Is it REAL? Research Evaluation for Audiology Literature

By Catherine V. Palmer, Elaine Mormer, Amanda Ortmann, David Byrne, Yi Ye, and Lindsey Keogh

© Dave Cutler/Images.com

Robyn Cox wrote a Page Ten article in the August 2004 *Hearing Journal* entitled "Waiting for evidence-based practice for your fittings? It's here!" She provided a real "how to" guide in terms of defining a clinical question, defining an acceptable level of evidence, and searching the literature electronically to find the research articles that answer that clinical question. Many of you know that she also had a more detailed article published on this topic in *JAAA* in 2005.

What we are talking about here, however, is a little different. We're going to discuss how a clinician needs to approach the one or two articles that an electronic search (e.g., PubMed, CINAHL, etc.) produces. Or, more commonly, how the clinician needs to approach research (or often marketing) materials presented by manufacturers and other individuals at meetings or during visits to the clinic.

Beginning students and clinicians often ask us, "But if it's published, shouldn't it be good?" First, we need to consider that there are many levels of "publication." Given the ease of desktop publishing and electronic dissemination, many things look published. Something may be published, but not subjected to peer review (i.e., a critical review by other scientists prior to acceptance for publication). On the other hand, a publication doesn't necessarily have to be peer-reviewed to be relevant to clinical decision making. And, even when something is peer-reviewed, the clinician must still look at what was done critically. The question isn't necessarily if it's good or bad but whether it is clinically relevant. The clinician needs to take responsibility for making this judgment, and that is more or less the main point of this article.

What we're going to discuss are some of the tools the clinician needs to determine a publication's practical relevance for a particular patient and setting. To make the job even

For the past 3 years, Jerry Yanz, PhD, senior staff audiologist at Starkey Laboratories, has organized a session at the American Academy of Audiology Convention that has focused on how hearing aid manufacturers use research evidence in the design and marketing of their products. Representatives of various manufacturers talk about their companies' product features and provide supporting evidence on the feature's function and its potential for patient benefit in the real world. The session is always popular, as the audience is excited to hear about all the new hearing aid features, and sometimes there is even a little heckling to liven things up.

This past April, Jerry invited Catherine Palmer, PhD, of the University of Pittsburgh to moderate the session. Following the lead of the previous moderator, Ruth Bentler, PhD, she challenged presenters to discuss their data in such a manner that the audience could be critical consumers. Their assignment was not just to present a few PowerPoint slides with bar charts and mean data, but to include such things as a discussion of the study design, estimation of appropriate sample sizes, variability of the data, statistical significance and practical significance—in other words, a higher standard than is usually required of presenters at this type of session. I suspect that the request to describe data in this degree of depth and detail had a few of the panelists checking and rechecking their alphas, partial etas, omega squares, and all other Greek things. Maybe they even reviewed a few notes from long forgotten stat classes.

To prepare the audience for what was to follow, Catherine started things off with a primer on evidence-based practice, some specific research design goals, and

a few details about the meaning of statistical and practical significance. The presenters from industry included Tim Trine, PhD (Starkey Labs), Francis Kuk, PhD (Widex), Victor Bray, PhD (Sonic Innovations), Stefan Launer, PhD (Phonak, Inc.), Tom Powers, PhD (Siemens, Inc), Don Schum, PhD (Oticon), and Laurel Christensen, PhD (GN ReSound). The presenters all rose to the occasion and what resulted was a great session, with lots of good data that actually could be critically consumed by the audience.

As you all know, evidence-based practice is a hot item in the selection and fitting of hearing aids today. Since Catherine's session was so informative and well received, it seemed to me that HJ *should try to get the word out to the rest of the world. What better way than for Catherine to write a Page Ten article on this topic, I thought (what, me biased?). It really didn't take any convincing, and an October publication date was set.*

With the help of the people in her research lab at Pittsburgh, Catherine put together an outstanding article—so outstanding that we decided to move it from Page Ten to this month's Cover Story to give it a bit more room to breathe, allow us more flexibility in the layout, and, most importantly, to encourage an even larger readership. It's not a typical Cover Story, but I'm betting it's one you'll want to keep!

Thanks, Catherine and Team Pitt, for a great contribution.

Gus Mueller
HJ Contributing Editor

tougher, much of the information clinicians must evaluate isn't published at all, but rather is presented in a variety of venues (papers/summaries from manufacturers, local and national presentations, etc.).

MAKE IT REAL

The REAL in the title of this article is an acronym our lab came up with that stands for "Research Evaluation for Audiology Literature." We think the clinician needs to be familiar with five aspects of a study or research report, and we will discuss each of them in detail:

- Level of evidence
- Rationale for the sample size chosen
- Variability in the data
- Determination of statistical significance
- Practical significance of the findings

These are all items that should be (but not always are) included in the reporting of a study. Clinicians need to ask the researcher or presenter about any of these items that are missing from a presentation. They also need to know how to interpret them in light of their specific clinical goals for particular patients.

At this point, you might be thinking you need a refresher class in statistics. The good news is that you don't need to calculate any of these items; that's the job of the investigator. You only need to look for these items and then understand their impact on interpreting the data. We'll go through each one with some examples, and that should help you understand how all of this works.

By the way, now that a clinical doctorate (AuD) is required for entry into the audiology profession, many academic programs are requiring some level of statistics courses, experimental design, and participation in research projects by their students with the emphasis on creating critical consumers of the literature. Any reasonable patient would expect that his or her "Doctor of Audiology" is reading and critically applying the current literature to any recommended treatments. Of course, regardless of your title, if you are providing hearing aids for the treatment of hearing loss, you, as a professional, have to be a critical consumer of the information the manufacturer is providing about product features. Patients are not going to know if having 10 or 20 channels of compression will make any difference for their particular hearing loss in their particular communication situations.

LEVELS OF EVIDENCE

The tables from Robyn Cox's 2004 article regarding levels of evidence are very useful, and we've reprinted them here. There are a few ways to categorize the design of an experiment, but the systems shown in Tables 1 and 2 are widely used. Table 1 describes the quality of the design and Table 2 provides a way to "grade" a recommendation that is based on the evidence. Systematic reviews are wonderful for clinicians because they mean someone else has taken the time to find, evaluate, and report on the literature that answers a particular clinical question. It would be great if we had more of these in audiology.

A meta-analysis is really a systematic review where you can combine studies in order to increase the power of the data, which makes the findings more compelling. There are very few meta-analyses in our field because researchers often use quite different outcome measures that cannot be combined. In other cases, a given hearing aid feature might be too new for well-designed studies to have been conducted and published.

Generally, the higher the grade, the better, and it is certainly better to have a well-designed study with a high grade. But, the

SUBJECT SELECTION

The clinician will want to know many details about the subjects included in a study when determining if its findings are relevant to a given clinical situation. Here are some important questions to ask:

- How and from where were the subjects selected?
- Did the subjects get paid?
- Do the subjects get to keep the "new" hearing aids?
- What types of biases were in the study (e.g., using the term "new" compared to "old" hearing aid.)?
- What was the age range?
- What was the level of education of the subjects (will this generalize to a larger population)?
- What was the socioeconomic status of the participants (were they all college students, college graduates)?
- Were both men and women included?
- What types, degrees, and configurations of hearing loss were included?
- Were these active individuals with high communication demands?
- Were the subjects blinded to the treatment condition?
- Was the researcher blinded to the treatment condition when measuring outcomes?

Table 1. *Describing the quality of the design (from Cox, 2004).*

(1) Systematic reviews and meta-analyses of randomized controlled trials
(2) Randomized controlled trials
(3) Non-randomized intervention studies
(4) Descriptive studies (cross-sectional surveys, cohort studies, case-control designs)
(5) Case studies
(6) Expert opinion

Table 2. *Grades of recommendation (from Cox, 2004).*

(A) Consistent level 1 or 2 studies
(B) Consistent level 3 or 4 studies or extrapolations from level 1 or 2 studies
(C) Level 5 studies or extrapolations from level 3 or 4 studies
(D) Level 6 evidence or troubling inconsistencies or inconclusive studies of any level

clinician needs to know more than the design and grade to decide if a study will be clinically relevant. Before worrying about the statistical issues we listed, the clinician needs to make sure the study is asking the question and using the treatments and outcome measures that are of interest to him/her clinically. As reviewed in Robyn's article, it's important to define the treatments and outcomes that are of interest to *you*. This makes a literature search much more efficient since it will come up with only relevant articles.

When you are confronted with data in a presentation or a manufacturer's literature, you must make sure that the treatments being compared are of clinical interest (e.g., Do you care if a new technology is better than a technology that isn't even available any more? Are you more interested in knowing how this new technology works compared to how normal-hearing individuals function or compared to a technology that you have been recommending?) Make sure the outcome measures are relevant to your patient.

And, is there real-world validity? For example, does the fact that treatment A is better than treatment B for listening to words in a background of white noise in a sound booth with speakers at specific locations tell you how your patient is going to do while moving around at a cocktail party listening to connected speech in a background of multitalker babble?

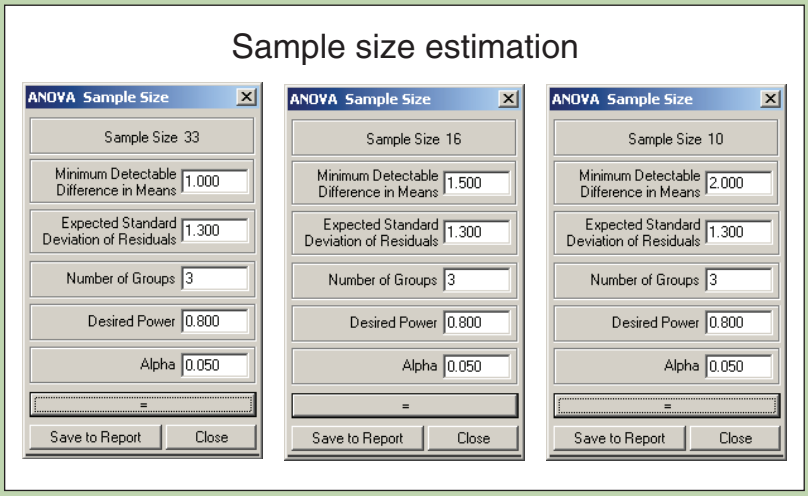
Remember, new treatments can be tested at several levels. When a new treatment (signal processing, etc.) is being tested it will first be tested for *efficacy*. This asks if the treatment does what it is intended to do under ideal conditions (e.g., usually an anechoic chamber or test booth with specific signals and noises). This is an important first step, but it does not generally tell clinicians what they need to know to apply the treatment to real patients. The clinician is looking for the level of investigation that tests *effectiveness* (i.e., Does the treatment work in the real world?).

HOW MANY STUDY PARTICIPANTS IS ENOUGH?

Let's talk about sample size. The "sample" refers to the individual "subjects" in the study. The politically correct term for these subjects is currently "participants." The first thing to consider about the sample is whether or not these people reflect the population in which you are interested. If the research is providing findings from an older male population with moderate to severe hearing loss primarily from noise exposure, will this translate into how your younger patients with mild hearing loss are going to function?

It's always important to consider how many participants were in the study. There seems to be a common belief that a number around 25 is required. This isn't true. When a study is designed appropriately, the investigator calculates before the study is started how many individuals will be needed to find the answer to the question being asked. This is called a "sample size calculation." The clinician doesn't have to understand everything that goes into this calculation; he or she simply needs to ask the presenter if a sample size calculation was conducted and what was the result (e.g., How many subjects were needed in each group?). If enough clinicians ask, hopefully we'll see this become part of every presentation. It certainly should be part of every published study. For some studies you might need only five people in each group, but for others you may need 40 or more in each group.

We promised there wouldn't be much math, but this might be starting to sound a bit confusing, so let's use an example from Tim Trine's talk at our AAA meeting. Take a look at Figure 1. This is a nice example because it shows the various items that go into estimating the needed number of subjects in each group. It also shows that if you change one of the numbers, that changes how many individuals you need in every group. So, for investigators



(courtesy of Tim Trine, Starkey Labs).

Figure 1. *Sample size estimation calculations.*

to calculate how many participants they need per group, they must have an estimate of the difference they are expecting (this comes from previous data or a reasonable guess if there are no previous data).

The investigator also needs to know the expected variability in the data (e.g., "expected standard deviation of residuals," Figure 1). Again, this will come from either related data in the literature or from a reasonable estimate based on the investigator's knowledge of the items under study. The number of groups that will be compared also is important to this calculation. For example, whether you are comparing two or three different treatments affects how many people you'll need in each treatment group).

The fourth box in Figure 1 is labeled "desired power." Power is a number between 0 and 1 that (when multiplied by 100) refers to the percentage of samples that lie in the region of rejection (meaning that they are different from the sample). The investigator sets a level of power that is appropriate for the purposes of the study (e.g., power could be a number like 0.2, 0.6, 0.8, etc.).

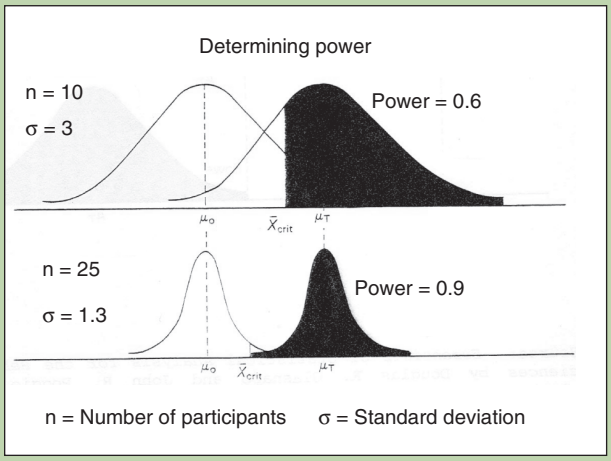


Figure 2. *Understanding power (e.g., Is the design powerful enough to find a difference if one really exists?).*

Figure 2 provides an example of different levels of power that could be chosen in an investigation. You can see that as the "power" increases (to 0.9 in Figure 2), the investigator will be more likely to see a difference between the two populations if one truly exists. In our literature, 0.8 is very often chosen as the level of power. This means if the two population samples are truly different, you will find this difference 80% of the time.

The last item shown in Figure 1 is alpha. You're probably familiar with this, as it is used in data analysis to determine if there is a significant difference between groups. But, the researcher actually chooses the alpha level before the experiment starts, and this choice also affects how many people will be needed in each experimental group. The alpha level refers to how sure the investigator wants to be that the results don't happen by chance. So, an alpha of 0.05 means there is a 5% chance that you will say there is a difference between treatments when there isn't one. Most studies in our field are designed with a 0.01 or 0.05 alpha level, meaning that the investigator is okay with finding a difference between treatments when none really exists 1% or 5% of the time, respectively.

While it is not necessary for clinicians to understand power values and alpha levels to interpret a study, it is important that they have a general understanding that every study is designed to allow some margin of error. These choices dictate the final number of participants needed for the study (i.e., sample size).

In our panel discussion, Stefan Launer provided a nice example of how an "under-powered" study can lead to incorrect findings. Table 3 shows a printout from a statistics program. The bolded number is called the "p-value," which is compared directly to the alpha level to determine if there is a difference between experimental groups. As we mentioned before, the investigator chooses the alpha level to calculate how many participants are needed in the study, and this is the level that has to be achieved to say there is a difference between groups.

SOME KEY TERMS TO REMEMBER

Alpha: Probability level chosen by the researcher. Alpha = 0.05 means there is a 5% chance that a performance difference demonstrated between two sample study groups does not truly exist in the general population.

Effect size: The amount of difference the researcher is actually interested in detecting with a given study. For example, one study may be looking to detect a difference of 1.5 dB between groups (small effect size), whereas another may be interested in only a 20-dB or greater difference between groups (large effect size).

p value: The probability that a difference in sample group results would occur by chance even when there is no true difference in population groups (see alpha). Typically chosen as p = 0.01 or 0.05.

Power: The probability that study results will find a difference between sample group performance scores when a true difference between population scores does in fact exist. Greater power (0.80 and above) is desirable in interpreting research results.

Practical significance: A finding can meet the requirements of statistical significance but not have practical value. For example, a study with a large enough sample size might find a statistically significant difference in scores of 5%, when a 5% difference has no real impact on a listener's overall function.

Sample size: A sample is a subset of a population used to make inferences about the characteristics or behavior of the population. The size of the sample is the number of observations or measurements made. Typically, a larger sample size leads to more precise estimates of various population properties (see power).

Statistical significance: A finding considered to be statistically significant means that it is highly unlikely the finding would have occurred by chance, based on the chosen alpha level and effect size.

So, for this example investigation (upper portion of Table 3) an alpha of 0.05 was set and the calculated p-value was 0.052. In order to say there is a significant difference between groups, the final statistical result would have to be 0.05 or less. Therefore, 0.052 reveals that there is no significant difference between groups. (With statistics it either is or isn't significant. "Sort of significant" is not an option.)

So, in this example, the question now becomes, is there really no difference, or was there not sufficient power to see a difference that truly exists? In the lower portion of Table 3, 10 subjects per group are used in the same study and you can see that now the study found a significant difference between groups (i.e., 0.003, which is smaller than 0.05). Launer set up this example for the purpose of instruction and worked backwards to calculate the power that was achieved in each design. In the design with five subjects per group, the resulting power was only 0.56. That means that if the two population samples are truly different, you will find this difference only 56% of the time. When 10 subjects participated in each group, the resulting power was 0.87, meaning that a difference would be seen 87% of the time.

This example illustrates why it is important for the desired power to be determined as part of the original sample size calculation. This enables the study to be designed so that it will begin with the appropriate number of participants. You don't want too few subjects because you might have incorrect results, and you don't want too many subjects because it is a waste of time and resources.

Table 3. *The results for five and ten participants per group in an example study.*

T-Test								
5 participants								
				95% Confidence interval of the difference				
	Mean	Std. dev.	Std. error of mean	Lower	Upper	t	df	p
A vs B	-1.29	1.05273	4,708	-2.597	.01714	-2.74	4	**0.052**
10 participants								
				95% Confidence interval of the difference				
	Mean	Std. dev.	Std. error of mean	Lower	Upper	t	df	p
A vs B	-1.19	.9293	.2938	-1.862	-.53319	-4.08	9	**0.003**

Courtesy of Stefan Launer, Phonak, Inc.

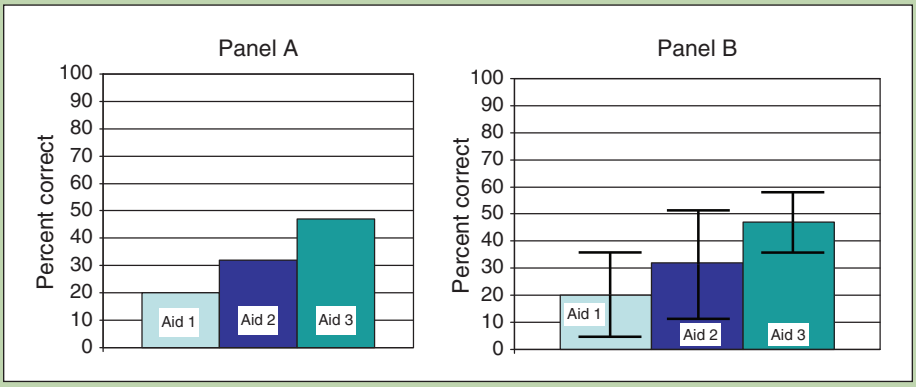


Figure 3. *Panel A: Data without standard deviation bars; Panel B: Data with standard deviation bars.*

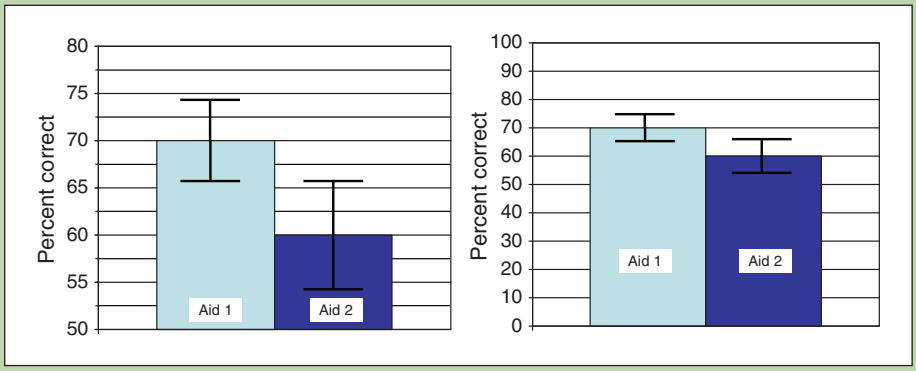


Figure 4. *The same data displayed with two different scales on the y-axis.*

LOOKING AT THE DATA

When looking at data, there are three important issues to be concerned with: variability, statistical significance, and practical significance.

Let's start with variability. You all know that today it's easy to put together a pretty bar chart in PowerPoint. You've no doubt been at presentations when you've been encouraged to notice that one bar is bigger than another, and sometimes, that's the extent of the teaching point. Our contention is that if the variability isn't represented on the graph (usually through standard deviation or standard error of the mean bars), then the chart is pretty meaningless.

To illustrate why the degree of variability is important for the clinician's interpretation of how meaningful the data are to his or her patients, look at Figure 3 (Panel A). If these were the results of three different hearing aid treatments (let's say hearing aids from three different companies), one would be pretty impressed with the differences and assume that Hearing Aid 3 was far better than the other two. You might even order a pair of them for your next patient.

In Panel B of Figure 3, however, we now see there is a lot of variability in the data. Although the mean differences are fairly large, the groups overlap in performance significantly. Notice that the bars overlap the mean results between groups. So, performance with each of the three hearing aids actually was similar for some subjects. The clinician would be much more convinced to order Hearing Aid #3 if the standard deviation bars didn't overlap with the other mean values.

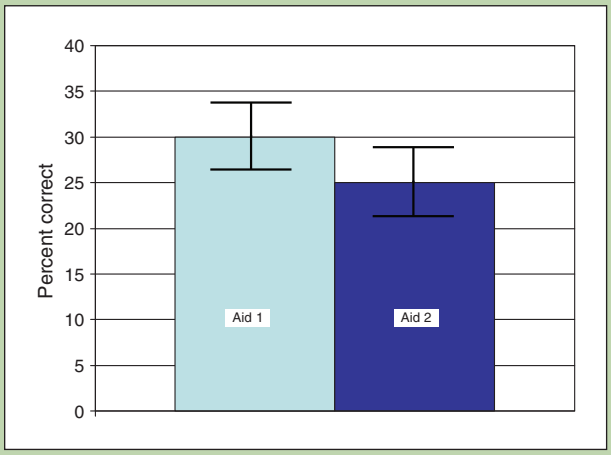


Figure 5. *Watch for the y-axis when interpreting results and choosing treatments.*

A WORD FROM AN EBP GURU

Kudos to manufacturers for increasingly turning to carefully designed research to motivate innovations in hearing aid design. Nevertheless, the dispenser is still the one responsible for independently reviewing evidence to come up with the best possible recommendation for each patient.

Practitioners are fortunate that Catherine Palmer is able to take a novel topic like evidence-based practice and break it into manageable chunks, as we see in this article. I encourage you to spend a little time with this cover story and try out some of the ideas soon. And, when you read about research evaluating new technology, ask yourself some questions:

For example, how many subjects participated? How were they recruited? Are they representative of *your* patients? Is there a potential for bias in the way the study was conducted? If so, was there appropriate blinding of both subjects and data collectors? If the new technology was statistically better than the old, was the effect size large enough to warrant the additional cost of this technology?

Once we get into the habit of looking at data this way, we will discover that EBP is something we really can do!

Robyn Cox

When you see large variability, such as that shown in Panel B of Figure 3, it is worth giving some thought to the cause. Was the outcome measure that was used a good test, or do you expect a lot of variability? Was the sample chosen so individuals were fairly similar (similar hearing levels, similar etiologies, etc.) or would the differences between subjects predict variable performance? Of course, these are items that should have been controlled in the study design. But they are not always controlled and so are worth some thought by the clinician since he or she is trying to apply these results to specific patient populations.

A separate point to consider is how the data are displayed. The clinician should always take a second to orient him/herself to the y-axis of the graph when data are being presented. One reason is that it can be very misleading to change the scale to point out differences.

Look at Figure 4. At first glance, it would be easy to interpret this set of graphs as showing greater differences between the two treatments shown on the left graph and smaller differences between the two treatment conditions shown on the right graph. These actually are all the same data. In the left graph, the scale goes from 50% to 80% whereas the scale on the right graph goes from 0 to 100%. We would suggest that generally the appropriate way to plot data is to use the entire range of the test being used. That is, if it were possible for participants to score anywhere from 0% to 100% on a given test, that is what should be used on the y-axis.

Figure 5 illustrates another problem you encounter when the full scale of possible results isn't used. In Figure 5, we can see that Hearing Aid 1 is better than Hearing Aid 2. (We'll assume that these have been tested to verify that they really are statistically different.) But, when the clinician looks at the y-axis, it becomes evident that although the two treatments are different, neither will be very helpful for the patient (providing scores of 25% and 30% correct). So, although Hearing Aid 1 is better in this comparison, there may be some other hearing aid that should be considered since neither of these treatments is providing even 50% speech understanding. Now, if we only showed you a slide with a "5% better" bar for Hearing Aid 1, you might assume that Hearing Aid 1 was the better choice when really neither result was particularly promising.

PROBABILITY AND STATISTICAL SIGNIFICANCE

Statistical significance really tells us that the results didn't happen by chance. Remember that earlier we discussed the selection of an alpha level—something like 0.05 for example. Recall that the researcher sets the alpha level before the study begins, so that is the level that must be achieved (that level or smaller) in the statistical analysis in order to say the treatments tested are truly different. Table 3 provided examples of tests of significance (often referred to as "p" value) in the bolded numbers in the table.

For the clinician, statistical significance can be considered the first step in interpreting differences between treatments. If there is no statistical difference between treatments (and you have verified that enough participants were included to achieve the power needed to find a difference if one existed and that extreme variability in the data caused by poor outcome measures or a poorly selected subject group didn't impact the ability to see a difference), then you can proceed confidently knowing that these

two particular treatments aren't different. This certainly would influence clinical decisions and recommendations.

In terms of published and presented data, we generally encounter treatments that have been found to be significantly different. After all, most treatment studies are designed to find differences that promote a particular new technology or signal processing scheme. Interestingly, clinicians would like to know when treatment results *aren't* different from each other, but non-significant findings are often difficult to publish and are rarely in the interest of manufacturers to present.

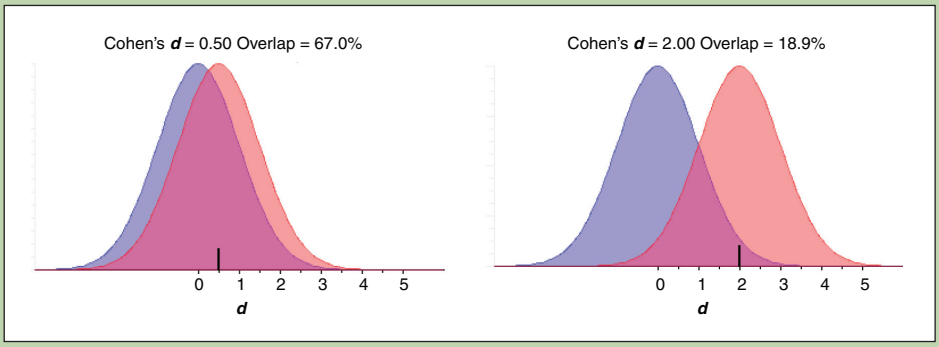


Figure 6. *Understanding Cohen's* ***d****.*

There are times, however, when a finding of no significant difference would be of interest and would be the goal of the investigation. We were involved in a study like this when we wanted to compare a new auto-adaptive directional hearing aid to a "gold standard." In this study, we were comparing new treatments (auto-adaptive directional hearing aids) to how normal-hearing individuals function (gold standard) in the same listening environment. So, not being different from normal-hearing individuals would be considered an excellent outcome.

As technology advances, I think we'll see more studies using normal function as the comparison condition, so we can expect to see more results where no difference is very positive and will help the clinician make recommendations of specific technology. It would be very motivating to be able to tell patients that a group of people similar to them using a particular signal processing scheme functioned as well as normal-hearing individuals.

But, let's talk now about how to think about results when there *is* a significant difference. Certainly, statistical significance between two treatments is the first step. But there is an important second step, which is to determine if this difference is *practically* significant as well. We talked about clinical relevance earlier (e.g., determining if the population being tested is similar to your patient, if the test set-up mimics the real-world environment, etc.). But *practical* significance tells us if the statistical significance is going to translate into a meaningful difference for the patient. That's our next topic.

WHEN IS A FINDING *PRACTICALLY* SIGNIFICANT?

When a statistical analysis is conducted, there are some well-accepted methods to look at *practical* significance. For example, a researcher may report Cohen's ***d***, which tells us the effect size of the difference (e.g., how much "non-overlap" there is between the two results). Or a researcher may choose to report partial eta squared (or eta squared or omega squared), which tells us the portion of the variance that is accounted for by membership in a particular treatment group. For example, if the partial eta squared is 56% then we might interpret the data to say that 56% of the outcome result is accounted for by being in the group that listened through the adaptive directional microphones.

Cohen's ***d*** is commonly used, so let's look at Figure 6 to get a better understanding of what it tells us and how we relate that

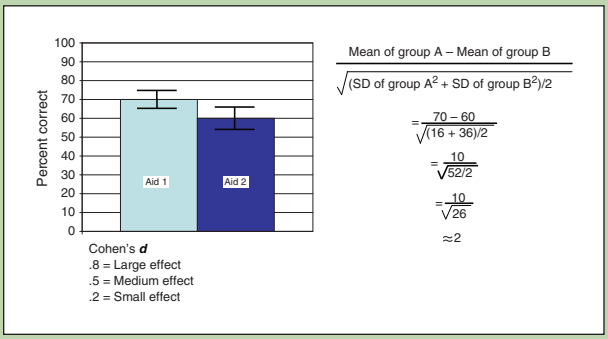


Figure 7. *Calculating effect size.*

to practical significance. Figure 6 provides two examples. In the first, Cohen's ***d*** is 0.5. This is determined by setting the peak of the first set of data at 0 and then calculating the distance between peaks. From the table in Figure 7 (below the graph), we can see that a Cohen's ***d*** of 0.5 is considered a medium effect size.

In the second example (right side of Figure 6), the Cohen's ***d*** is 2.0. You can see that there is less overlap between the two sets of data (i.e., the means are further apart). This is considered a large effect size. So, the bigger the "non-overlap" value, the bigger the Cohen's ***d***, and the larger the effect size. Only 18.9% of the outcome data overlap in the example with a Cohen's ***d*** of 2.0, whereas 67% of the outcome data overlap in the first example (left side of the figure). Less overlap gives you more clinical confidence that the treatment being evaluated will actually produce better results than the comparison treatment.

Two treatments may show a statistically significant difference, but there might be a small effect size. Clinicians have to use their expertise to decide if a small effect size would be important enough to the patient's ability to function (perhaps in noise, lectures, etc.) to justify recommending a particular treatment. And, we can't forget that the new treatment might cost much more than the comparison treatment.

We talked about efficacy and effectiveness research earlier. Practical significance has to do with yet another "e" word—efficiency (e.g., Is the treatment worth it?). If there is a large effect size (or in the case of partial eta squared, if the result accounts for a substantial portion of the variance), then the clinician may be more motivated to use this treatment. But, this doesn't mean there might not be times when a small effect size is very motivating. For example, with certain signal processing strategies used

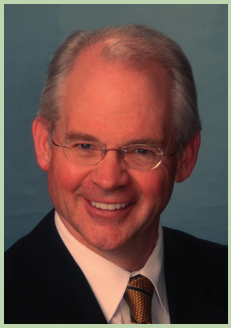
ON EVIDENCE-BASED HEARING AID DESIGN

The motivation for the first AAA session on evidence-based hearing instrument design in 2006 was to open a dialog between the hearing aid industry and hearing professionals and to provide a venue for information exchange, questions, answers, and honest critique. The emphasis was on improving practitioners' understanding of current hearing aid technologies and how industry uses evidence-based principles to develop those technologies. At the root of this effort was the belief that understanding arises not from marketing, but from solid scientific principles and evidence.

Accompanying the recent application of evidence-based practice (EBP) principles in the fitting of hearing aids, manufacturers have adopted an analogous evidence-based design (EBD) approach to developing hearing aids. In our summary of the 2007 session, Ruth Bentler and I defined EBD as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the design of hearing instruments by integrating individual clinical and engineering expertise with the best available external clinical and engineering evidence from systematic research."

The systematic use of EBD in the hearing aid industry serves two purposes. First, it provides a way to substantiate marketing claims about product features. This is a benefit to manufacturers. Second, and more important, it offers the best assurance that new features will meet patient needs and do what they are designed to do, before they ever reach the market. This is a benefit to patients.

The active participation at these sessions of key manufacturers—Oticon, Phonak, ReSound, Siemens, Sonic Innovations, Starkey, and Widex; of strong researchers—Ruth Bentler in 2006 and 2007 and Catherine Palmer in 2008; and a ready and able audience, has been much appreciated. It has been heartening to see the level of interest among practitioners in how the industry applies scientific principles to the design of hearing instruments. Speaking for the industry, I encourage our customers to continue this ongoing exchange, active questioning, and critique. We will all be better for it.



Jerry Yanz

with children a small difference can translate into a large difference in terms of speech and language development or educational achievement.

So, at this point, we would like to introduce a little math related to calculating effect size. It is really simple and can be very empowering for the clinician trying to interpret findings where the researcher has provided no information about practical significance (which happens more often than not). Figure 7 provides the data that were in Figure 4, but now the effect size has been calculated. So long as you have the means of the two

Several of the authors and others involved with this article are pictured in the University of Pittsburgh audiology lab. From left are Jenifer Powers, Kim Rorabaugh, Elaine Mormer, Janelle Kisiday, Catherine Palmer, Yi Ye, and Lindsey Keogh.

groups (taken right from the graph) and the standard deviation of each mean (taken right from the graph) and a calculator, you can calculate the effect size.

Just to walk you through the calculation, you subtract the mean of Group B from the mean of Group A and then divide by the square root of Group A's standard deviation squared plus Group B's standard deviation squared divided by two. Try it with the example in Figure 7.

In the example in Figure 7, the effect size (Cohen's ***d***) is 2.0, which translates into a large effect size. The Cohen criteria for small, medium, and large effect sizes are provided below the graph. If the researcher provides information about practical significance using partial eta squared, then there will be a statement along the lines of "50% of the variance is accounted for by membership in a specific treatment group." (Cohen's ***d*** can be translated into partial eta squared and vice versa).

PUTTING IT ALL TOGETHER

At this point, you may be thinking that in addition to being a little suspicious of every presentation you hear, you'll have to enroll in a statistics class, subscribe to 10 journals, and then close your office so you have time to read them.

It's really not that bad. To start with, the key is to follow Robyn's guidelines that we talked about earlier (available in the *HJ* archives at www.thehearingjournal.com). This will ensure that you end up reading only the articles that are relevant to your specific clinical question (for the rest of the journals you can skim topics to keep up to date). If you find that the subject population doesn't reflect your patients or the test set-up doesn't apply to how your patients will be using the technology, you probably can stop right there, because the results aren't going to be very relevant to your clinical situation.

As clinicians, it's our responsibility to be critical consumers of the published literature (or conference presentations) so we can provide the best possible treatment options for our patients—in terms of effectiveness and efficiency. This does take time and we have to fit that into our schedule. It helps when several clinicians can attack different questions and share their information in clinic staff meetings or journal clubs.

As we mentioned before, published systematic reviews are very helpful to the clinician, and we need more of them. But, you will continue to be faced with presentations reporting the results of studies comparing new technology with old, and you need the tools to interpret this information. Because the publication process takes so long, the clinician is forced to make decisions about using new technology and signal processing well before any information is published in the peer-reviewed literature.

It also is helpful (and much more efficient for the person trying to interpret the data) when the author or presenter provides the items described in this review (description of the design, sample size calculation, variability of the data, statistical significance, and practical significance). The more often clinicians ask for these items, the more likely those presenting data will be to include them routinely.

So that's it. As Gus said in his introduction, this isn't a typical *HJ* Cover Story. As you can tell, we feel strongly about clinicians being critical consumers of research. This approach is critical to the ethical care of patients.

We hope this article provides you with tools to critically evaluate research on various treatments (e.g., hearing aid features) for your patients. Our goal is to empower you to make the very best decisions for the hearing-impaired patients who rely on your advice.

Acknowledgments

Thanks go out to Tim Trine and Stefan Launer for letting us use some of the examples they used in their talks. We were assisted in trying to make this information useful and usable for students and clinicians by the AuD students in our lab at the University of Pittsburgh, including Jenifer Powers, Dave Jedlicka, Janelle Kisiday, Meghan Mueller, Reem Mulla, Kim Rorabaugh, Nicole Beninati, Anga Lao, Dionne Russell, and Whitney Hurley. Thanks to Jerry Yanz for organizing the evidence-based hearing aid design course at AudiologyNOW! each year. It's a great session for manufacturers, researchers, and clinicians. Special thanks to Gus Mueller who took what we thought was a useful article and made it even more useful with very helpful editing.

Catherine V. Palmer, PhD, is an Associate Professor at the University of Pittsburgh and Director of Audiology at the University of Pittsburgh Medical Center. **Elaine Mormer**, MA, is a doctoral student at the University of Pittsburgh and the Audiology Clinical Coordinator in the Department of Communication Science and Disorders at the University of Pittsburgh. **Amanda Ortmann**, MS, is a doctoral student at the University of Pittsburgh and an Audiologist at the Washington University Medical Center. **David Byrne**, MS, is a doctoral student at the University of Pittsburgh and a Research Audiologist at the National Institute for Occupational Safety and Health. **Yi Ye**, MD, is a doctoral student and Research Assistant at the University of Pittsburgh. **Lindsey Keogh**, AuD, is an Audiology Doctoral Fellow at the University of Pittsburgh and Veterans Administration Medical Center. Readers may contact Dr. Palmer at Audiology, 4th floor, Eye and Ear Institute, 203 Lothrop Street, Pittsburgh, PA 15213; at 412/647-6089 or 412/647-2455 (fax); or at palmercv@upmc.edu.