ and dB SPL found in our data. Table A1 shows that the probability of a false alarm increases in Model 1 as intensity decreases, whereas in Model 2 it remains constant at 0.25. Table A1 also shows that the function relating probability of a hit to dB SPL differs in the two models. Finally, Figure A2, bottom panel, which plots the psychometric functions for both models, shows that the psychometric function for Model 1 is much flatter than the one for Model 2. Thus, the type of strategy used by an observer can dramatically affect the form of the psychometric function. Note that in both cases the false-alarm rate, averaged over the entire session (assuming equal numbers of presentations at each stimulus intensity), is virtually identical. This, then, is an example in which two different strategies can produce the same false-alarm rate but dramatically different psychometric functions. If we apply the standard correction for guessing, the theoretical form of the psychometric function becomes

$$p(Y/N) + [1 - p(Y/N)] \left[ \int_{-\infty}^{x} N(\mu, \sigma) \right], \hspace{1cm} (A2)$$

where $p(Y/N)$ is the false-alarm rate, $N(\mu, \sigma)$ is the normal distribution with $M = \mu$ and $SD = \sigma$, and $x$ is the intensity of the stimulus. Note that if $p(Y/N)$ is known, there remain two parameters to fit. The parameter values of the smooth curves fit to the psychometric functions in Figure A2, bottom panel, are $\mu = 30.7$ dB and $\sigma = 35$ dB for Model 1, and $\mu = 40$ dB and $\sigma = 17$ dB for Model 2. Note that the slope of the presumed underlying psychometric function in Model 2 (slope = $1/\sigma$) is less than half of the value of that in Model 1. Note also that the 50% point on the presumed underlying psychometric function ($\mu$) differs by almost 10 dB in the two models. Thus, both the slope of the psychometric function and the estimate of threshold can be affected by the strategy used by the observer.

Received October 30, 1989
Revision received June 15, 1990
Accepted June 18, 1990


The Publications and Communications Board of the American Psychological Association announces the appointment of Robert J. Sternberg, Yale University, as editor of *Psychological Bulletin* for a 6-year term beginning in 1991. Beginning immediately, manuscripts should be directed to

Robert J. Sternberg
Yale University
Department of Psychology
P.O. Box 11A Yale Station
New Haven, Connecticut 06520-7447