PART I

Overview and Foundational Issues
Rationale and Standards of Evidence in Evidence-Based Practice

OLIVER C. MUDFORD, ROB MCNEILL, LISA WALTON, AND KATRINA J. PHILLIPS

What is the purpose of collecting evidence to inform clinical practice in psychology concerning the effects of psychological or other interventions? To quote Paul’s (1967) article that has been cited 330 times before November 4, 2008, it is to determine the answer to the question: “What treatment, by whom, is most effective for this individual with that specific problem, under which set of circumstances?” (p. 111). Another answer is pitched at a systemic level, rather than concerning individuals. That is, research evidence can inform health-care professionals and consumers about psychological and behavioral interventions that are more effective than pharmacological treatments, and to improve the overall quality and cost-effectiveness of psychological health service provision (American Psychological Association [APA] Presidential Task Force on Evidence-Based Practice, 2006). The most general answer is that research evidence can be used to improve outcomes for clients, service providers, and society in general.

The debate about what counts as evidence of effectiveness in answering this question has attracted considerable controversy (Goodheart, Kazdin, & Sternberg, 2006; Norcross, Beutler, & Levant, 2005). At one end of a spectrum, evidence from research on psychological treatments can be emphasized. Research-oriented psychologists have promoted the importance of scientific evidence in the concept of empirically supported treatment. Empirically supported treatments (ESTs) are those that have been sufficiently subjected to scientific research and have been shown to produce beneficial effects in well-controlled studies (i.e., efficacious), in more natural clinical environments (i.e., effective), and are the most cost-effective (i.e., efficient) (Chambless & Hollon, 1998). The effective and efficient criteria of Chambless and Hollon (1998) have been amalgamated under the term “clinical utility” (APA Presidential Task Force on Evidence-Based Practice, 2006; Barlow, Levitt, & Bufka, 1999). At the other end of the spectrum are psychologists who value clinical expertise as the source of evidence more highly, and they can rate subjective impressions and skills acquired in practice as providing personal evidence for guiding treatment (Hunsberger, 2007). Kazdin (2008) has asserted that the schism between clinical researchers and practitioners on the issue of evidence is deepening. Part of the problem, which suggests at least part of the solution, is that research had concentrated on empirical evidence of treatment efficacy, but more needs
to be conducted to elucidate the relevant parameters of clinical experience.

In a separate dimension from the evidence–experience spectrum have been concerns about designing interventions that take into account the uniqueness of the individual client. Each of us can be seen as a unique mix of levels of variables such as sex, age, socioeconomic and social status, race, nationality, language, spiritual beliefs or personal philosophies, values, preferences, level of education, as well as number of symptoms, diagnoses (comorbidities), or problem behavior excesses and deficits that may bring us into professional contact with clinical psychologists. The extent to which there can be prior evidence from research or clinical experience to guide individual’s interventions when these variables are taken into account is questionable, and so these individual differences add to the mix of factors when psychologists deliberate on treatment recommendations with an individual client.

Recognizing each of these three factors as necessary considerations in intervention selection, the APA Presidential Task Force on Evidence-Based Practice (2006, p. 273) provided its definition: “Evidence-based practice in psychology (EBPP) is the integration of the best available research with clinical expertise in the context of patient characteristics, culture, and preferences.” The task force acknowledged the similarity of its definition to that of Sackett, Straus, Richardson, Rosenberg, and Haynes (2000) when they defined evidence-based practice in medicine as “the integration of best research evidence with clinical expertise and patient values” (p. 1).

Available research evidence is the base or starting point for EBPP. So, in recommending a particular intervention from a number of available ESTs, the psychologist, using clinical expertise with collaboration from the client, weighs up the options so that the best treatment for that client can be selected. As we understand it, clinical expertise is not to be considered as an equal consideration to research evidence, as some psychologists have implied (Hunsberger, 2007). Like client preferences, the psychologist’s expertise plays an essential part in sifting among ESTs the clinician has located from searching the evidence.

Best research evidence is operationalized as ESTs. Treatment guidelines can be developed following review of ESTs for particular populations and diagnoses or problem behaviors. According to the APA (APA, 2002; Reed, McLaughlin, & Newman, 2002), treatment guidelines should be developed to educate consumers (e.g., clients and health-care systems) and professionals (e.g., clinical psychologists) about the existence and benefits of choosing ESTs for specific disorders over alternative interventions with unknown or adverse effects. Treatment guidelines are intended to recommend ESTs, but not make their use mandatory as enforceable professional standards (Reed et al., 2002). The declared status, implications, misunderstanding, and misuse of treatment guidelines based on ESTs continue to be sources of controversy (Reed et al., 2002; Reed & Eisman, 2006).

Our chapter examines the issues just raised in more detail. We start with a review of the history and methods of determining evidence-based practice in medicine because the evidence-based practice movement started in that discipline, and has led the way for other human services. Psychologists, especially those working for children and young people, tend to work collaboratively with other professionals and paraprofessionals. Many of these groups of colleagues subscribe to the evidence-based practice movement through their professional organizations. We sample those organizations’ views. The generalizability to psychology of methods for evidence-based decision making in medicine is questionable, and questioned. Next we examine criteria for determining the strength of evidence for interventions in psychology. These criteria are notably different to those employed in medicine, particularly concerning the relative value to the evidence base of
research in psychology that has employed methods distinct from medicine (e.g., small-N design research). The controversies concerning treatment guidelines derived from ESTs are outlined briefly. The extent to which special considerations exist regarding treatment selection for children and adolescents is then discussed. Finally, we highlight some of the aspects of EBPP that require further work by researchers and clinicians.

EVIDENCE-BASED MEDICINE

Evidence-based practice can be perceived as both a philosophy and a set of problem-solving steps, using current best research evidence to make clinical decisions. In medicine, the rationale for clinicians to search for best evidence when making diagnostic and treatment decisions is the desire or duty, through the Hippocratic Oath, to use the optimal method to prevent or cure physical, mental, or social ailments and promote optimal health in individuals and populations (Jenicek, 2003). This section of the chapter provides an overview of the history of evidence-based practice in medicine (EBM), as well as a description of the current methodology of EBM. A critical reflection on the process and evidence used for EBM is also provided, leading to an introduction of the relevance of evidence-based practice (EBP) to other disciplines, including psychological interventions.

History of Evidence-Based Practice in Medicine

The earliest origins of EBM can be found in 19th-century Paris, with Pierre Louis (1787–1872). Louis sought truth and medical certainty through systematic observation of patients and the statistical analysis of observed phenomena; however, its modern origins and popularity are found much more recently in the advances in epidemiological methods in the 1970s and 1980s (Jenicek, 2003). One of the key people responsible for the emergence and growth of EBM, the epidemiologist Archie Cochrane, proposed that nothing should be introduced into clinical practice until it was proven effective by research centers, and preferably through double-blind randomized controlled trials (RCTs). Cochrane criticized the medical profession for its lack of rigorous reviews of the evidence to guide decision making. In 1972, Cochrane reported the results of the first systematic review, his landmark method for systematically evaluating the quality and quantity of RCT evidence for treatment approaches in clinical practice. In an early demonstration of the value of this methodology, Cochrane demonstrated that corticosteroid therapy, given to halt premature labor in high-risk women, could substantially reduce the risk of infant death (Reynolds, 2000).

Over the past three decades, the methods and evidence used for EBM have been extended, refined, and reformulated many times. From the mid-1980s, a proliferation of articles has instructed clinicians about the process of accessing, evaluating, and interpreting medical evidence; however, it was not until 1992 that the term evidence-based medicine was formally coined by Gordon Guyatt and the Evidence-Based Working Group at McMaster University in Canada. Secondary publication clinical journals also started to emerge in the early to mid-1990s, with the aim of summarizing original articles deemed to be of high clinical relevance and methodological rigor (e.g., Evidence-Based Medicine, ACP Journal Club, Evidence-Based Nursing, Evidence-Based Mental Health).

Various guidelines for EBM have been published, including those from Evidence-Based Working Group (Guyatt et al., 2000), the Cochrane Collaboration, the National Institute for Clinical Excellence (NICE), and the British Medical Journal Clinical Evidence group, to name but a few. David Sackett, one of the strong proponents and authors in EBM, describes it as an active
clinical decision-making process involving five sequential steps: (1) convert the patient’s problem into an answerable clinical question; (2) track down the best evidence to answer that question; (3) critically appraise the evidence for its validity—closeness to truth; impact—size of the effect; and applicability—usefulness in clinical practice; (4) integrate the appraisal with the practitioner’s clinical expertise and the patient’s unique characteristics, values, and circumstances; and (5) evaluate the change resulting from implementing the evidence in practice, and seek ways to improve (Sackett et al., 2000, pp. 3–4). The guidelines for EBM are generally characterized by this sort of systematic process for determining the level of evidence for treatment choices available to clinicians, while at the same time recognizing the unique characteristics of the individual’s characteristics, situation, and context.

It is not difficult to see the inherent benefits of EBM, and health professionals have been quick to recognize the potential benefit of adopting it as standard practice. Citing a 1998 survey of UK general practitioners (GPs), Sackett and colleagues (2000) wrote that most reported using search techniques, with 74% accessing evidence-based summaries generated by others, and 84% seeking evidence-based practice guidelines or protocols. The process of engaging in EBM requires some considerable understanding of research and research methods, and there is evidence that health professionals struggle to use EBM in their actual practice. For example, Sackett et al. (2000) found that GPs had trouble using the rules of evidence to interpret the literature, with only 20% to 35% reporting that they understood appraising tools described in the guidelines. Clinicians’ ability to practice EBM also may be limited by lack of time to master new skills and inadequate access to instant technologies (Sackett et al., 2000). In addition to these practical difficulties, there have also been some criticisms of EBM’s dominant methodology.

An early criticism of EBM, and nearly all EBP approaches, is that it appears to give greatest weight to science with little attention to the “art” that also underlies the practice of medicine, nursing, and other allied health professions. For example, Guyatt and colleagues cited attention to patients’ humanistic needs as a requirement for EBM (Evidence-Based Medicine Working Group, 1992). Nursing and other health-care disciplines note that EBP must be delivered within a context of caring to achieve safe, effective, and holistic care that meets the needs of patients (DiCenso, Cullum, Ciliska, & Guyatt, 2004).

Evidence-based practice is also criticized as being “cookbook care” that does not take the individual into account. Yet a requirement of EBP is that incorporating research evidence into practice should consistently take account of the patient’s unique circumstances, preferences, and values. As noted by Sackett et al. (2000), when these three elements are integrated to inform clinical decision making, “clinicians and patients form a diagnostic and therapeutic alliance which optimizes clinical outcomes and quality of life” (p. 1).

One of the key issues in the use of EBM is the debate around what constitutes best evidence from the findings of previous studies and how these various types of evidence are weighted, or even excluded, in the decision-making process. The following section first outlines the way in which evidence currently tends to be judged and then provides a more in-depth analysis of each type of evidence.

Levels and Types of Evidence

Partly through the work of Archie Cochrane, the quality of health care has come to be judged in relation to a number of criteria: efficacy (especially in relation to effectiveness), efficiency, and equity. Along with acceptability, access, and relevance, these criteria have been called the “Maxwell Six” and have formed the foundation of decision making around service provision and funding in the United Kingdom’s National Health Service (Maxwell, 1992). Other health systems
around the world have also adopted these criteria to aid in policy decision making, in both publicly and privately funded settings. Despite this there has been a tendency for evidence relating to effectiveness to dominate the decision-making process in EBM and, in particular, effectiveness demonstrated through RCTs.

There has been considerable debate about the ethical problems arising from clinicians focusing too much on efficacy and not enough on efficiency when making decisions about treatment (Maynard, 1997). In any health system with limited resources, the decision to use the most effective treatment rather than the most efficient one has the potential to impact on the likelihood of having sufficient resources to deliver that treatment, or any other, in the future. By treating one person with the most effective treatment the clinician may be taking away the opportunity to treat others, especially if that treatment is very expensive in relation to its effectiveness compared to less expensive treatment options.

According to current EBM methods, the weight that a piece of evidence brings to the balance of information used to make a decision about whether a treatment is supported by evidence can be summarized in Table 1.1. Although there are subtle variations in this hierarchy between different EBM guidelines and other publications, they all typically start with systematic reviews of RCTs at the top and end with expert opinion at the bottom.

Before discussing systematic reviews, meta-analyses, and clinical guidelines, it is important to understand the type of study that these sources of evidence are typically based on: the RCT. RCTs are a research design in which the participants are randomly assigned to treatment or control groups. RCTs are really a family of designs, with different components such as blinding (participants and experimenter), different randomization methods, and other differences in design. Analysis of RCTs is quantitative, involving estimation of the statistical significance or probability of the difference in outcomes observed between the treatment and control groups in the study. The probabilities obtained are an estimate of the likelihood of that size difference, or something larger, existing in the population.

There are a number of international databases that store and provide information from RCTs, including the Cochrane Central Register of Controlled Trials (CENTRAL; mrw.interscience.wiley.com/cochrane/cochrane_clcentral_articles_fs.html), OTseeker (www.otseeker.com), PEDro (www.pedro.fhs.usyd.edu.au), and the Turning Research Into Practice (TRIP; www.tripdatabase.com) database. Another effort to increase the ease with which clinicians can access and use evidence from RCTs has come through efforts to standardize the way in which they are reported, such as the CONSORT Statement and other efforts from the CONSORT Group (www.consort-statement.org).

The major strength of well-designed RCTs and other experimental designs is that the researcher has control over the treatment given to the experimental groups, and also has control over or the ability to control for any confounding factors. The result of this control

### Table 1.1 Typical Hierarchy of Evidence for EBM

<table>
<thead>
<tr>
<th>Level of Evidence</th>
<th>Type of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (High)</td>
<td>Systematic reviews or meta-analysis of randomized controlled trials (RCTs) OR Evidence-based clinical practice guidelines based on systematic reviews of RCTs</td>
</tr>
<tr>
<td>2</td>
<td>At least one well-designed RCT</td>
</tr>
<tr>
<td>3</td>
<td>Well-designed quasi-experimental studies</td>
</tr>
<tr>
<td>4</td>
<td>Well-designed case control and cohort studies</td>
</tr>
<tr>
<td>5</td>
<td>Systematic reviews of descriptive and qualitative studies</td>
</tr>
<tr>
<td>6</td>
<td>A single descriptive or qualitative study</td>
</tr>
<tr>
<td>7 (Low)</td>
<td>The opinion of authorities and/or expert committees</td>
</tr>
</tbody>
</table>

*Source: Adapted from Melnyk and Fineout-Overholt (2005, p. 10)*
is the ability, arguably above all other designs, to infer causation—and therefore efficacy—from the differences in outcomes between the treatment and control groups.

RCTs, and other purely experimental designs, are often criticized for having poor ecological validity, where the conditions of the experiment do not match the conditions or populations in which the treatment might be delivered in the real world. Low ecological validity can, although does not necessarily, lead to low external validity where the findings of the study are not generalizable to other situations, including the real world. Another issue with RCTs is that they are relatively resource intensive and this leads to increased pressure to get desired results or to suppress any results that do not fit the desired outcome (Gluud, 2006; Simes, 1986).

In social sciences, RCTs are not always possible or advisable, so many systematic reviews and meta-analyses in these areas have less restrictive inclusion criteria. Commonly, these include studies that compare a treated clinical group with an untreated or attention control group (Kazdin & Weisz, 1998; Rosenthal, 1984). In these situations, a quasi-experimental design is often adopted. There is also evidence that RCTs are not necessarily any more accurate at estimating the effect of a given treatment than quasi-experimental and observational designs, such as cohort and case control studies (Concato, Shah, & Horwitz, 2008).

Systematic reviews provide a summary of evidence from studies on a particular topic. For EBM this typically involves the formation of an expert committee or panel, followed by the systematic identification, appraisal, and synthesis of evidence from relevant RCTs relating to the topic (Mehnky & Fineout-Overholt, 2005). The result of the review is usually some recommendation around the level of empirical support from RCTs for a given diagnostic tool or treatment. The RCT, therefore, is seen as the gold standard of evidence in EBM.

Systematic review and meta-analysis are closely related EBP methods to evaluate evidence from a body of sometimes contradictory research to assess treatment quality. There is a great deal of overlap between methods, stemming in part from their different origins. The systematic review arises from EBM and the early work of Cochrane and more recently Sackett and colleagues (2000). Meta-analysis originated in the 1970s, with the contributions of Glass (1976) in education and Rosenthal (1984) in psychology being central. Meta-analyses are a particular type of systematic review. In a meta-analysis, measures of the size of treatment effects are obtained from individual studies. The effect sizes from multiple studies are then combined using a variety of techniques to provide a measure of the overall effect of the treatment across all of the participants in all of the studies included in the analysis.

Both approaches use explicit methods to systematically search, critically appraise for quality and validity, and synthesize the literature on a given issue. Thus, a key aspect is the quality of the individual studies and the journals in which they appear. Searches ideally include unpublished reports as well as published reports to counteract the “file drawer” phenomenon: Published findings as a group may be less reliable than they seem because studies with statistically nonsignificant findings are less likely to be published (Rosenthal, 1984; Sackett et al., 2000).

The principal difference between systematic review and meta-analysis is the latter includes a statistical method for combining results of individual studies that produces a larger sample size, reduces random error, and has greater power to determine the true size of the intervention effect. Not only do these methods compensate for the limited power of individual studies resulting in a Type I error, the failure to detect an actual effect when one is present, they can also reconcile conflicting results (Rosenthal, 1984; Sackett et al., 2000).

The criticisms of meta-analysis mostly relate to the methodological decisions made during the process of conducting a meta-analysis,
often reducing the reliability of findings. This can result in meta-analyses of the same topic yielding different effect sizes (i.e., summary statistic), with these differences stemming from differences in the meta-analytic method and not just differences in study findings (Flather, Farkouh, Pogue, & Yusuf, 1997). Methodological decisions that can influence the reliability of the summary statistic produced from a meta-analysis include the coding system used to analyze studies, how inclusive the study selection process was, the outcome measures used or accepted, and the use of raw effect sizes or adjusted sample sizes (Flather et al., 1997). Some meta-analyses are conducted without using a rigorous systematic review approach, and this is much more likely to produce a mathematically valid but clinically invalid conclusion (Kazdin & Weisz, 1998; Rosenthal, 1984).

There are also some criticisms of the meta-analysis approach, specifically from child and adolescent psychology literature. Meta-analyses can obscure qualitative differences in treatment execution such as investigator allegiance (Chambless & Hollon, 1998; Kazdin & Weisz, 1998) and are limited by confounding among independent variables such as target problems, which tend to be more evident with certain treatment methods (Kazdin & Weisz, 1998).

Evidence-based clinical practice guidelines are also included in this top level of evidence, with the caveat that they must be primarily based on systematic reviews of RCTs. The purpose of these practice guidelines is to provide an easy-to-follow tool to assist clinicians in making decisions about the treatment that is most appropriate for their patients (Straus, Richardson, Glasziou, & Haynes, 2005).

There are some issues with the publication and use of evidence-based clinical practice guidelines. One problem is that different groups of experts can review the same data and arrive at different conclusions and recommendations (Hadorn, Baker, Hodges, & Hicks, 1996). Some guidelines are also criticized for not being translated into tools for everyday use. One of the major drawbacks of clinical guidelines is that they are often not updated frequently to consider new evidence (Lohr, Eleazer, & Mauskopf, 1998).

In addition to individual researchers and groups of researchers, there are a large number of organizations that conduct systematic reviews, publish clinical practice guidelines, and publish their findings in journals and in databases on the Internet, including the Cochrane Collaboration (www.cochrane.org/), the National Institute for Clinical Evidence (NICE; www.nice.org.uk/), the Joanna Briggs Institute (www.joannabriggs.edu.au/about/home.php), and the TRIP database (www.tripdatabase.com/index.html). For example, the Cochrane Collaboration, an organization named after Archie Cochrane, is an international network of researchers who conduct systematic reviews and meta-analyses and provide the results of these to the research, practice, and policy community via the Cochrane Library.

The overall strengths of systematic reviews, meta-analyses, and clinical practice guidelines relate partly to the nature of the methods used and partly to the nature of RCTs, which were discussed earlier. The ultimate goal of scientific research is to contribute toward a body of knowledge about a topic (Bowling, 1997). Systematic reviews and clinical practice guidelines are essentially trying to summarize the body of knowledge around a particular treatment, which is something that would otherwise take a clinician a very long time to do on their own, so it is easy to see the value in a process that does this in a rigorous and systematic way.

There are a number of overall weaknesses for this top level of evidence for EBM. For example, the reliability of findings from systematic reviews has been found to be sensitive to many factors, including the intercoder reliability procedures adopted by the reviewers (Yeaton & Wortman, 1993). It is possible for different people using the same coding system
to come up with different conclusions about the research evidence for a particular treatment. The selection of comparable studies is also a major issue, and especially in matching comparison groups and populations. The control and treatment groups in RCTs and other studies are usually not identical and this raises issues, particularly in the use of meta-analyses, where the treatment effects are being combined (Eysenck, 1994). The mixing of populations can also cause problems, where the treatment may only be effective in one or more specific populations but the review has included studies from other populations where the treatment is not effective. The overall effect of mixing these populations is to diminish the apparent effectiveness of the treatment under review (Eysenck, 1994).

Other issues with this top level of evidence relate to the information not included. Once a review or guideline is completed it must be constantly updated to check for new evidence that may change the conclusions. There are also issues around the information not included in the review process, including non-RCT studies, gray literature, such as technical reports and commissioned reports, and unpublished studies. Studies that do not get into peer-reviewed journals, and therefore usually get excluded from the review process, are often those that did not have significant results. The effect of this is that reviews will often exaggerate the effectiveness of a treatment through the exclusion of studies where the treatment in question was found to be ineffective. For this reason there has been a recent move to introduce what is called the “failsafe N” statistic. This is the hypothetical number of unpublished (or hidden) studies showing, on average, no effect that would be required to overturn the statistically significant effects found from review of the published (or located) studies results (Becker, 2006).

Quasi-experimental designs are the same as RCTs but the groups are not randomly assigned, there is no control group, or they lack one or more of the other characteristics of an RCT. In many situations randomization of participants or having a control group is not practically and/or ethically possible. The strengths of quasi-experimental designs come from the degree of control that the research has over the groups in the study, and over possible confounding variables. A well-designed quasi-experimental design has the important characteristic of being able to infer some degree of causation from differences in outcomes between study groups.

Case control studies identify a population with the outcome of interest (cases) and a population without that outcome (controls), then collects retrospective data to try to determine their relative exposure to factors of interest (Grimes & Schulz, 2002). There are numerous strengths of case control studies. They have more ecological validity than experimental studies and they are good for health conditions that are very uncommon (Grimes & Schulz, 2002). There is a relatively clear temporal sequence, compared to lower level evidence, which allows some degree of causality to be inferred (Grimes & Schulz, 2002). They are also relatively quick to do, relatively inexpensive, and can look at multiple potential causes at one time (Grimes & Schulz, 2002). The weaknesses of case control studies include the inability to control potentially confounding variables except through statistical manipulations and the reliance on participants to recall information or retrospectively collating information from existing data; the choice of control participants is also difficult (Grimes & Schulz, 2002; Wacholder, McLaughlin, Silverman, & Mandel, 1992). All of this often leads to a lot of difficulty in correctly interpreting the results of case control studies (Grimes & Schulz, 2002).

Cohort studies are longitudinal studies where groups are divided in terms of whether they receive or do not receive a treatment or exposure of interest, and are followed over time to assess the outcomes of interest (Roberts & Yeager, 2006). The strengths of cohort studies include the relatively clear
temporal sequence that can be established between the introduction of an intervention and any subsequent changes in outcome variables, making the establishment of causation possible, at least to some degree. The limitations of cohort studies include the difficulty in controlling extraneous variables, leading to a relatively limited ability to infer causality. They are also extremely resource intensive and are not good where there is a large gap between treatment and outcome (Grimes & Schulz, 2002).

Descriptive studies involve the description of data obtained relating to the characteristics of phenomena or variables of interest in a particular sample from a population of interest (Melnyk & Fineout-Overholt, 2005). Correlational studies are descriptive studies where the relationship between variables is explored. Qualitative studies collect nonnumeric data, such as interviews and focus groups, with the analysis usually involving some attempt at describing or summarizing certain phenomena in a sample from a population of interest (Melnyk & Fineout-Overholt, 2005). Descriptive studies in general sit low down on the ladder of evidence quality due to their lack of control by the researcher and therefore their inability to infer causation between treatment and effect. On the other hand, much research is only feasible to conduct using this methodology.

Qualitative research in health services mostly arose from a desire to get a deeper understanding of what the quantitative research finding meant for the patient and provider (Mays & Pope, 1996). It asks questions such as “How do patients perceive . . . ?” and “How do patients value the options that are offered?”

Expert opinion is material written by recognized authorities on a particular topic. Evidence from these sources has the least weight in EBM, although to many clinicians the views of experts in the field may hold more weight than the higher level evidence outlined above.

Despite its criticisms and methodological complexity, EBM is seen by most clinicians as a valid and novel way of reasoning and decision making. Its methods are widely disseminated through current clinical education programs at all levels throughout the world. As proposed by Gordon Guyatt and colleagues in 1992, EBM has led to a paradigm shift in clinical practice (Sackett et al., 2000).

CURRENT STATUS OF EBP
MOVEMENTS ACROSS HEALTH
AND EDUCATION PROFESSIONS

Evidence-based practice was initially in the domain of medicine, but now most human service disciplines subscribe to the principles of EBP. For example, EBP or the use of empirically supported treatment is recommended by the APA, Behavior Analyst Certification Board (BACB), American Psychiatric Association, National Association of Social Workers, and General Teaching Council for England. For U.S. education professionals, EBP has been mandated by law. The No Child Left Behind Act of 2001 (NCLB; U.S. Department of Education, 2002) was designed to make the states, school districts, principals, and teachers more answerable for the performances shown by the students for whom they were providing education services. Along with an increase in accountability, the NCLB requires “schoolwide reform and ensuring the access of children to effective, scientifically based instructional strategies and challenging academic content” (p. 1440). The advent of the NCLB and the resulting move toward EBP occurred because, despite there being research conducted on effective and efficient teaching techniques (e.g., Project Follow Through: Bereiter & Kurland, 1981; Gersten, 1984) and a growing push for accountability from the public (Hess, 2006), this seldom was translated into actual practice. Education appears to have been particularly susceptible to implementing programs that were fads, based on little more than personal ideologies and good marketing. Much money and time has been lost by school districts adopting programs that have no
empirical support for their effectiveness, such as the adoption of known ineffective substance abuse prevention programs (Ringwalt et al., 2002) or programs that are harmful for some students and their families, such as facilitated communication (Jacobson, Mulick, & Schwartz, 1995). This leads not only to resource wastage, but an even greater cost in lost opportunities for the students involved.

It seems like common sense for EBP to be adopted by other disciplines, as it is difficult to understand why reputable practitioners, service providers, and organizations would not want to provide interventions that have been shown to be the most effective and efficient. Yet, despite this commonsense feeling, the codes of conduct and ethical requirements, and mandated laws, many disciplines (including medicine) have found it difficult to bridge the gap between traditional knowledge-based practice and the EBP framework (Greenhalgh, 2001; Stout & Hayes, 2004). Professions such as social work (Roberts & Yeager, 2006; Zlotnick & Solt, 2006), speech language therapy (Enderby & Emerson, 1995, 1996), occupational therapy (Bennett & Bennett, 2000), and education (Shavelson & Towne, 2002; Thomas & Pring, 2004) have all attempted to overcome the difficulties that have arisen as they try to move EBP from a theoretical concept to an actual usable tool for everyday practitioners. Although some of these disciplines have unique challenges to face, many are common to all.

Many human services disciplines have reported that one of the major barriers to the implementation of EBP is the lack of sound research. For example, speech language therapy reports difficulties with regard to the quality and dissemination of research. A review of the status of speech language therapy literature by Enderby and Emerson (1995, 1996) found that there was insufficient quality research available in most areas of speech language therapy; however, they did find that those areas of speech therapy that were associated with the medical profession (e.g., dysphasia and cleft palate) were more likely to have research conducted than those associated with education (e.g., children with speech and language disorders and populations with learning disabilities). There continues to be a lack of agreement on speech language therapies effectiveness (Almost & Rosenbaum, 1998; Glogowska, Roulstone, Enderby, & Peters, 2000; Robertson & Weismer, 1999).

In addition to the lack of research, Enderby and Emerson were also concerned about the manner in which health resources were being allocated for speech language therapists. They found that of the resources allocated to speech language therapy approximately 70% were being used with children with language impairments, despite there being limited quality evidence of the effectiveness of speech language therapy with the population at that time. They also identified that dysarthria was the most commonly acquired disorder, but had very little research outside of the Parkinson’s disease population. So, again, many resources have been allocated to programs that have no evidence of effectiveness. This questionable allocation of resources is seen in other fields also. For example, a number of studies have found that people seeking mental health treatments are unlikely to receive an intervention that would be classified as EBP and many will receive interventions that are ineffective (Addis & Krasnow, 2000; Goisman, Warshaw, & Keller, 1999).

A number of organizations and government agencies within a variety of fields try to facilitate the much-needed research. Examples in education include the United States Department for Children, Schools, and Families (previously Department for Education and Skills), the National Research Council (United Kingdom), and National Teacher Research Panel (United Kingdom). Similarly, in social work the National Association of Social Workers (United States) and the Society for Social Work and Research (United States) both facilitate the gathering of evidence to support the use of EBP within their field;
however, even if these organizations generate and gather research demonstrating the effectiveness of interventions, this does not always translate into the implementation of EBP. As mentioned previously, resource allocation does not necessarily go to those treatments with evidence to support them due to a lack of effective dissemination on the relevant topics. Despite medicine being the profession with the longest history of EBP, Guyatt et al. (2000) stated that many clinicians did not want to use original research or they fail to do so because of time constraints and/or lack of understanding on how to interpret the information. Rosen, Proctor, Morrow-Howell, and Staudt (1995) found that social workers also fail to consider empirical research when making research decisions. They found that less than 1% of the practice decisions were justified by research.

Although the lack of research-based decision making is concerning, many professions are attempting to disseminate information in a manner that is more user friendly to its practitioners. Practice or treatment guidelines have been created to disseminate research in a manner that will facilitate EBP. These guidelines draw on the empirical evidence and expert opinion to provide specific best-practice recommendations on what interventions or practices are the most effective/efficient for specific populations (Stout & Hayes, 2004). There are a number of guideline clearinghouses (National Guidelines Clearinghouse [www.guideline.gov/] and What Works Clearinghouse [ies.ed.gov/ncee/wwc/]). Associations and organizations may also offer practice guidelines (e.g., American Psychiatric Association).

Although there appears to have been emphasis placed on EBP, there is still much work to be done in most human service fields, including medicine, to ensure that there is appropriate research being conducted, that this research is disseminated to the appropriate people, and that the practitioners then put it into practice. Reilly, Oates, and Douglas (2004) outlined a number of areas that the National Health Service Research and Development Center for Evidence-Based Medicine had identified for future development. They suggest that there is a need for a better understanding of how practitioners seek information to inform their decisions, what factors influence the inclusion of this evidence into their practice, and the value placed, both by patients and practitioners, on EBP. In addition, there is a need to develop information systems that facilitate the integration of evidence into the decision-making processes for practitioners and patients. They also suggest that there is a need to provide effective and efficient training for frontline professionals in evidence-based patient care. Finally, they suggest that there is simply a need for more research.

EVIDENCE IN PSYCHOLOGY

Although RCT research is still held up as the gold standard of evidence in medical science, in other clinical sciences, especially psychology, best research evidence has been less focused on RCTs as being the only scientifically valid approach. Randomized controlled trials are still held in high regard in clinical psychology, but it has long been recognized that alternative designs may be preferred and still provide strong evidence, depending on the type of intervention, population studied, and patient characteristics (APA Presidential Task Force on Evidence-Based Practice, 2006; Chambless & Hollon, 1998).

Research Methods Contributing to Evidence-Based Practice in Psychology

We have described and discussed RCTs in the previous section on EBM. The issues concerning RCTs and their contribution to the clinical psychology evidence base are similar. The inclusion of evidence of treatment efficacy and utility from research paradigms other than RCT methods and approximations thereto has
been recommended for clinical psychology since the EBPP movement commenced (Chambless & Hollon, 1998). The APA Presidential Task Force on Evidence-Based Practice (2006) findings support the use of qualitative studies, single-case or small-N experimental designs, and process-outcome studies or evaluations, in addition to the RCT and systematic review techniques employed in EBM. A more in-depth discussion of some of these methods will be provided.

Psychologists and members of other professions that are more influenced by social science research methods than medicine may consider that findings from qualitative research add to their EBP database. Although qualitative research is not accepted as legitimate scientific study by many psychologists, and not usually taught to clinical psychologists in training, it may well have its place in assessing variables regarding clinical expertise and client preferences (Kazdin, 2008). Research questions addressable through qualitative research relevant to EBPP are similar to those mentioned for EBM.

Another important source of scientific evidence for psychologists, educators, and other nonmedical professionals is that derived from small-N research designs, also known as single-case, single-subject, or N = 1 designs. These alternative labels can confuse, especially since some single-case designs include more than one subject (e.g., multiple baseline across subjects usually include three or more participants). The most familiar small-N designs are ABAB, multiple baseline, alternating treatments, and changing criterion (Barlow & Hersen, 1984; Hayes, Barlow, & Nelson-Gray, 1999; Kazdin, 1982). Data from small-N studies have been included as sources of evidence, sometimes apparently equivalent to RCTs (Chambless & Hollon, 1998), or at a somewhat lower level of strength than RCTs (APA, 2002).

Strength of evidence from small-N designs. Small-N designs can be robust in terms of controlling threats to internal validity; however, external validity has often been viewed as problematic. This is because participants in small-N studies are not a randomly selected sample from the whole population of interest. The generality of findings from small-N studies is established by replication across more and more members of the population of interest in further small-N studies. A hypothetical example of the process of determining the generality of an intervention goes like this: (a) A researcher shows that a treatment works for a single individual with a particular type of diagnosis or problem; (b) Either the same or another researcher finds the same beneficial effects with three further individuals; (c) Another researcher reports the same findings with another small set of individuals; and so on. At some point, sufficient numbers of individuals have been successfully treated using the intervention that generality can be claimed. Within a field such as a particular treatment for a particular disorder, small-N studies can be designed so results can be pooled to contribute to an evidence base larger than N = 1 to 3 or 4 (Lord et al., 2005).

Even if there is an evidence base for an intervention from a series of small-N studies, every time another individual receives the same treatment, the clinician in scientist-practitioner role evaluates the effects of the intervention using small-N design methods. Thus, every new case is clinical research to determine the efficacy and effectiveness of this treatment for that individual.

Clinical psychologists can be cautioned that physicians and medical researchers have a similar-sounding name for an experimental design with their “N of 1 trials.” The definition of N of 1 trials can vary somewhat but typically they are described as “randomised, double blind multiple crossover comparisons of an active drug against placebo in a single patient” (Mahon, Laupacis, Donner, & Wood, 1996, p. 1069). These N of 1 trials are continued until the intervention, usually a drug, in question shows consistently better effects than its comparison treatment or control condition.
Then the intervention is continued, or discontinued if there were insufficient beneficial effects. The $N$ of 1 trials can be considered top of the hierarchy of evidence in judging strength of evidence for individual clinical decisions, higher even than systematic reviews of RCTs (Guyatt et al., 2000). This is because clinicians do not have to generalize findings of beneficial effects of the treatment to this patient from previous researched patients. Consistent beneficial findings from many $N$ of 1 trials add up to an “$N$ of many” RCT, thus providing evidence of the generality of findings. An $N$ of 1 trial is quite similar to the alternating treatments small-$N$ design employed in clinical psychology, especially applied behavior analysis; however, they are not identical and the differences may be overcome for some psychosocial interventions only with careful planning, but many interventions could not be assessed by fitting them into an $N$ of 1 trial format. Further details are beyond the scope of this chapter; however, close study of the methodological requirements of both $N$ of 1 designs and small-$N$ designs can show the similarities and differences (Barlow & Hersen, 1984; Guyatt et al., 2000; Hayes et al., 1999; Kazdin, 1982; Mahon et al., 1996).

Some state that small-$N$ designs may be most suitable for evaluating new treatments (Lord et al., 2005), others that single-case studies are most suitable for determining treatment effects for individuals (APA Presidential Task Force on Evidence-Based Practice, 2006). We do not disagree with either, except to point out that it has been argued that small-$N$ studies can contribute much more to EBPP than these two advantages. In the following section, we review how ESTs can be determined from an evidence base consisting entirely of small-$N$ studies.

Criteria for Assessing Efficacy

Lonigan, Elbert, and Johnson (1998) tabulated criteria for determining whether an intervention for childhood disorders could be considered well-established (i.e., efficacious) or probably efficacious (i.e., promising). For the former, they recommended at least two well-conducted RCT standard studies by independent research teams or a series of independent well-designed small-$N$ studies with at least nine participants carefully classified to the diagnostic category of interest showing that the intervention was better than alternative interventions. The availability of treatment manuals was recommended. For “promising” treatments, the criteria were relaxed to allow nonindependent RCTs, comparing treatment to no treatment, or a minimum of three small-$N$ studies. These criteria followed those established at the time for psychological therapies in general (Chambless & Hollon, 1998).

We have discussed the determination of empirical support from RCTs already. The next sections will examine how evidence is derived from small-$N$ studies: First, how evidence is assessed from individual research reports; and second, how the evidence can be combined from a group of research articles addressing the same topic of interest.

Evaluating Evidence From Small-$N$ Research Designs

How do those assessing the efficacy of a treatment from small-$N$ designs measure the strength of the design for the research purpose and whether, or to what extent, a beneficial effect has been demonstrated? Chambless and Hollon (1998) recommended reviewers to rate single-case studies on the stability of their baselines, use of acceptable experimental designs, such as ABAB or multiple baseline designs, and visually estimated effects. Baselines need not always be stable to provide an adequate control phase; there are other valid designs (e.g., changing criterion, multielement experimental designs); and visual estimates of effects are not necessarily reliable (Cooper, Heron, & Heward, 2007). Validity
and, probably, reliability of reviews using only Chambless and Hollon’s (1998) criteria could not be assured.

Attempts to improve reliability and validity of judgments about the value of small-N design studies have included establishing more well-defined criteria. Quite detailed methods for evaluating small-N studies have been published. For example, Kratochwill and Stoiber (2002) describe the basis of a method endorsed by the National Association of School Psychologists (NASP) designed for evaluating multiple research reports that used single-case designs for establishing evidence-based recommendations for interventions in educational settings. The system is more inclusive of design variations beyond those recommended by Chambless and Hollon (1998), and reviewers are instructed to code multiple variables, including calculating effects sizes from graphical data. The coding form extended over 28 pages. Shernoff, Kratochwill, and Stoiber (2002) illustrated the assessment procedure and reported that, following extensive training and familiarity with the coding manual, they achieved acceptable agreement among themselves. They stated that the process took 2 hours for a single study, although, so our own graduate students report, it takes much longer if multiple dependent and independent variables or hundreds of data points had been features of the research article with which they chose to illustrate the NASP procedure. The NASP method was constructively criticized by Levin (2002), and is apparently being revised and expanded further (Kratochwill, 2005).

Meanwhile, others have developed what appear to be even more labor intensive methods for attempting to evaluate objectively the strength of evidence from a series of small-N designs. As an example of a more detailed method, Campbell (2003) measured every data point shown on published graphs from 117 research articles on procedures to reduce problem behaviors among persons with autism. Each point was measured by using dividers to determine the distance between the point and zero on the vertical axis. Campbell calculated effect sizes for three variables: mean baseline reduction, percentage of zero data points, and percentage of nonoverlapping data points. Use of these statistical methods may have been appropriate considering the types of data Campbell examined; nonzero baselines of levels of problem behavior followed by intervention phases in which the researchers’ goal was to produce reduction to zero (see Jensen, Clark, Kircher, & Kristjansson, 2007, for critical review of meta-analytic tools for small-N designs). Nevertheless, the seemingly arduous nature of the task and the lack of generalizability of the computational methods to reviewing interventions designed to increase behaviors are likely to mitigate wider acceptance of Campbell’s (2003) method.

A final example of methods to evaluate the strength of an evidence base for interventions is that outlined by Wilczynski and Christian (2008). They describe the National Standards Project (NSP), which was designed to determine the benefits or lack thereof of a wide range of approaches for changing the behaviors of people with autism spectrum disorders (ASD) aged up to 21 years. Their methods of quantitative review enabled the evidence from group and small-N studies to be integrated. Briefly, and to describe their method for evaluating small-N studies only, their review rated research articles based on their scientific merit first. Articles were assessed to determine whether they were sufficiently well-designed in terms of experimental design, measurement of the dependent variables, assessment of treatment fidelity, the ability to detect generalization and maintenance effects, and the quality of the ASD classification of participants. If the article exceeded minimum criteria on scientific merit, the treatment effects were assessed as being beneficial, ineffective, adverse, or that the data were not sufficiently interpretable to decide on effects. Experienced trained reviewers were able to complete a review
for a research article in 1 hour or less with interreviewer agreement of 80% or more.

Inevitable problems for all these systems for review of single case designs arise from the necessity of creating “one size fits all” rules. For example, the evaluative methods of both the NASP (Kratochwill & Stoiber, 2002) and NSP projects (Wilczynski & Christian, 2008) include consideration of level of interobserver agreement for determining partly the quality of measurement of the dependent variables. The minimum acceptable level of agreement is specified at, say, 70% or 80%, which makes it relatively convenient for reviewers to determine from reading the published research article that they are rating; however, it has been known for more than 30 years that an agreement percentage is rather meaningless without examination of how behaviors were measured, what interobserver agreement algorithm was employed, and the relative frequency or duration of the behaviors measured (Hawkins & Dotson, 1975). Another example concerns coding rules for evaluating the adequacy of baseline measures of behavior. Whether the criterion for a baseline phase of highest scientific merit is a minimum of 3 points (NASP) or 5 points (NSP), there will be occasions when the rule should be inapplicable. For example, in Najdowski, Wallace, Ellsworth, MacAleeese, and Cleveland (2008), after more than 20 observational sessions during intervention showing zero severe problem behavior, a return to baseline in an ABAB design raised the rate of problem behavior to more than four per minute, which was higher than any points in the first A-phase. To continue with the baseline to meet NASP or NSP evaluative criteria would have been unnecessary to show experimental control and dangerous for the participant.

These examples indicate why more detailed and codified methods for reviewing small-N designs quantitatively are, as yet, not firmly established. Although there may be satisfactory methods for fitting RCTs’ scientific merit and size of effects into databases for meta-analyses, that appears to be more problematic with small-N designs, given their flexibility in use (Hayes et al., 1999).

Volume of Evidence From Small-N Studies Required to Claim That an Intervention Is Evidence Based

Having rated the strength of evidence from individual research articles, the next step is to determine whether a group of studies on the same topic between them constitute sufficient evidence to declare that an intervention is an empirically supported treatment or a promising or emerging intervention. We discuss only empirically supported treatment criteria here. Consensus on what minimum criterion should apply has yet to be reached. Chambless and Hollon (1998) originally recommended a minimum of two independent studies with three or more participants \(N \geq 3\) showing good effects for a total of \(N \geq 6\) participants. Lonigan et al. (1998) required three studies with \(N \geq 3\); that is, beneficial effects shown for \(N \geq 9\) participants. Since then, the bar has been raised. For instance, Horner et al. (2005) proposed that the criteria for determining that an intervention is evidence-based included a minimum of five small-N studies, from three or more separate research groups, with at least 20 participants in total. Wilczynski and Christian (2008) used similar criteria with \(N \geq 6\) studies of strongest scientific merit totaling \(N \geq 18\) participants with no conflicting results from other studies of adequate design. Others have recommended similar standards: Reichow, Volkmar, and Cicchetti (2008) described a method for evaluating research evidence from both group and small-N designs, as had Wilczynski and Christian (2008). Finally, Reichow et al. (2008) set the criterion for an established EBP at \(\geq 10\) small-N studies of at least “adequate report strength” across three different locations and three different research teams with a total of 30 participants, or, if at least five studies of “strong report strength” with a total of 15 or more participants existed, that could substitute for the 10-study criterion.
The rationale for selecting the numerical criteria is usually not stated. Thus, all we can state is that some systems (e.g., Reichow et al., 2008) are more conservative than others (e.g., Lonigan et al., 1998). Sometimes conservative criteria may be appropriate, for example, when the costs of treatment are high, and/or the intervention is exceedingly complex requiring highly skilled intervention agents, and/or the benefits for the intervention are less than ideal (i.e., it reduces problems to a more manageable level, but does not eliminate them), and/or some negative side effects have been observed. On the other hand, if relatively few resources are required to implement an effective and rapid intervention without unwanted side effects, fewer well-conducted studies may be needed to persuade consumers that the intervention is empirically supported, and therefore worth evaluating with the individual patient. This brief discussion should advise readers to examine and consider criteria carefully when reviewers claim that a particular intervention is an empirically supported treatment for a particular disorder or problem for a particular population.

The discussions on evidence from small-N designs have left much out. For instance, Reichow et al. (2008) and Wilczynski and Christian (2008) developed algorithms for assessing the strength of evidence at different levels, although we have outlined only the highest levels. Both groups have also reported algorithms for determining ESTs from mixed methods (e.g., RCTs and small-Ns). Wilczynski and Christian (2008) report rules for incorporating conflicting results into the decision-making process about overall strength of evidence (De Los Reyes & Kazdin, 2008).

Level of Specificity of Empirically Supported Treatments

The issue to be discussed next concerns the unit of analysis of the research evidence. We illustrate by examining an example provided by Horner et al. (2005) in which they assessed the level of support for functional communication training (FCT; Carr & Durand, 1985). Functional communication training is an approach to reducing problem behaviors that teaches an appropriate nonproblematic way for individuals to access reinforcers for the problem behavior that have been identified through functional assessment. Horner and colleagues cited eight published research reports that included 42 participants who had benefited in FCT studies across five research groups. The evidence was sufficient in quantity for it to be concluded that FCT is an empirically supported treatment, exceeding all criteria reviewed earlier except that two more studies would have been needed to reach the number of studies $\geq 10$ by the criteria of Reichow et al. (2008).

It might reasonably be asked: “For which population is FCT beneficial?” Perusal of the original papers cited by Horner et al. (2005) shows that 21/42 participants’ data were reported in one of the eight cited studies (Hagopian, Fisher, Sullivan, Acquisto, & LeBlanc, 1998), with the oldest participant being 16 years old, and none reported to have a diagnosis of autism. Thus, applying Horner et al.’s criteria, it cannot be concluded from the studies cited that FCT is an empirically supported treatment for participants with autism or for participants older than 16, regardless of diagnosis. As an aside, eight participants across the other seven studies were reported to have autism, and three participants in total were aged over 20 years; however, the literature on FCT that Horner et al. (2005) included did not appear to have been obtained from a systematic search, so it is possible that there has been sufficient research to show that FCT is an empirically supported treatment for subgroups, and perhaps autism and adults are two of those.

TREATMENT GUIDELINES

Treatment guidelines specifically recommend ESTs to practitioners and consumers. Alternative descriptors are clinical practice and
best practices guidelines (Barlow, Levitt, & Bufka, 1999). The view of APA was that guidelines are not intended to be mandatory, exhaustive, or definitive... and are not intended to take precedence over the judgment of psychologists. APA’s official approach to guidelines strongly emphasizes professional judgment in individual patient encounters and is therefore at variance with that of more ardent adherents to evidence-based practice. (Reed et al., 2002, p. 1042)

It is apparent that many health-care organizations, insurance companies, and states in the United States interpret the purpose of lists of ESTs and treatment guidelines differently (Gotham, 2006; Reed & Eisman, 2006). They can interpret guidelines as defining what treatments can be offered to patients and, via manualization, exactly how treatment is to be administered, by whom, and for how long. The requirement for manualization allows funders of treatment to specify a standard reimbursement for the treatment provider. Thus, the empirically supported treatment movement was embraced by governments and health-care companies as it was anticipated to be a major contributor to controlling escalating health-care costs.

Many practicing psychologists were less enthusiastic about the empirically supported treatment and EBPP movements (see contributions by clinicians in Goodheart et al., 2006; Norcross et al., 2005). General concerns included that requirements to use only ESTs restrict professionalism by reframing psychologists as technicians going by the book mechanically; restricting client choice to effective interventions that have been granted empirically supported treatment status higher than others, only because they, like drugs, are relatively easy to evaluate in the RCT format. Prescription of one-size-fits-all ESTs may further disadvantage minorities, and people with severe and multiple disorders for whom there is scant evidence available. There were also concerns that the acknowledged importance of clinical expertise, such as interpersonal skills to engage the client (child) and significant others (family) in a therapeutic relationship, would be ignored.

Contrary to the pronouncements from the APA (2002, 2006), guidelines have been interpreted or developed that “assume the force of law” in prescribing some interventions and proscribing others (Barlow et al., 1999, p. 155). Compulsion of psychologists in practice to follow treatment guidelines has been reported to occur in the United States by giving immunity from malpractice lawsuits to those who use only ESTs, and increasing the vulnerability of those who do not to litigation and increased professional indemnity insurance (Barlow et al., 1999; Reed et al., 2002). Some guidelines, especially those produced by agencies or companies employing psychologists, have been viewed as thinly veiled cost-cutting devices justified with a scientistic gloss.

Ethical Requirements

For many human service professional organizations, EBPP and ESTs have become an ethical requirement. The APA’s Ethical Principles of Psychologists and Code of Conduct document mentions the obligation to use some elements of EBPP; for example, “Psychologists’ work is based upon established scientific and professional knowledge of the discipline” (American Psychological Association, 2010, p. 5). Other professional groups appear to be more prescriptive with regard to EBPP. The BACB’s Code for Responsible Conduct, for example, recommends EBPP with statements such as, “Behavior analysts rely on scientifically and professionally derived knowledge when making scientific or professional judgments in human service provision” (Behavior Analyst Certification Board, 2004, p. 1). The BACB also require the use of ESTs:

a. The behavior analyst always has the responsibility to recommend scientifically...
supervised most effective treatment procedures. Effective treatment procedures have been validated as having both long-term and short-term benefits to clients and society. b. Clients have a right to effective treatment (i.e., based on the research literature and adapted to the individual client). c. Behavior analysts are responsible for review and appraisal of likely effects of all alternative treatments, including those provided by other disciplines and no intervention. (Behavior Analyst Certification Board, 2004, p. 4)

As shown by the previous examples, each of these organizations have variations in how they have included EBP and empirically supported treatment into their codes of conduct and ethical statements; however, they both include the basic tenets of EBP: combining the best research information with clinical knowledge and the preferences of the individuals involved.

**CHILDREN, ADOLESCENTS, AND EVIDENCE-BASED PRACTICE IN PSYCHOLOGY**

Despite there being ESTs for a number of disorders and problem behaviors manifesting in children and adolescents, there are more than 500 treatments in use with children (Kazdin, 2008), most of which are unresearched. Chambliss and Ollendick (2001) listed 108 ESTs for adults, compared with 37 for children, suggesting that research on therapies for young people is lacking relative to adults. Further research with child and adolescent populations has been prioritized by APA (APA Presidential Task Force on Evidence-Based Practice, 2006). Researching the effects of treatments for children brings special difficulties (Kazdin & Weisz, 1998). Regarding practice, children do not typically self-refer for mental, psychological, or behavioral disorders, nor are they active seekers of ESTs or preferred interventions. Parents or guardians tend to take those roles, either independently or following recommendations from family, or health or education professionals. Children and youth cannot legally provide informed consent for treatments or for participation in research studies. These are among the developmental, ethical, and legal factors that affect consideration of EBPP with children.

Lord et al. (2005) discussed the challenges of acquiring rigorous evidence regarding efficacy of treatments for children with complex, potentially chronic behavioral/psychological disorders. Contributors to the article were researchers from multiple disciplines assembled by the National Institutes of Health in 2002. They wrote about autism spectrum disorders specifically, but acknowledged that the issues may have relevance to other child and youth problems.

Parents may be unwilling to consent to randomization studies in case their child is assigned to what parents perceive to be a less preferred treatment alternative, particularly when the intervention is long term and early intervention is, or is perceived to be, critical, such as early intensive behavioral intervention for pervasive developmental disorders. Lord et al. (2005) noted that ethical concerns may prohibit RCTs of promising interventions when randomization to no treatment or treatment of unknown effects is required by the evaluation protocol. Additional factors that reduce the internal validity of group comparison studies of psychosocial interventions include that parental blindness to the intervention allocated to their children is nigh on impossible, diffusion of treatment through parent support groups is likely, parents may choose to withdraw their children from no treatment or treatment as usual groups and obtain the experimental intervention or an approximation to it from outside the study, and children with severe disorders will often be subjected to multiple interventions of unknown benefit, provided with varying...
fidelity that may interact with one another to produce uninterpretable beneficial, neutral, or adverse effects (Smith & Antolovich, 2000).

Parents seek out information on their child’s disorder and intervention recommendations through Internet or parent organizations, sometimes accepting advice from professionals. Mackintosh, Meyers, and Goin-Kochel (2006) received 498 responses to a Web-based survey of parents with children with autism spectrum disorders and found that the most oft-cited sources of information were books (88%), Web pages (86%), other parents (72%), and autism newsletters (69%). Lagging somewhat as sources of advice were professionals other than educators or physicians (57%). Physicians, education professionals, and family members were cited as sources of information by fewer than half the parents who responded to the survey.

Multiple fad treatments have been recommended for childhood onset disorders, and their adoption by families and some professionals, including psychologists, wastes resources and time that could have been spent profitably by employing ESTs (Jacobson, Foxx, & Mulick, 2005). Although some of our examples of supported and unsupported treatments have related to children with autism spectrum disorders (Romanczyk, Arinstein, Soorya, & Gillis, 2003), the problem of treatment selection uninformed by research affects children with other difficulties and their families. Some interventions for children with developmental and other disabilities have been found ineffective or harmful (Jacobson et al., 2005), and the same occurs for children with ADHD (Waschbusch & Hill, 2003). (See also Chapter 2 of this volume for a further discussion of this point by Waschbusch, Fabiano, and Pelham.) We believe that clinical psychologists working with young people ought to have a professional ethical obligation to inform themselves and others about empirically unsupportable treatments as well as ESTs for their clients.

LIMITATIONS OF THE EVIDENCE BASE REGARDING EVIDENCE-BASED PRACTICE IN PSYCHOLOGY

There is a relatively large body of evidence concerning efficacy of treatments (Kazdin & Weisz, 1998, 2003), but far less on treatment utility, effectiveness, and efficiency. There is evidence that the utility of some efficacious treatments has been demonstrated, but further study is needed before the general statements can be made about the similarity or difference between outcomes from controlled research and clinical practice (Barlow et al., 1999; Hunsley, 2007).

There is less evidence concerning dimensions of clinical expertise and client characteristics, culture, and preferences that are relevant to beneficial treatment outcomes (Kazdin, 2008). Employment of qualitative research methods may help us to understand clients’ experiences of psychological treatments. The APA Task Force has suggested that clinical expertise is made up of at least eight components, including assessment and treatment planning, delivery, interpersonal skills, self-reflection, scientific skills in evaluating research, awareness of individual and social factors, the ability to seek additional resources where necessary, and having a convincing rationale for treatment strategies (APA Presidential Task Force on Evidence-Based Practice, 2006). Qualitative methods may provide evidence regarding clinical expertise also. Improvement of two-way communication between psychologists who are primarily researchers and those who identify more as practitioners would assist dissemination of ESTs, collaboration in clinical utility studies of efficacious treatments, and facilitate research into clinical expertise and barriers to the adoption of ESTs by clinicians (Kazdin, 2008).

Lilienfeld (2005) and McLennan, Wathen, MacMillan, and Lavis (2006) recommended further research on interventions in child
psychopathology on two fronts: (1) increased research on promising or new but plausible interventions; and (2) on research to combat questionable, but potentially harmful, interventions. Attention is needed to increase training of clinical psychologists in ESTs and cessation of training to use treatments that have been shown to be harmful, ineffective, or less effective. Decreasing the demand and use of treatments that have not been evaluated scientifically or have been found ineffective or harmful may be another strategy for helping clinical psychologists orient more to treatments that do work (i.e., ESTs).

Woody, Weisz, and McLean (2005) report that APA accredited clinical psychologist training programs taught and supervised interns in fewer ESTs in 2003 than they had in 1993. The training of clinical psychologists should include sufficient study of research methodology so that career-long learning, and contribution to research in practice concerning ESTs can be enhanced (Bauer, 2007; Kazdin, 2008). Considering the APA definition of EBPP, trainees need supervised practice in incorporating clients’ preferences, values, and cultural considerations as well as to develop clinical expertise (Collins, Leffingwell, & Belar, 2007). Further training for university-based psychologist training faculty should be studied.

Psychologists in practice may wish or feel forced to adapt to employing EBPP, but willingness or compulsion to do so is not the same as becoming immediately competent to use an empirically supported treatment effectively. The typical workshop format for introducing new techniques is as ineffective for professionals (Gotham, 2006), as it is for direct care staff (Reid, 2004). Skill-based training can be effective when trainees practice an intervention method in the natural clinical environment with differential feedback from the trainer on their performance of the skill. This should occur after workshops that introduce the rationale and method, and include in vivo or videotaped demonstration of the skill by the trainer (i.e., modeling). Frequent follow-up observations by the trainer, again with objective feedback to the trainee, can facilitate maintenance of the newly acquired skills (Reid, 2004). Gotham (2006) identified barriers to implementation of EBPP, and provided an example of how to implement an empirically supported treatment statewide despite obstacles. McCabe (2004) wrote quite optimistically for clinicians about the challenges of EBPP, offering advice in a step-by-step form to psychologists. Reorienting and training clinicians is an area for further clinician-researcher collaboration that requires emphasis.

We have not included discussion of an area of EBPP that has, to date, received less attention than evidence-based treatment, which is “evidence-based assessment” (Kazdin, 2005; Mash & Hunsley, 2005). An initial assessment with clinical utility will identify what is the disorder or problem behavior so that it points the psychologist in the right direction for identifying the range of available ESTs for this client. Evidence-based assessment includes identifying reliable and valid ongoing measures that show the effects of an intervention with the individual client (Kazdin, 2005). Typically, empirically supported treatment reviews identify treatments for DSM-type classifications, the nosological approach; however, an idiographic functional approach to assessment may lead to better problem-EST match (Sturmey, 2007).

It was mentioned earlier that some professional groups are concerned at the lack of research on the outcomes of their assessment and treatment methods. Because clinical and other psychologists have extensive training in research methods, we can assist other professions to assess the evidence for their intervention methods. Interdisciplinary collaborations also may help elucidate the interaction effects of behavioral or psychological ESTs with interventions with presently unknown effects delivered by other professions.
CONCLUDING REMARKS

The issues concerning evidence-based clinical practice of psychology for children and adolescents are arguably even more complex than for adult clients. All readers would welcome the day when, for any referred psychological, behavioral, or mental problem, psychologists with clinical expertise were able to offer a wide range of effective and safe interventions. Then the young person, with more or less help from those who care for them, could select the effective treatment that suited their culture, preferences, and resources. The literature reviewed for this chapter suggests that, generally speaking, the health care, including clinical psychology, and education professions are starting on the road to addressing the aspirational targets of the EBP movement. The lack of unanimous agreement within professions that the evidence-based practice movement is desirable and what constitutes evidence for beneficial interventions is unsurprising; however, with more evidence for safe, acceptable, and effective interventions such as contained in the present volume and more education for consumers, professionals, and the public, eventually the naysayers and peddlers of unsupported treatments may find they have no raison d’être (and also no income!).

REFERENCES


Concato, J., Shah, N., & Horwitz, R. (2008). Randomized, controlled trials, observational studies, and the
Overview and Foundational Issues

hierarchy of research designs. New England Journal of Medicine, 342, 1887–1892.


