Design and Evaluation Issues in Universal Newborn Hearing Screening Programs

Questions de conception et d'évaluation touchant aux programmes de dépistage universel de la surdité chez les nouveau-nés

Martyn L. Hyde, PhD and Krista Riko, MSc
University of Toronto, Toronto, Ontario

Abstract

In this paper, a perspective is offered on the purpose, design, operation, and evaluation of universal newborn hearing screening (UNHS) programs. The goals are to inform and to stimulate discussion. The endorsement of UNHS is broad but by no means universal. It is suggested that while the direct evidence base for UNHS effectiveness is modest at best, the case for screening can be strengthened by a shift of emphasis from speech and language development to the infant's ability to hear. A programmatic view of early hearing detection and intervention programs is offered that emphasizes integration of screening, audiologic assessment, diagnosis, intervention, and referral. A strong consensus on goals, clearly defined objectives, a powerful information system, a family-centred style of care, and strong, ongoing program evaluation are appropriate in a high-quality program. Also important are attention to the definition and prevalence of the target disorder, population coverage, and operating characteristics of specific screening test strategies. Audiologic assessment is based mainly on the auditory brainstem response to tonepip stimuli, but an approach that integrates all audiometric information is necessary. Early hearing detection and intervention initiatives present remarkable challenges and opportunities for professionals to improve knowledge and skills, revise and rationalize practices, and develop interdisciplinary cooperation.

Abrégé

Dans cet article sont mis en perspective le but, la conception, l'exécution et l'évaluation des programmes de dépistage universel de la surdité chez les nouveau-nés (DUSN). Il s'agit d'informer et de stimuler la discussion. Bien que l'adhésion au DUSN soit importante, elle n'est en aucun cas universelle. Certains suggèrent que s'il n'existe pas de preuve directe suffisante pour établir, dans le meilleur des cas, une efficacité moyenne du DUSN, on peut sans doute renforcer les arguments en faveur d'un tel dépistage en se penchant davantage sur la capacité auditive d'un enfant plutôt que sur l'acquisition du langage et le développement de la parole. L'article présente la manière dont les programmes de dépistage de la surdité et d'intervention précoces sont exécutés et préconise l'intégration des activités liées au dépistage, à l'évaluation audiologique, au diagnostic, ainsi qu'à l'intervention et aux services de spécialistes consultés. Un large consensus concernant les buts visés, des objectifs clairement définis, un système d'information solide, des soins axés sur la famille et une évaluation rigoureuse et régulière du programme sont tous des éléments qui contribuent à instaurer un programme de qualité. Sont également importants la définition donnée au trouble à l'étude ainsi que sa fréquence, la proportion de la population desservie et les caractéristiques opérationnelles des stratégies spécifiques liées aux tests de dépistage. L'évaluation audiologique repose principalement sur une approche qui évalue les potentiels évoqués auditifs avec stimulus tonal, mais une approche englobant toutes les informations audiométriques est nécessaire. Les initiatives de dépistage auditif et d'intervention précoces présentent des défis considérables pour les professionnels et sont l'occasion idéale pour eux d'améliorer leurs connaissances et leurs compétences, de réviser et de rationaliser leurs façons de faire et d'engager une collaboration interdisciplinaire.

Key words: universal newborn hearing screening, early hearing detection and intervention, program evaluation, program design, sensitivity, specificity, screening protocol, diagnostic protocol

A perspective on selected aspects of the purpose, design, operation, and evaluation of Universal Newborn Hearing Screening (UNHS) programs is offered herein. The goal is to inform and stimulate discussion. The viewpoints are based on experiences over the last 20 years in operating a large auditory brainstem response (ABR) based screening and diagnostic assessment program for high-risk infants, participation in various working groups relating to screening and childhood hearing healthcare, and in clinical research on early identification of hearing loss.

UNHS has been endorsed in the United States (National Institutes of Health, 1993) and Europe (Grandori & Lutman, 1999). Several positive position statements from professional bodies have appeared, such as that from the American Academy of Pediatrics (1999). The influential Joint Committee on Infant Hearing (JCIH) has recently published a thoughtful set of guidelines relating to early detection and intervention (Joint Committee on Infant Hearing, 2000). UNHS programs are legislatively mandated in at least 32 American states, and the first stages of a national program are underway in the United

Kingdom. High-level American and Canadian task forces on preventive health care are evaluating the strength of the scientific evidence for universal hearing screening. In Canada, a major model program has been funded in Alberta; the government of Ontario has recently announced a comprehensive provincial program for newborn/infant hearing screening and communication development; and many ad hoc local initiatives in either universal or targeted (high-risk) screening are in place or under development. The federal government has just convened a Working Group on Childhood Hearing to plan a National Workshop in the near future, and to follow up on that workshop.

Despite this activity, we are unaware of any comprehensive, provincial program of early identification and intervention for hearing loss, either universal or targeted, in Canada. At present, there is a patchwork of local initiatives, doubtless with diverse goals, methods, and outcomes. Moreover, the importance of UNHS programs is far from universally accepted.

The Case for Universal Hearing Screening

A reasonable goal for UNHS is to ensure that all infants with significant congenital hearing loss or early-onset hearing loss have an equal opportunity to be identified promptly and to receive appropriate and effective hearing health care and family support services. Clearly, there is ample evidence that the current situation in Canada falls lamentably short of this. The reasons are many and include late detection, delay in diagnostic assessment and in initiation of intervention, and ineffective "intervention." Intervention is defined as any act of service provision intended to change the child's communicative development. It is likely that even today there is not a high level of awareness among healthcare professionals at large about the importance, the feasibility and the benefits of early identification and intervention for hearing loss in infancy. There is an acute need for education and information programs for the public at large, for special interest groups and for a wide range of professionals in the areas of health care, education, and social services.

What do "prompt" identification and intervention mean? For years, the JCIH has recommended identification and audiological assessment by three months of age, and initiation of intervention by six months, wherever possible. Combining various guidelines, primary reports, and other information, what might be called a "1, 2, 3, 4" target plan is suggested: identification of hearing loss by one month of age, confirmation by two months, completed audiological assessment and diagnosis by three months, and nonmedical intervention started by four months, where indicated and elected. We are not convinced of the net merit of fitting hearing aids to children under about four months of age. Medical intervention should be initiated as soon as it is indicated, and any nonmedical component should begin as soon as there is no medical contraindication. For premature newborns, age targets would be adjusted to a 40-week term. For those who have extended stays in Neonatal Intensive Care Units (NICUs), the clock might start ticking at the time of discharge from the NICU or step-down unit. These crude performance targets seem ambitious but we believe they are reasonable and in most cases will conform to the JCIH guidelines.

Why should scarce health care dollars be spent on programs to achieve the stated goal? The usual response from professionals is that undetected and unmanaged hearing loss can compromise speech, language, and cognitive development. While there is substantial evidence to support that statement, there is only limited evidence that early identification makes any long-term difference to speech and language development. In the past, this has resulted in high-level federal health services advisory groups declining to recommend universal hearing screening as part of a preventive public health strategy (US Preventive Services Task Force, 1996). However, the situation has changed recently in two significant respects: First, the technology for quick, simple, noninvasive, and fairly reliable hearing screening in newborns and infants has vastly improved. Second, there are now several substantive studies showing that early intervention is effective in improving speech and language development, and that it is important to intervene by the age of six months (Mayne, Yoshinaga-Itano, Sedey, & Carey, 2000; Yoshinaga-Itano, Coulter, & Thomson, in press; Moeller, 2000; Yoshinaga-Itano, Sedey, Coulter, & Mehl, 1998). Despite the enthusiastic response from clinicians, in the jargon of serious, evidence-based evaluation, such study designs could at best constitute Level II evidence for intervention effectiveness. Level I evidence would require a well-designed and well-executed randomized controlled trial, which is by now questionable ethically and not practicable medico-legally (see Goldbloom, 1997 and Woolf et al., 1990 for a discussion of hierarchy of scientific evidence for and against preventive manoeuvres). There are several methodological reservations about the studies to date, especially with respect to the representativeness of participants and the control of potential confounding variables. In order to achieve at least 'fair (Grade B)' evidence for UNHS programs, there would have to be evidence of a substantive effect from large, well-designed controlled trials without randomization, or from well-designed cohort or case-control analytic studies, preferably from more

than one centre or research group. On balance, unfortunately, it is likely that methodologically stringent review would conclude that, at present, the available scientific evidence for effectiveness of UNHS in relation *specifically to speech and language development* does not meet these standards.

We would like to suggest a different viewpoint. While the putative effect of early intervention on speech and language development is important, it has been overemphasized in arguments for or against UNHS. This is unfortunate, because as the current situation implies, good evidence is hard to come by. Speech and language outcomes are medium to long term, they are inherently complex and multivariate, functional measures are of limited validity and reliability, and speech and language development is influenced by a host of nonlinear, interactive factors that confound a clean demonstration of the effects of early intervention. There is, though, a simpler rationale for early detection and intervention that seems to have been largely overlooked. It starts with accepting that hearing is a normal and fundamental attribute of the human. Notwithstanding special perspectives within the deaf culture, these authors believe that a newborn child has an inherent right to hear, just as the child has a right to breathe. In health services jargon, hearing loss is a "primary health outcome" of the underlying disease or disorder. It is something that is directly experienced by the individual. It is not an "intermediate" outcome, such as elevated cholesterol levels are in the development of overt cardiovascular disease. For the affected child, the hearing loss imposes sensory distortion and sensory deprivation. These authors find the concept of willingly and needlessly imposing up to several years of that condition on a young child very disturbing.

In contrast to the situation relating to speech and language development, there is overwhelming evidence that hearing aids and cochlear implants improve hearing sensitivity. No randomized trial is required to prove that a device with 40 dB gain will render audible a signal that is 20 dB subthreshold. It is a matter of physics and physiology, and is easy to comprehend. So, one has a primary health outcome, hearing loss, and incontrovertible evidence of effective intervention (such as a hearing aid). The sensory deprivation imposed by the hearing loss is ameliorated or negated as soon as the intervention is provided. In the authors' opinion, this is a more solid basis to justify UNHS programs. Any actual advantage that subsequently accrues for speech and language development is icing on the cake.

There is another facet: the rights of the child to equal access to healthcare services. Few adults with a sudden hearing loss would deny themselves access to medical treatment, hearing aids or other supports. So why should intervention be denied to the newborn? Is this some novel form of discrimination? To this, the health services expert might point out that the affected individual is not symptomatic, and the burden of proof of benefit from screening and early intervention is stronger when the affected individual has not sought care (see Feightner, 1992). The response is that one has answered the effectiveness question, and that infants have neither the experience nor the means to complain at their lack of something (i.e., hearing) that they have never experienced and understood. Therefore, the onus is increased upon society to restore the same access to care for the infant as would exist for a symptomatic adult with similarly impaired hearing. *To fail to act proactively to determine infant hearing status is to deny, by inaction, the basic right to hear, as well as equity of access to care.*

At this point, the health services expert might point out that there are downsides to screening, not just the direct costs, but the impact of false-positive screening tests on the parent-child bond. These are important considerations that must be addressed and are discussed later, but they are not solid counter-arguments. To the primary clinician, it may seem that in justifying UNHS programs to health services authorities, one is forced to address bizarre questions such as whether it matters or not to have functioning primary sense organs during an explosive period of learning and neuro-developmental plasticity! Perhaps the very desirable trend towards evidence-based healthcare practices could lead to a flawed position by a series of individually reasonable steps. Every algorithm has its limits. It is speculated that the quality of evidence issue might be less compelling if one were talking about a more "impact-provoking" sense, such as vision. Imagine the conversation: "Yes, there is a small chance that your child cannot see, and yes, we have a cheap, painless and reliable screening test for that, and yes, we can restore at least some vision easily if there is a problem, but you see, we cannot screen everyone because there is no strong evidence that earlier seeing has any important effect on infant development...."

This line of argument must acknowledge that the formal evaluation of scientific evidence is limited in focus and is subject to strict and uniform methodological standards that must be applied evenhandedly across many areas of healthcare. Societal values and broad, contextual elements (such as the current quality of care) do not impinge on such evaluations. However, the results of formal evaluations must themselves be assessed within that broader context.

Early Identification Alternatives

What program design is likely to be the most cost-effective way of achieving the defined goal? A systematic program of targeted screening and early intervention in high-risk infants is relatively easy to justify and would probably be endorsed even by the health services evaluation community. The prevalence of hearing loss is at least ten times higher in the high-risk group. However, it is commonly believed that even the most careful targeted screening will identify at best 50% of children who have congenital hearing loss, because the other 50% have no risk factor by current assessment methods. The recent discovery of the genetic mutational basis for certain nonsyndromic hearing losses (Steel, 1998) raises the possibility of "gene-chip" screening, but that is a complex social issue. At present, the 50% level is only achievable with diligent and exhaustive risk-assessment.

Given targeted high-risk screening, is it necessary to screen the 90+ % of babies who have no measurable risk factor, to identify the roughly 50% of children with presumed congenital hearing loss who are not high-risk, or is there some other way of identifying affected children not at risk? One approach might be a vigorous and aggressive education program to promote earlier identification by parents and professionals. If parents were alert to early communicative milestones and physicians were more inclined to act on parental report, to refer for competent audiologic assessment, or to administer themselves a reliable screening test, perhaps a high proportion of nil-risk affected children could be identified without mass screening. To the authors' knowledge, this question has not been studied properly, but it would be difficult to study in a controlled manner. However, it is hard to conceive of being able by these means to identify children with impaired hearing reliably enough, early enough and across all population segments, especially those who are disadvantaged socio-economically.

What Do We Want To Detect?

UNHS programs are directed at detecting a specific, defined hearing impairment or *target disorder* in health services jargon. An individual with the target disorder is called a *case*. It is essential to define the target disorder precisely, because the definition will affect many aspects of the program design and performance. The most obvious variables that contribute to the target disorder definition are hearing loss type, laterality, severity, and frequency profile. The factors that govern the inclusion criteria are the effect of the definition on the absolute number of cases, the ability to differentiate cases from non-cases, the effectiveness of intervention for the cases at the low-disorder limit of the definition, and the so-called *marginal yield* and *marginal benefit*. This use of the word marginal here is technical jargon, and refers to the difference between the status quo and the proposed program. For example, the marginal yield of a UNHS is the difference in the number of cases identified with the UNHS and some other alternative such as a targeted program or the status quo.

There is good evidence that, on average, the delay in detection and intervention increases as the hearing loss decreases (Fortnum & Davis, 1997; Harrison & Roush, 1996) so in terms of time alone, the marginal benefit increases with decreasing hearing loss. Acting in opposition to this is the likelihood that as hearing loss severity increases, so does its absolute impact. Even a small time gain at a high level of impact may be important. For a child with bilateral profound loss, in the past the suspicion would have evolved at best over a period of several months. Further, it would have been difficult to be sure about hearing levels until at least six months of age, more in the presence of multiple disorders or developmental delay. It is not difficult to do much better than that with modern UNHS.

As the lower severity limit decreases, the individual impact of the hearing loss probably decreases, but even mild hearing losses can cause disadvantage (Bess, Dodd-Murphy, & Parker, 1998). At low severity limits the number of cases probably increases rapidly, and the marginal gain in intervention time relative to the status quo increases. These arguments suggest that the hearing loss severity criterion should be as low as possible, consistent with the existence of a significant impact of disorder and an effective intervention. At some low severity criterion, the prevalence will increase massively upon entering the region of "normal" threshold variation. There is little evidence for a significant disadvantage from slight hearing losses, and even less for effective intervention. The current range of minimum hearing loss criteria is 20 to 40 dB HL.

The frequency range of the target disorder is based on the significance of hearing loss at a given frequency in relation to its impact on the affected child, its prevalence, and its importance in intervention decisions, diagnosis and prognosis. An emphasis on speech development might lead to a focus on the 1 to 2 kHz region. Current knowledge about the population distribution of hearing loss profiles over frequency and patterns of progression in the first months of life is very limited. It is inferred from studies in older children that most congenital, sensorineural hearing loss will express at 1 to 2 kHz, and that the proportion expressing exclusively at low

frequencies is very small. Similarly, it is assumed that hearing loss tends to progress from high to low frequencies. This de-emphasizes low-frequency losses, which is of interest because both otoacoustic emission (OAE) and ABR screening have validity problems at low frequencies and particularly more so for the former.

When defining the limits of the target disorder, it is important to focus on the impact of the disorder and not be overly influenced by current opinion about the capabilities of screening tests. For example, the finding that OAEs are not good screening tools for 500 Hz hearing losses has nothing to do with whether it is important to detect 500 Hz hearing losses, which is a deeper question. There is a temptation to define the target disorder in such a way that one's favourite screening method will turn out to have really good performance. The next step is to forget that the undetected cases even exist.

There is contention about whether unilateral hearing loss merits detection. Some argue that the target disorder should be bilateral, which simplifies matters and might be justifiable on cost grounds. The overall systems development and infrastructure costs are similar for unilateral and bilateral criteria, but some operating costs would increase for unilateral criteria, other things being equal. The use of a unilateral criterion at least doubles the referrals for confirmatory and diagnostic assessment, for a given severity criterion. In our view, there are good reasons to include unilateral, permanent hearing losses. To assume a child with one good ear is not disadvantaged is misguided. A hearing aid or other intervention may be helpful. It is necessary to quantify the hearing loss, diagnose it and develop a prognosis. It is necessary to monitor, to assess progression. Parents need to be encouraged to seek medical and audiologic input when there is any ear-related complaint or behavioural change suggesting a hearing problem. Transient, conductive loss in the contralateral ear can render a child with unilateral sensorineural hearing loss unable to hear. A more diligent and aggressive management approach to common middle ear disease is needed in the child with preexisting unilateral sensorineural hearing loss.

The Programmatic View

A discussion of program design is now in order. The first point is that UNHS programs are poorly named. The purpose of the program is not to screen, but to deliver timely and effective intervention options to those in need. There is more than cosmetic significance to the fact that the JCIH refers to an early hearing detection and intervention (EHDI) system. In this much better name, which will be used throughout the rest of this article, the detection of hearing, not hearing loss, is emphasized. This gives a positive character and reduces the tendency towards labeling of individuals who do not pass screening tests. The "intervention" is made explicit, because it is intervention that is the whole point of the exercise. This is not a trivial matter. There is more to a name than meets the eye. Terminology has a habit of affecting thinking in subtle ways. Inadvertent focus on screening itself can lead to insufficient attention being paid to other critical elements of the overall program, throughout program planning, operation and quality management. For example, existing facilities for audiometric confirmation, diagnosis and intervention almost certainly will be inadequate to deal appropriately with the output of newly-implemented universal screening. There is not only a bulge in the demand for follow-up services after initiating screening, but henceforth those services will be dealing with much younger children, necessitating changes in practices. A substantial upgrading and rationalization of diagnostic and intervention services is likely to be required, and those services must be tightly integrated with the screening. Thus, screening must be seen as a necessary but not sufficient component of an overall system, and a balanced approach that integrates and links all the crucial steps into a coherent and seamless whole is necessary for success. Because the process is a serial one, with tight linkages and conditional branching, the entire enterprise is only as good as its weakest link.

An EHDI program as part of a hearing health care system for young children will now be considered. The overall EHDI program will have various components, including a universal screening component. What is the glue that makes this more than a disconnected set of activities? What is it that makes a *program*, as opposed to just an activity? There are several key elements. A program has a defined goal and specific, quantitative, evaluable objectives. A program has a *structure*, a set of infrastructure elements (e.g., staff, clients, buildings, equipment, supplies and funding) that, by delivering defined *processes* (e.g., test protocols and tracking procedures) facilitate attainment of desired *outcomes*. Most large programs benefit from modular design, with defined components; this approach facilitates definition of subprograms, specification of subprogram objectives, and formulation of decision rules and processes that link subprograms. Also, a good program has the evaluation of its own structure, processes and outcomes as an integral part of its design. Finally, the most successful programs usually have a style, a character, and a pervading theme. Three important aspects of the EHDI program "glue" will be discussed briefly.

Commonality of Goals and Objectives

EHDI programs have several features in common with other complex programs. Ultimately, the keys to success lie in the conduct and attitudes of program personnel. The challenge of program design is to facilitate and encourage high-quality performance. There are several important ingredients. For the greatest success, every program activity must be seen to serve a common goal, so that there is a minimization of disjoint or even conflicting activity. Misunderstandings and poor communication abound in complex human systems. It follows that efforts must be made to ensure that everyone involved in the program is on the same team and reading from the same page. This is often very difficult to achieve. It is helpful to make a strong effort to continuously inform program staff about the program goals and achievements; it is remarkable how isolation, lack of information and "seeing the elephant through a keyhole" can create alienation and inappropriate action. It is also very important that each and every member of the program staff feels a sense of involvement and personal responsibility; genuine consensus-building processes in program development and ongoing operational evaluation can encourage this.

It is helpful to define and disseminate very specific program objectives and to implement high-profile processes for measurement of program and subprogram performance. Most people like to feel that they are doing a good job, and they welcome proof of it and the means to improve. Most healthcare professionals will put the welfare of their clients as the top priority. In both aspects, it helps to have quantitative measures of performance. Furthermore, performance measures are essential for external accountability. This relates back to the careful definition of program objectives. What is not carefully defined cannot be evaluated. Retrospective redefinition of fuzzy program objectives in order to wring apparent success out of failure is an art that is alive and well, but it does not ultimately serve the care recipient.

Information Systems

A comprehensive, integrated, well-planned information system (IS) is perhaps the most important nonhuman ingredient in a successful EHDI program. Core IS functions include storage of subject identification and description, parent/caregiver compliance, test schedules, test occurrence information (e.g., time place, tester, etc) and test results, similar information for re-screens and confirmatory tests as necessary, logging and timing of referral generation, occurrence, report reception and outcome, storage of risk data to identify children at risk for progressive hearing loss and in need of periodic re-screening, intervention logging and outcome data, quality management data including parental satisfaction measures, period activity reports and program evaluation reports, and so on.

The IS defines whether a child exists in the program, what should happen to the child and when, whether it did happen, who did what and where, and what the result was. In a very real sense, if the data are not in the IS, then there *are* no data. Thus, a screening test may have been done, but if its occurrence and outcome are not entered, then effectively it did not occur. Thus, a test should be considered as done if and only if its occurrence and outcome are entered into the IS. The data entry is what brings the work done to fruition. The information is the life blood of the program.

The IS can have a very strong effect on program quality. It may include all manners of data controls that aid in the collection and storage of necessary data in a reliable manner. Database fields can be made mandatory if appropriate, specific data values can be required or rejected, data-conditional error messages and commands can be generated, options to change data values can be restricted, cross-checking can be done, default values chosen, and so on. The system should also contain sophisticated process checks, measures to detect and correct errors associated with input mistakes. This would include not just the wrong value of a measurement, for example, but more profound errors such as an attempt to create an inappropriate appointment type or date or to discharge a client without having ensured that certain supports have been established. Administrative override ability to break database end-user rules will be essential for selected, critical data items.

In a large-scale program, IS development effort can be substantial. Commercial software is widely available and can seem expensive, but it is easy to underestimate the costs of developing a custom system, and difficult to develop a reliable and comprehensive system. System flexibility and tailoring to local needs can be a very big issue. In some cases, such as very large-scale programs for an entire province, compatibility with existing information systems may be a critical factor in the selection or the design of the system.

Weak ISs tend to be generic, but may contain well-developed methods for handling key, standard items of information. The challenge for software developers is to balance system costs, integrity, and context-specificity. The most powerful ISs are highly context-specific and are absolutely intimately linked to microscopic details of program operations. The array of logical decision rules and the degree of process control that may be built into a sophisticated IS can be formidable. In-

deed, the functional specification of the IS can be very useful and influential in program design. The need to specify precise, comprehensive and pedantic rules for data management helps to clarify many details of test protocols and other important program processes.

It is prudent to plan the IS to take account of unforeseen situations and overlooked items of information. For example, if up to six ABR thresholds might be measured per ear at the confirmatory assessment, it would be wise to allow for ten. Database volume is rarely an important consideration these days, but in contrast, the logical linkages among data items and the underlying database table structure are very important in facilitating value-specific actions and other database queries such as for report generation. It is often much easier to ignore a system option than to add a new capability as an afterthought once the IS has been implemented. In addition, it is appropriate to build in a capability of epidemiological or audiological investigations that might not seem important in the urgency to implement a system, and which go beyond ordinary program evaluation. This is usually a matter of careful thought about what the areas of both local interest and more generally limited knowledge are, and incorporating extra fields so that unanticipated variables and various methods of data exploration and statistical analyses can be accommodated readily.

Quite apart from the core content of the various data entry screens that will be presented at various steps in the program, it is important to give attention to the design of the graphical user interface. While it seems a mundane area, poor interface design can frustrate users, decrease data entry efficiency, and increase error rates or missing data rates despite administrative data controls. It is really important that the users like the look and feel of the database interface. It is helpful if the data entry screens are as simple and uncluttered as possible and if they match the format of any hardcopy forms used in program administration or in audiological data collection.

Program Style or Theme

There is a clear distinction between what must be done and the way in which it is done. While it may seem a hackneyed refrain, the importance of a child-centred and family-centred program style cannot be overestimated. There are several strong reasons for this. The family or caregiver is actually the agent who decides whether an appointment, data collection or test procedure will occur or not. They are crucial partners in the daily process of maintaining and promoting effective intervention. They can be important contributors of additional information about hearing-related and intervention-related behaviours. If the family is not onside and involved throughout the EHDI process, the value of every program step is in question. Furthermore, the family and the child form a unit, and the child's hearing loss clearly may have a profound effect on the entire group. Therefore, the unit needing intervention options is the family, not just the affected child.

Parents appear to want perfectly reasonable things from EHDI programs, but they may not get them (Bamford, Davis, Hind, McCracken, & Reeve, 2000). At the initial stages of screening failure, they need sensitivity, emotional support, and good information about what the result means or does not mean, what happens next, and what their choices are. At confirmation and diagnosis the same ingredients are required, but in more depth and in different balance. The intervention options can be a problematic area, of course, and the advice must be culturally sensitive as well as consistent with professional integrity. The parent may elect an avenue that is not endorsed by the audiologist. The right to do that underlies the concept of family-centred care. Information and persuasion are legitimate means of influence, whereas the authoritarian imposition of the professional's preferred management approach is not. It is important to remember that in EHDI programs intervention will normally be offered well before six months of age, far earlier than in the past. Usually, there will be some time to continue discussion, observe actual effectiveness, and change course if necessary.

In the authors' view, the best programs give parents an array of choices in an unbiased and evidence-based manner. There is no place for doctrinaire dogma in a high-quality program. This means that the various intervention options should be presented in an even-handed and informational manner, with the commonly quoted strengths and weaknesses of each avenue identified as opinions if there is not scientific evidence that can be referenced. Written materials are virtually essential. The audiologist informs and may recommend, as is the case for the physician. The parent or caregiver decides. Some parents do not want too much information and may wish to be guided. That is a professional judgement call.

Program Evaluation

A good program has a strong quality evaluation component built in. Comprehensive program evaluation is multifaceted and includes systematic review of program structure, processes, and outcomes. It includes assessment of the clarity and appropriateness of the program goals and objectives, the performance of each and every distinct program step, including test protocol clarity and personnel compliance, test timeli-

ness and outcomes, referral performance, data management, program style consistency, staff satisfaction, client satisfaction, infrastructure aspects such as staffing, training, equipment calibration, field performance, maintenance and repair, and many other facets. Some of this falls within the purview of routine, internal program management, but it is good practice to schedule periodic external, more systematic appraisals.

Program evaluation, like any other activity, can be good or bad. A good evaluation subprogram is seen as a positive influence, not threatening or punitive. For those being evaluated it should be seen as based on a deep understanding of the purpose of the program and of the realities of front-line delivery. Evaluations conducted in a rote fashion, or by consultants who blatantly do not understand the core business, or who appear to have a covert agenda (e.g., cost-cutting), are often destructive. For a good evaluation, it is necessary to get people to open up and tell it like it is, not to clam up or cover their backs.

Cosmetic evaluations tend to miss the point, find irrelevant flaws and lead to either no changes, ineffective changes, or actually harmful changes. Good evaluations zero in on real issues and engage program personnel *and program recipients* in finding consensus solutions that work. They do not stop at diagnosing faults, but propose concrete and practicable solutions wherever possible. The measure of an evaluation program is not how painstaking it seems or even how many difficulties it unearths, but whether its conclusions are valid and lead to actual and sustainable improvements in program output or efficiency. Evaluation without corrective action and subsequent determination of the effects of that action might as well not be undertaken. To this end, the evaluation subprogram itself must have specific objectives that are evaluated.

Process and Outcome Evaluation

Usually, a program evaluation is directed at the assessment of program *effectiveness*. In the jargon of health services research, effectiveness is a measure of whether an intervention *does* actually work, in contrast to efficacy, which is a measure of whether the intervention *can* work, at least in principle. Note that an efficacious intervention may turn out not to be effective. Coverage and follow-up compliance problems would be obvious causes of this: it does not matter how good a treatment is, for example, if you do not ever get to receive it.

There are at least two flavours of program assessment: administrative and evaluative. The administrative view is more immediate and short-term. A typical thrust would be to ask if the overall quarterly rate of screening is satisfactory. This type of routine reporting is typically based on activity within fiscal periods, and there is a presumption that the quantity of activity is an index of value. What is being counted here is simply the occurrence of specific events. Contrast this with a truly evaluative question, such as "Is the program achieving its stated objectives?" This question is usually harder to answer. Ultimately, the program exists to deliver successful interventions, that is, to have a sustained, positive impact on the affected child and family. The results of the interventions may take a long time to develop. For example, a long-term outcome might be that a child with a substantial hearing loss achieves near-normal speech perception performance and speech and language developmental milestones at age three years. Long-term outcomes can be difficult to relate directly to program events, not only because they take some time to happen but also because they may be influenced by a host of factors beyond program control, such as a family breakdown or a serious, unrelated childhood illness. If available, short-term outcomes can be immediately useful in program quality improvement. An example of a short-term outcome is a measure of family satisfaction with the program.

Because many outcomes tend to be delayed and confounded by nonprogram factors, it is common to use measures of program processes as substitutes or proxies for true outcome. Examples include screening coverage, the proportion of diagnostic assessments that are successful and timely, prompt enrollment in a family-centred intervention plan, and so on. These are not true outcomes, but in so far as they are genuine mediators of the ultimate endpoint, they are useful in program evaluation. Their actual value, of course, depends on the validity and strength of their relationship to ultimate outcomes. These types of measures are sometimes called *process outcomes*, as distinct from *outcome outcomes*. It is worth noting that because the proper fitting of a hearing aid on a young infant has a very strong and immediate relationship to the improvement of hearing sensitivity, then if that improvement is accepted as a genuine primary health outcome (as was proposed earlier), the process event of fitting the aid *properly* is virtually synonymous with achievement of that primary health outcome. The relationship to subsequent speech and language development milestones is, of course, much less direct, and so fitting the aid is only a (statistical) process proxy as far as speech and language outcomes are concerned.

Process outcomes are quantified by *indicators*, which are usually quantitative variables that can be derived from program activity records. A simple example of an indicator of

screening performance is the proportion of live births successfully screened by one month corrected age. Periodic computation of this type of indicator is an essential part of ongoing program evaluation, above and beyond the administrative performance analysis. An evaluation indicator should be defined for all major program steps that are crucial to the achievement of the program goals and objectives. Target values for performance indicators are called *benchmarks*. These are defined in the program objectives. The benchmarks must be specific and quantitative, otherwise their achievement cannot be verified. That is why, unlike goal statements, which are often nonspecific and thematic, program objectives must be absolutely crystal clear and precise. For useful examples of indicators and benchmarks for EHDI programs, see the JCIH guidelines (Joint Committee on Infant Hearing, 2000).

Finally, another very important component of program evaluation is assessment of protocols and the protocol compliance among program personnel. A high-quality program will be highly structured and all the major steps will be governed by documented protocols. It is very difficult to evaluate and improve activities that are not well-defined, so protocols are an essential part of a program's evaluation. An evidence-driven, consensus-style process should have derived the protocols. If there are areas of practice where the evidence is inconclusive, then options and flexibility must be built into the protocol. While it may seem at first sight that explicit protocols infringe upon concepts such as "clinical judgement", that is a fallacy if the protocol is developed properly. A good protocol enshrines and crystallizes what is agreeable as a desirable, effective course of action. If the protocol is deficient, it must be revised as part of the evaluation process, in close consultation with those who actually are required to deliver it. A common reason for noncompliance with a protocol is the belief that it was developed by people who did not really understand the front-line realities. Unfortunately, that may often be the case.

Program Context Effects

EHDI programs do not live in isolation. They are a component of the health care fabric of a community. Efficient and effective EHDI programs will reflect the values of the culture in which they are embedded, link effectively with allied components of child health care, early educational systems and family social supports, and capitalize on existing mechanisms and infrastructure elements. This means that EHDI program structures and processes may differ radically from one culture to another or from one geographic region to another. For large programs at the state or provincial level, heterogeneities of culture, demographics, socioeconomics and health care delivery systems are such that there may be substantial differences in the way things are done in different parts of the program. It is desirable on equity grounds that even if the mechanisms differ from place to place, the objectives remain as constant as possible. Pragmatically, with finite funding it may be necessary to adjust objectives to reflect what is feasible given local conditions. In a sense, a large program may be viewed as a set of linked, parallel subprograms, each with individual characteristics. The planning challenge then is to balance consistency and context-dependency, under resource constraint.

Part of the program context is the status quo with respect to availability of personnel with appropriate skills. It is to be expected that the introduction of a large EHDI program will have many implications for change in professional practices. One example is that prior to the program, there will be a certain amount of effort being expended to deal with children who are identified late. A desired goal of the EHDI program is clearly to minimize the need for such activities, and it is important to redirect the expertise and resources to maximum benefit within the new system. Another example is the area of audiological expertise with diagnostic ABR testing. With EHDI, there will be a need for more and better ABR testing, and a need to adapt to an evidence-based, protocol-driven practice pattern with a strong evaluation component. There are also clear implications for training and continuing education programs, and perhaps a case for subspecialty certification.

Screening Test Protocols

Having outlined the importance of the overall, programmatic viewpoint, a discussion of a few specific areas and issues arising in them is in order. The emphasis is on screening, with a few comments on audiologic assessment. The performance achieved by a given screening process is dependent ultimately on four fundamental parameters: prevalence, coverage, sensitivity, and specificity.

Prevalence

The prevalence of the target disorder is the proportion of a defined *target population* for screening who actually have the disorder at the intended time of screening. The target population is reasonably defined as the set of all live births within a defined time period (i.e., a *birth cohort*). The true prevalence of congenital hearing loss and the pattern of evolution of hearing loss in the first year of life are not known exactly, and depend on many factors. Only high-quality, large-sample, lon-

gitudinal studies, based on electrophysiological measures such as the ABR, can provide really good descriptions of these phenomena. What are currently available are somewhat speculative estimates of congenital and early-onset hearing loss prevalence and progression patterns. These are based on back-projections from findings in older children (e.g., at two years or more), as well as on small-sample case studies, coupled with very new information from recently-established universal screening programs.

While there is much variation in published figures, typical quoted prevalence values are about 1/1000 for bilateral, at least severe, congenital, sensorineural hearing loss. This probably increases to at least 2/1000 if unilateral losses are included and to about 5/1000 if all sensorineural losses of at least mild degree are included (Stein, 1999). Clearly, the definition of the target disorder is a major determinant of its prevalence. The first question is whether the disorder is defined as bilateral only, or may include unilateral loss. The authors' data from a large sample of high-risk ABR screening suggest that prevalence approximately doubles if unilateral loss is included. Inclusion of conductive loss increases prevalence, by an amount that depends on test timing and the target severity criterion. The larger the loss criterion, the smaller the prevalence increment.

For a given target disorder, real prevalence changes across populations are to be expected. The causes include population demographics, the quality and style of perinatal intensive care, geographic factors such as access to acute care facilities, socioeconomic factors such as poverty levels and nutrition during pregnancy, drug and alcohol use, and other factors related to culture and ethnicity, especially intermarriage and care-seeking behaviours. All these variables relate either to the risk or predisposing factors for hearing loss in the population, or to the effectiveness with which expressed risk is managed.

Published prevalence estimates would vary substantially due to sampling variation, even if there were no true differences. When statistically estimating the true value of a small proportion, such as the prevalence of congenital hearing loss, a very large sample size is needed for a precise estimate. For example, suppose a random sample of 10,000 newborns screened were to yield 20 cases of significant hearing loss. The best estimate (i.e., point estimate) of the prevalence would be $p = 20/10{,}000 = 0.002$ or 0.2%. The 95% confidence interval width is approximately root $(16p/n)$ = root (16 x 0.002/10,000), which equals 0.0018 or 0.18%, almost as large as the prevalence estimate itself! In practice, a sample size sufficient to yield at least 50 and preferably 100 cases is necessary to achieve reasonably precise estimates of prevalence, for true prevalences less than about 5%.

It is important to consider variations in prevalence when designing a program, modeling the flow of cases through any proposed system, and developing economic projections of the direct and indirect program costs and benefits. When looking at published data, the similarities between the test populations and care practices reported and the local situation must be reviewed carefully. In particular, multivariate *sensitivity analysis* should be employed in prediction of program performance. Here, the term sensitivity is different from that used in conjunction with specificity (see later). Sensitivity analysis involves determining how stable the quantitative predictions of a model are when the values of the parameters fed into the model are changed, usually by modest amounts (i.e., perturbations). When there is more than one parameter, as is the case in modeling EHDI programs, there may be interactions among parameter perturbations. Also, it may be found that overall performance depends much more strongly on one parameter than on another. Such findings can influence strongly the program rationale, design and evaluation procedures.

Coverage

Coverage is the proportion of the target population that actually receives a successful screening test in the desired time frame. It is a function not of the test itself but of the actual delivery of the test in an effective and timely manner. High coverage is a very important indicator of program quality. The keys are diligence in identification and tracking, gaining parental compliance for testing, scheduling and delivering the test or tests. There is a law of diminishing returns, and it may take a lot of effort and ingenuity to push coverage beyond 75%. However, the benefit of universal screening over at-risk screening depends critically on doing better than about 50% (the at-risk proportion), so 75% is not at all good, representing access to at best only half the possible additional cases. A common reason for limited predischarge coverage is unpredictable discharge or transfer of newborns from the screening hospital.

Coverage directly affects the overall, net operational sensitivity of the screening system. A screening test may have an excellent intrinsic ability to detect the target disorder (i.e., sensitivity), but if the test is not done on some babies then all true cases who are untested are missed and are effectively false negatives. All the probabilities for delivery of screening and intervention concatenate and multiply. For example, a system that achieves 85% coverage using a screening protocol with 95% sensitivity has *effective* screening sensitivity of only

81% (i.e., .85 x .95 = .81). Furthermore, if the numbers just given apply and the success rate for follow-up of screening failures were 80%, then the actual net effective probability of intervention delivery in the target population falls to only about 65%! Thus, despite excellent screening test quality and coverage and follow-up that looks at least respectable, the point is rapidly reached where only two out of three babies in need get to intervention. That is the nature of a process with a series of steps, each conditional upon the last.

In program design, it is easy to focus on test performance and lose sight of the seemingly mundane, nontechnical, nonclinical factors such as access optimization processes, parental compliance promotion methods, tracking and scheduling systems, follow-up promotion approaches, and so on. All play an important role in the overall process of delivering intervention to all those in need. To optimize system performance, efforts must be directed to whichever program component is rate-limiting the overall performance. If and when the administrative processes are tuned to excellent performance, then the intrinsic operating characteristics (sensitivity and specificity) of the screening protocol itself become the performance-limiting factors.

Sensitivity and Specificity

Sensitivity and specificity are intrinsic properties of the screening protocol itself. The nature of sensitivity and specificity, and their relationships, have been described in several audiology-oriented texts (Hyde, Davidson, & Alberti, 1991; Jacobson & Jacobson, 1987). Sensitivity is the probability that an individual with the target disorder will fail the screen (i.e., positive screen). Specificity is the probability that an individual without the target disorder will pass the screen (i.e., negative screen). In a subject with the target disorder, the screen must be either positive or negative, so the sensitivity (i.e., true positive rate [TPR]) and the false negative rate (FNR) must sum to 1.0. When the patient does not have the target disorder, the specificity (i.e., true negative rate [TNR]) and the false-positive rate (FPR) must also sum to unity.

The usual model underlying sensitivity and specificity involves a statistical distribution of a test measure such as a signal-to-noise ratio for an OAE or an ABR. There are two distributions, one if the target disorder is present and another if it is absent, and they usually overlap. The screening result is governed by a decision *criterion*, which is a specific value of the response measure. Sensitivity and specificity are respectively the area under the disorder-present distribution for outcomes less than the criterion, and under the disorder-absent distribution for values greater than the criterion.

Any value of the criterion will yield some sensitivity and specificity pair. The stricter the statistical response (OAE or ABR) detection criterion, the more likely the test will be judged response-negative (i.e., positive screen). This is true whether or not the target disorder is present, but the probability of a positive screen will be much greater in the presence of hearing loss. Thus, as the detection criterion gets stricter, the sensitivity and the false-positive rate increase, so sensitivity and specificity (1 - FPR) vary inversely. Each is a number between zero and unity; tossing a coin instead of testing would give a sensitivity and specificity of 0.5. The perfect test has sensitivity and specificity equal to unity (zero FNR and FPR), but that implies no overlap in distributions, which is almost never the case. In subjective terms, the more separated the two distributions are, the better the test is.

Because sensitivity and specificity depend on the test criterion, a sensitivity or specificity figure by itself says little about how good the screen is. Both sensitivity and specificity are needed. But even that is not enough, when comparing the performance of two or more screening protocols. It is easy to see which is best if the sensitivity or specificity was held constant, but what if all the values are different? Which test is best? There are two approaches. One is to assign numerical costs to the errors (which may be very difficult) and calculate statistically *expected costs* associated with each sensitivity and specificity pair. A more sophisticated method is to incorporate the effects of all possible criterion values and compare the screens using the *relative operating characteristic* (ROC), which is a plot of sensitivity against FPR for various values of the criterion (see Hyde et al., 1991; Swets, 1988).

The actual sensitivity of a screening test is a very difficult quantity to determine directly because in order to know who truly has the target disorder, everyone who is screened, regardless of the screening outcome, must also be given a gold standard test such as high-quality diagnostic ABR, or visual reinforcement audiometry where feasible (Hyde, Riko, & Malizia, 1990; Norton et al., 2000). When the time interval between the screen and the gold standard is lengthy, less than perfect follow-up and actual changes in hearing status may bias the results. Sensitivities in the range of 85 to 95% are probably achievable with high-quality screening procedures, for hearing losses that are moderate or greater, but lower values are to be expected for lesser degrees of loss.

An obvious cause of false-negative screens is an auditory system disorder that is more rostral than the level assessed by the screening technology. For OAE, for example, disorders located at a neuronal level higher than the outer hair

cell could cause these errors, such as in the situation of auditory neuropathy (Sininger, Hood, Starr, Berlin, & Picton, 1995). For the ABR, it is conceivable that extensive thalamic or cortical lesions could go undetected. The actual prevalences of such conditions are unknown, but probably very small.

Another important contributor to overall screening error rates is the statistical decision process underlying the automated screener (Hyde, Sininger, & Don, 1998). Typically, the response detection criteria are adjusted so that the probability of false detection (of an OAE or ABR) is very low, usually below 1%. With respect to this source of error, the screening test FNR is set very low, so in effect the sensitivity is set very high. For the ABR, a possible cause of false-negative screens is false-positive response detection due to electromyogenic artifact. It is essential not only that the signal processing and statistical decision algorithms in automated screeners are optimized with respect to artifact management, but also that screening personnel are reasonably adept at recognizing and managing situations that promote high electromyogenic artifact. The most common is an overtly moving or fussing baby.

While sensitivity is obviously important, for low-prevalence disorders it is the false positive rate that usually ends up controlling program feasibility and cost-effectiveness. This is the case because it takes so many screens to yield a single true case, and even with a very good specificity, most of the children who screen positive will be false-positive. It is easy to fall into the trap of believing that children who fail an initial screen (e.g., OAE) are likely to have hearing loss, but calculating with typical prevalence, sensitivity and specificity values, the probability that an individual child who fails the screen actually has the target disorder (the positive predictive value of the screen) is unlikely to be greater than 10 %. A child who fails a screen should be considered at increased risk of hearing loss.

It is common to see FPR benchmarks set at 3% (American Academy of Pediatrics, 1999), which equates to a specificity target of 97%. It is not clear why 3% should be chosen, but it may reflect some assumption about the best possible result with OAE. A quantitative rationale for setting an acceptable FPR would be based mainly on the costs of false-positive errors. The costs include possible parental anxiety or distress, re-screens, or diagnostic assessments. Because the FPR and FNR vary inversely, their costs must be weighed. The overall costs of detecting a child with hearing loss may or may not be outweighed by the benefits (negative costs) accruing from early detection. These benefits are difficult to assess monetarily, so at present a quantitative rationale for setting the maximum FPR does not appear to exist. What is acceptable for now may be governed not by fundamental arguments, but pragmatically by what appears to be readily achievable.

The main factors affecting the FPR include resolving middle-ear fluid within the first 24 hours of life. The only way to solve that difficulty is to delay the screen. A more general problem is environmental noise or movement of the baby such that signal detection criteria for detecting the OAE or ABR are not satisfied, but with adverse conditions not bad enough to trigger any automatic warning from the screening device. Also, immaturity or reversible pathology of the auditory brainstem pathways may cause a few automated auditory brainstem response (AABR) screening failures, but resolve in the first few weeks or months of life. The prevalence of these disorders is not known, but is small. They are not true screening errors, but they have the same effect.

It is important to note that automation of screening devices reduces interpretive errors but does not abolish the need to try and achieve the best possible test conditions, which involves simple and practical skills at recognizing and solving common problems. While non-audiologist personnel can use automated screeners, this does not mean that a superficial approach to training and test procedures can be taken. The overall performance of the screening program will depend strongly on the quality of the training programs for screening personnel, and their motivation to improve. This is consistent with the observation that false positive rates for OAE screening, for example, may differ dramatically from place to place and tend to improve substantially with time.

ABR versus OAE

There appear to be no substantial differences between transient evoked OAE and distortion product otoacoustic emission (DPOAE) methods in terms of their sensitivity and specificity for detecting hearing loss at middle to high frequencies, as reflected in large-sample ROC analysis (Norton et al., 2000). The present authors speculate that DPOAE screening may be made more efficient in terms of how long it takes to get an acceptable test result in a given acoustical environment, and more robust in terms of the range of adverse test environments that can be tolerated. This view is based on the relative simplicity and power of statistical detection procedures for DPOAE. However, a disadvantage of OAE screening relative to the ABR is its relative insensitivity and high false-positive rate for hearing loss at 1 kHz and below (Norton et al., 2000). The main cause of this is increased noise at lower frequencies. It is feasible in principle to screen at 500 Hz with the ABR, using tonepip stimuli with or without notch filtered or high-pass ipsilateral noise masking (Stapells, 2000). How-

ever, these authors are unaware of any quantitative evidence on the performance of automated low-frequency ABR screening in an EHDI context. Most ABR screeners utilize click stimuli. These authors are also unaware of any well-validated automated response detection algorithm for tonepip ABRs.

It seems possible that the concept that one screening test or multi-test protocol will be best for all babies is simplistic. The weight of various factors may indicate different procedures for different subgroups. An example of this is a relatively strong case for AABR screening versus automated otoacoustic emission (AOAE) screening, in NICU graduates. In that population, the a priori likelihood of hearing loss is much higher (i.e., by a factor of at least ten) than in the non-NICU population. It is a general principle in testing that the higher the probability of disorder, the more definitive is the test required. An analogy is the switch from otoneurologic ABR screening to a direct path to magnetic resonance imaging when the base probability of an acoustic neuroma in an adult is high. The other factor favouring AABR in NICU graduates is that the likelihood of a retrocochlear disorder is increased; these will be missed by OAE screening, but are likely to fail AABR screening. Disorders of neural synchrony such as auditory neurophathy, are included here.

In general, the arguments favouring OAE screening emphasize speed and simplicity, relative to AABR screening. These arguments are undermined by the now commonplace practice of multiple OAE screening, in an attempt to control the intrinsic high false-positive rate of the single OAE screen. Personnel skill requirements are similar, and improvement in FPR rates over time suggests that there is clearly a skill factor in OAE screening. So, the real distinctive feature of the AABR screen is the need to place scalp electrodes. Bearing in mind that if all initial screens were AABR, there would be reduced need for rescreening and all its overhead costs, one would not be surprised to see a movement towards more general use of AABR as the initial screen in future EHDI programs.

Multi-Test Protocols

It is common to encounter multi-test screening protocols (Gravel et al., 2000). The overall performance of a multi-test protocol depends on the sensitivity and specificity of the component tests, the combination rules that link the tests, and the correlation structure among the test outcomes (Hyde et al., 1991; Turner, 1988). Some examples follow.

Repeated Tests

It is the norm to repeat a given test, such as an AOAE, with the rule that the child fails overall if both tests are failed (i.e., *refer* outcome). If the tests were statistically independent, overall sensitivity would be the square of single-test sensitivity. The overall specificity is (1-fpr^2). Thus, if the single-test sensitivity and specificity were respectively 0.95 and 0.90, say, the overall sensitivity and specificity would be 0.90 and 0.99. By this maneuver, the specificity is increased dramatically and the FNR is doubled. In practice, because of correlation among test outcomes, it would be more realistic to expect at best a halving of the overall FPR, and a slight decrease in sensitivity, with retest after an attempt to remedy the cause of failure.

To understand the actual effects of test repetition, the specific causes of false-positive and false-negative errors must be examined. If the source of the error is constant from test to test, then nothing is gained by repeating the test. An example would be an OAE retest given failure due to environmental noise, without some effective manoeuvre to quieten the child or to change the acoustical environment. If the cause of test failure were an occluded probe, retesting could be beneficial if the problem were poor positioning against the meatal wall, but not if there were canal occlusion due to cerumen. Retesting can only help if there is a reasonable effort and opportunity to remove the probable cause of error.

Combined Protocols

While a reasonable case can be made on cost and simplicity grounds to use OAE as a stage one screen, it can be hard to achieve very low FPRs with OAE. One approach is follow OAE screening failure with a (stage two) re-screen by AABR. This reduces the overall FPR, and with far fewer AABR tests than by doing single-step AABR screening in all babies. The principle is reasonable because the OAE is used to increase the prevalence of the target disorder to the point where the more definitive AABR is warranted. The overall performance of AOAE-AABR two-stage (i.e., "two technology") screening is reported to be much better than that for a series protocol based on the OAE (Gravel et al., 2000).

The time between the stage one and stage two screens is an opportunity for transient conditions such as middle-ear fluid to resolve or emerge. Only children who fail stage one are retested, so an emergent disorder is not detected. For some subjects, stage two will be negative because of disorder resolution, not because the first screen was false-positive. Is there an optimal time between screens? The longer the interval, the longer the parents have to live with a possible false-positive stage one and, other things being equal, the later the initiation of any needed intervention. Conversely, the greater the opportunities for resolution of transient middle-ear disorders and early neurodevelopmental abnormalities, and for expres-

sion of progressive or early-onset disorders. The probability and cost-benefit information needed to weigh these factors quantitatively is not yet available. Clearly, reasonable measures should be taken to allay parental anxiety. Currently, we presume that a two to four week interval between screens may be beneficial and is unlikely to compromise the timing of early intervention, as long as the overall process leading to intervention is reasonably efficient. In some environments, delaying the re-screen might compromise tracking and follow-up compliance, which would be a significant disadvantage.

Screening Parameters and The Target Disorder

It might be thought that the definition of the target disorder would determine many aspects of the screening test parameters. This is true for the ABR, but the linkage is not very strong for the OAE. When defining the AABR screening test protocol, the key question is: What is the target hearing loss range and what ABR stimulation and recording parameters will detect that loss? The important variables include stimulus type, level, rate, number, etc., as well as signal-processing parameters.

For the OAE, the key question is different and is expressed better as: What are the best parameters for eliciting an OAE, and given those parameters, what are the effects of various types, degrees and frequency profiles of hearing loss on the ability to detect an OAE? Here, the choice of parameters for OAE screening is oriented towards maximizing the likelihood of obtaining a clear OAE, without regard to the target disorder. Here, in a sense the cart (i.e., the screening test) is put before the horse (i.e., the target disorder) and it is fortunate that by and large, the hearing losses that are deemed to be of interest are such that OAE screening has reasonable performance. That would not be the case if the focus were upon say 50+ dB of hearing loss, or if it were 20+ dB, whereas the ABR test condition would simply be adjusted to reflect the modified target definition. Of course, it is quite possible that the known or presumed properties of the OAE have influenced the choice of target disorder definitions, either covertly or on pragmatic grounds.

Even for the ABR, several factors mediate the relationship between the target minimum hearing level and the screening stimulus parameters. For a stimulus of a given HL, at certain frequencies the absolute SPL at the tympanic membrane is on average some 10 to 15 dB greater in a neonate's ear than in an adult ear, depending on the meatal cavity volume. The ABR detection threshold is not identical to the perceptual threshold, even in an adult. There may also be temporal summation effects. Thus, for example, a 20 dB HL minimum loss target does not translate directly into a 20 dB HL or 20 dB nHL stimulus level for ABR-based screening. It would be advantageous to define screening levels in terms of actual SPLs in individual ears.

Audiometric Assessment

A basic principle of public health screening is that all children who do not pass the screen should have access to timely and effective services for confirmation, quantification, and diagnosis of the target disorder. Challenges here include: (a) the quality of individual tests in the assessment protocol, (b) the validity of the overall diagnostic assessment, especially the decision rules linking the various patterns of test outcomes to specific courses of action and diagnoses, and (c) the unprecedented level of interdisciplinary collaboration required, in order to achieve overall program effectiveness and a high quality of care.

Quality of Testing

The cornerstone of the audiological assessment is valid and accurate estimation of the puretone audiogram for all frequencies that are important in decision-making about service needs. Air conduction and bone conduction threshold estimates are required. The information that is strictly necessary for good diagnostic and intervention-related decision-making may not be as comprehensive as might be thought. Because of the effort and potential difficulty of getting both accurate and comprehensive estimates of the puretone audiogram using threshold tonepip ABR methods in a young infant, there is an acute need to define and justify precisely which information is absolutely necessary. A hierarchy of importance of audiometric information needs to be developed, so that audiometric effort can be directed as effectively and efficiently as possible. A basic question might be "What is the importance of the various frequencies and hearing loss severity estimates in the selection of a hearing aid?" For example, is 4 kHz more important than 500 Hz? If we have results at 500Hz and 2 kHz, how important is 1 kHz? How important is a 10 dB or 20 dB range of uncertainly in threshold estimation? Is the acceptable accuracy different for different frequencies and severities of hearing loss?

Tonepip ABR is currently the tool of choice for frequency-specific estimation of puretone audiometric thresholds in sleeping neonates or infants under about six months of age. The methods and results have been described in detail (Stapells, 2000). Given the proper technique, it is usually possible to obtain accurate audiometry, including by air and bone

conduction, for a wide range of hearing loss types, etiologies, severities and frequency profiles. In practice, there will be some infants for whom it is difficult to obtain the range and precision of threshold estimates desired within a reasonable time frame.

A significant concern in program delivery is how to achieve and maintain a very high quality of ABR threshold measurement. The immediate problem arises because tonepip ABR threshold testing is much more demanding than otoneurologic click ABR testing. Its use is not yet very widespread, and it is relatively unlikely to be well-taught in training and continuing education programs. Especially for low-frequency stimuli, the tonepip ABR waveform is completely different from the standard click response, and the signal-in-noise response recognition problem is exacerbated by the need to lower the high-pass cut-off frequency of the recordings. Artifact rejection management and the number and length of averages are other areas requiring skill. Because of variation across individuals in the size of the ABR and the level and characteristics of EEG noise, response detection protocols are necessarily adaptive, and interpretive criteria must take individual subject characteristics into account. There are as yet few valid computational tools that would assist subjective measurement strategies and interpretive judgements, for tonepip ABR measurements.

Validity of Overall Assessment

Confirmation of hearing loss and audiological diagnosis (as distinct from etiological diagnosis, the domain of the physician) are based on the ABR but include a broader palette of measurements. The set of audiometric procedures will currently include otoneurologic ABR, to assess the functional status of auditory brainstem pathways; DPOAE, as an index of cochlear function; high-frequency (i.e., 660Hz) otoacoustic immittance and acoustic reflexes, and visual reinforcement audiometry when feasible. Parent/caregiver behavioural reports can be useful if they are valid and timely. Note that most of these audiometric assessments should be occurring at less than three months of age, hence the emphasis on electrophysiological measures.

There is a clear need to ensure that not only are diagnostic protocols appropriate but that there is a sufficient number and distribution of assessment sites and staff. The staff must be adequately trained and possess appropriate and in several respects novel skills. Assessment in the context of an EHDI program is even more difficult and demanding than traditional audiological assessment in older children (i.e., aged one year and over). One dimension is that there are many areas of limited knowledge and uncertainty relating to assessment and management in young infants. Another dimension is the potential costs of error. It is bad enough to make an audiometric error in, say, a two-year old child referred for audiological assessment because of parental and pediatrician concern. Contrast that with making the error in a baby aged three months, who is presenting because of screening test failure. One problem scenario is a misinterpretation of ABR records and a false conclusion that hearing is within normal limits, only to have it confirmed a year later, after the parents have sought a second or third opinion, that the screening result was entirely correct. This can and does happen, and only a few such events seriously compromise the rationale for a universal program, which is predicated on accurate diagnostic assessment and effective intervention.

The basic principles of overall assessment can be extrapolated from older children and adults: integrating the airbone gap, tympanometry and reflexes, tonepip ABR and reflex thresholds, OAE and otoneurologic ABR, and so on. If there is a clear discrepancy, a selective ABR retest directing effort to very high accuracy and confirmation of previous findings is often warranted. The possibility of progression or fluctuation in hearing levels, even in bone conduction thresholds, should not be forgotten. Attention to risk factors may provide important additional diagnostic and prognostic information.

The findings of the various component tests in the diagnostic assessment should be integrated and the picture should be monitored and reassessed periodically. Discrepancies among test components should be pursued diligently. Provisional decisions on the basis of incomplete or inconsistent information should be reviewed regularly. All manner and types of evidence supporting the audiometric determination should be sought. There should be a clear and enduring intent to achieve accurate behavioural corroboration of electrophysiologic measures as soon as is possible, although of course it may not be possible. Parental report that is not consistent with the clinical audiometric picture should not be dismissed but should be fully debated and should trigger critical re-evaluation.

An example of a difficult situation is one in which the ABR shows a clear bilateral hearing loss but the parent insists that the child responds to sound. If the OAE and reflexes are normal then the clarity and reproducibility of the ABR threshold estimates and the normality of EEG noise levels should be confirmed. The parent should be involved in the validation or refutation process. There are many other problematic situations, and it is necessary to develop defensible procedures covering the major alternatives, but a full discussion is be-

yond the scope of this article. Gravel (2000) has offered some case examples that are cautionary and instructive.

The primacy of electrophysiological measurement in neonates and very young infants is justifiable, but only if those measurements are of high quality. Bad electrophysiology is even worse than bad behavioural testing, because of the illusion of objectivity. In fact, at present, ABR threshold estimation is a subjective art. Even given a high level of expertise, it should not be forgotten that the ABR is a statistical correlate of auditory perception, not a mediator of it. Certainly, there are several reasons why the ABR may under- or over-estimate hearing sensitivity substantially, albeit in a small proportion of cases. An especially problematic situation is that in which ABR morphology is highly abnormal, due to a retrocochlear disorder. This could lead to screening failure because the expected response waveshape does not occur. At diagnostic testing, a depressed or absent wave V might be seen. In this situation, tonepip threshold estimates based on wave V may be unreliable. It is prudent to repeat diagnostic ABR testing periodically, to consider testing with other evoked potentials such as steady-state or middle-latency potentials, and to give special weight to clues about true hearing levels from behavioural and acoustic reflex measures. By six months of age, the response to conservative hearing aid fitting may also give audiometric clues.

Need for Effective Collaboration

The additional complexities of audiological assessment in early infancy, as well as the pressures to intervene early, place an extraordinary emphasis on the achievement of a seamless, continuous and consistent pattern of care. There is really no place in a high-quality program for conflicting views from professionals about the accuracy of the audiometry or the relative merits of this or that type of test. It is important that key professionals involved express consistent views and behave as a coherent team. There will probably be a need to develop those views and come up with a strategy for consensus and guidelines development. The traditional problems of territoriality, dogma, inefficiency and plain ignorance have to be overcome, by a goal-directed, evidence-based, consensus-building process that may involve otolaryngologists, pediatricians, family physicians, audiologists, public health nurses, speech-language pathologists, educators and others. For example, it is pointless to initiate an EHDI program only to be met with one month delays for specialist referrals, outdated attitudes (e.g., "Let's wait and see, he'll probably grow out of it"), misinformed comments about hearing aids (e.g., "She won't benefit much because her other ear is pretty good"), or injudicious remarks (e.g., "We won't really know about her hearing until we can test her behaviourally"). The elimination of all of these well-known problems is perhaps one of the biggest challenges facing EHDI program planners. One difficulty is that the true diversity and extent of such problems is difficult to quantify and is not always reflected in standard program evaluation tools. Unwarranted delay is easier to detect, because the date of every important event should be recorded in the program information system, and delay and age criteria form part of the basic set of performance indicators. It is harder to detect and measure inappropriate remarks, but systematic exploration of parental/caregiver satisfaction with every major aspect of the process of care can be very revealing and should be undertaken routinely as part of the program quality management.

All in all, the planning and implementation of an EHDI program present remarkable challenges as well as opportunities to improve radically current levels of professional understanding and interdisciplinary collaboration, in hearing health care for young children.

Summary

UNHS is being endorsed and implemented widely but by no means universally. The modest evidence base for its effectiveness may be improved by a shift of emphasis from speech and language development to the right of the infant to hear.

It is essential to look beyond screening at the entire program for early detection and intervention. The programmatic view emphasizes integration of screening, audiologic assessment, diagnosis and intervention into a seamless, coherent whole. A consensus on goals, crystal clear objectives, a powerful information system, a family-centred style of care and strong, ongoing quality evaluation are necessary for success.

Specific screening areas needing attention include the definition of the target disorder, prevalence, coverage, and test sensitivity and specificity. Issues in each of these areas are discussed, including the relative merits of different protocols. Audiologic assessment is based predominantly on tonepip threshold ABR, but a strategic approach that includes many other sources of audiometric information is necessary to avoid errors.

Early identification and intervention programs present remarkable challenges and opportunities for professionals to improve knowledge and skills, revise and rationalize practices, and develop interdisciplinary cooperation.

Author Note

Please address all correspondence to M. L. Hyde, Suite 201, Mount Sinai Hospital, 600 University Avenue, Toronto, ON M5G 1X5; telephone 416-586-4510, fax 416-586-8739, or email mhyde@mtsinai.on.ca.

References

American Academy of Pediatrics. (1999). Newborn and infant hearing loss: Detection and intervention. Task Force on Newborn and Infant hearing. *Pediatrics, 103,* 527-530.

Bamford, J., Davis, A., Hind, S., McCracken, W., Reeve, K. (2000). Evidence on very early service delivery: What parents want and don't always get. In R. C. Seewald (Ed.), *A sound foundation through early amplification* (pp. 151-157). Staefa, Switzerland: Phonak AG.

Bess, F. H., Dodd-Murphy, J., & Parker, R. A. (1998). Children with minimal sensorineural hearing loss: Prevalence, educational performance and functional status. *Ear and Hearing, 19,* 339-354.

Feightner, J. W. (1992). Screening in the 1990s: Some principles and guidelines. In F. H. Bess & J. W. Hall III (Eds.), *Screening children for auditory function* (pp. 1-16). Nashville, TN: Bill Wilkerson Center Press.

Fortnum, H., & Davis, A. (1997). Epidemiology of permanent childhood hearing impairment in Trent Region, 1985-1993. *British Journal of Audiology, 31,* 409-446.

Goldbloom, R. B. (1997). Weighing the evidence: The Canadian experience. *American Journal of Clinical Nutrition, 65,* 584S-586S

Grandori, F., & Lutman, M. (1999). The European Consensus Development Conference on Neonatal Hearing Screening (Milan, May 15-16, 1998). *American Journal of Audiology, 8,* 19-20.

Gravel, J. (2000). Audiologic assessment for the fitting of hearing instruments: Big challenges from tiny ears. In R. C. Seewald (Ed.), *A sound foundation through early amplification* (pp. 33-46). Staefa, Switzerland: Phonak AG.

Gravel, J., Berg, A., Bradley, M., Cacace, A., Campbell, D., Dalzell, L., DeCristofaro, J., Greenberg, E., Gross, S., Orlando, M., Pinheiro, J., Regan, J., Spivak, L., Stevens, F., & Prieve, B. (2000). New York State Universal Newborn Hearing Screening Demonstration Project: Effects of screening protocol on inpatient outcome measures. *Ear and Hearing, 21,* 131-140.

Harrison, M., & Roush J. (1996). Age of suspicion, identification and intervention for infants and young children with hearing loss. A national study. *Ear and Hearing, 17,* 55-62.

Hyde, M. L., Davidson, M. J., & Alberti, P. W. (1991). Auditory test strategy. In J. T. Jacobson & J. L. Northern (Eds.), *Diagnostic audiology* (pp. 295-322). Austin, TX: Pro-Ed.

Hyde, M. L., Riko, K., & Malizia, K. (1990). Audiometric accuracy of the click ABR in infants at risk for hearing loss. *Journal of the American Academy of Audiology, 1,* 59-66.

Hyde, M. L., Sininger, Y., & Don, M. (1998). Objective detection and analysis of ABR: An historical perspective. *Seminars in Hearing, 19,* 97-113.

Jacobson, J., & Jacobson, C. (1987). Principles of decision analysis in high-risk infants. *Seminars in Hearing, 8,* 133-141.

Joint Committee on Infant Hearing. (2000). Year 2000 Position Statement: Principles and guidelines for early hearing detection and intervention programs. *American Journal of Audiology, 9,* 9-29.

Mayne, A. M., Yoshinaga-Itano, C., Sedey, A. L., & Carey, A. (2000). Part 1: Language – expressive vocabulary development of infants and toddlers who are deaf or hard of hearing. *The Volta Review, 100*(5), 1-28.

Moeller, M. P. (2000). Early intervention and language development in children who are deaf and hard of hearing. *Pediatrics, 106,* e43.

National Institute of Health. *(1993).* Early identification of hearing impairment in infants and children. NIH Consensus Statement (vol. 11). *Bethesda, MD: Author.*

Norton, S., Gorga, M. P., Widen, J. E., Folsom, R. C., Sininger, Y., Cone-Wesson, B., Vohr, B. R., Mascher, K., & Fletcher, K. (2000). Identification of neonatal hearing impairment: Evaluation of transient evoked otoacoustic emission, distortion product otoacoustic emission, and auditory brainstem response test performance. *Ear and Hearing, 21,* 508-528.

Sininger, Y. S., Hood, L. J., Starr, A., Berlin, C. I. & Picton, T. W. (1995). Hearing loss due to auditory neuropathy. *Audiology Today,* 7, 10-13.

Stapells, D. (2000). Frequency-specific evoked potential audiometry in infants. In R. C. Seewald (Ed.), *A sound foundation through early amplification* (pp. 13-31). Staefa, Switzerland: Phonak AG.

Steel, K. P. (1998). A new era in the genetics of deafness. *New England Journal of Medicine, 339,* 1545-1547.

Stein, L. (1999). Factors influencing the efficacy of universal newborn hearing screening. *Pediatric Clinics of North America, 46,* 95-106.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293.

Turner, R. G. (1988) Techniques to determine test protocol performance. *Ear and Hearing, 9,* 177-189.

U.S. Preventive Services Task Force. (1996). *Guide to clinical preventive service: Report of the U.S. Preventive Services Task Force* (2nd ed.). Baltimore, MD: Williams & Wilkins.

Woolf, S. H., Battista, R. N., Logan, A. G., & Wang, E. (1990). Assessing the clinical effectiveness of preventive maneuvres: Analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations. *Journal of Clinical Epidemiology, 43,* 891-905.

Yoshinaga-Itano, C., Coulter, D., & Thomson, V. (in press). The Colorado Newborn Hearing Screening Project: Effects on speech and language development for children with hearing loss. *Journal of Perinatology.*

Yoshinaga-Itano, C., Sedey, A., Coulter, D. K., & Mehl, A. L. (1998). Language of early and later identified children. *Pediatrics, 102,* 1161-1171.