CYBERPSYCHOLOGY & BEHAVIOR
Volume 5, Number 6, 2002

An Evaluation Model for Psychoeducational Interventions Using Interactive Multimedia

IVAN L. BEALE, Ph.D.

ABSTRACT

A review of published evaluations of psychoeducational interventions using interactive digital multimedia shows that evaluations typically address only some of the areas that collectively would constitute a comprehensive evaluation. There appears to be a need for an accepted standard for these evaluations, based on a general evaluation model that encompasses all relevant aspects of development, efficacy and use of multimedia interventions. A comprehensive model is proposed which includes special features of multimedia interventions that lead to unique evaluation requirements. The model integrates relevant aspects of program evaluation and clinical trials models in order to provide a unique model that includes all the evaluation domains relevant to digital multimedia interventions. These include evaluation of intervention theory, intervention design strategies, the formative process, intervention efficacy (process and outcome) and contextual issues such as evaluability assessment, reporting and stakeholder issues. The application of individual components of the model is illustrated with reference to problems in the evaluation literature on a particular type of digital multimedia intervention, electroencephalographic biofeedback for Attention-Deficit/Hyperactivity Disorder. The model should be useful for researchers planning evaluations of digital multimedia interventions, especially in the psychoeducational domain. This paper provides a theoretical and evidential background for the evaluation model, and includes a checklist and flowchart for the planning and conduct of the evaluation.

INTRODUCTION

"PSYCHOSOCIAL INTERVENTION" is a term applied to treatments that target psychological variables such as behaviors, cognitions, and emotions, within a social context. "Psychoeducational intervention" is a term applied to a subset of these interventions in which some educative process is involved. Examples include many types of individual applications of behavior modification, cognitive behavior therapy, and other psychotherapeutic models in which treatment involves learning of new skills or knowledge. Less-individualized or intrusive examples include group- or community-oriented health education programs.

In recent decades, computers have played an increasingly prominent role in the delivery of psychoeducational interventions ranging from health advice websites[1] to individualized treatment of anxiety disorders using virtual reality.[2] Specific applications of digital technology for psychological treatment have a long history, especially in the cybernetic domain of biofeedback therapy,[3] but there has now begun a proliferation of new applications of digital

HopeLab, Palo Alto, California.

interactive multimedia (IM) in psychoeducational interventions. The enthusiasm for these new developments will need to be balanced by careful appraisal of their theoretical foundations and efficacy as treatments, especially in view of the likely involvement of commercial interests and the potential for spurious claims. Whereas in the case of traditional forms of therapy, professionals could act as gatekeepers to advise the lay public about a particular treatment option, many digital-based interventions are freely on offer for direct purchase by the public for use in an unsupervised environment. Although claims of efficacy are commonly made in the advertising of such products, evidence from adequate evaluation research may be lacking. For example, few of the educational IM game products available to the public in retail stores and the Internet appear to have been subjected to adequate peer-reviewed evaluation studies, either prior to their release or subsequently.

Evaluation models for psycheducational interventions

How should digital IM interventions be evaluated? It is argued here that they are sufficiently different from more traditional types of intervention that models which have previously served for the evaluation of interventions do not fit well the requirements for evaluation of digital IM. There are two classes of evaluation model that typically have been used with psychoeducational interventions: program evaluation and clinical trials. Of the two, program evaluation is the more inclusive. It refers to the evaluation model generally used for interventions that are broadly applied to a large target population. Examples are government mental health initiatives such as early intervention programs or school-based suicide education programs. Evaluation measures frequently encompass a broad range, including perceptions of different groups of stakeholders, uptake of newly created services by targeted populations, and long-term societal outcomes such as rates of suicides or adjudicated delinquency. Sometimes case studies may be included to assess intervention outcomes and processes for individuals. Causal processes usually are implied rather than observed, and there may be little attempt to identify or measure moderating variables that may influence the impact of the intervention on individuals. Especially with large community-based programs, cost-benefit analysis is a major focus.[4,5]

The clinical trials model is used to evaluate either the efficacy or effectiveness of some treatment agent in relation to some specific disease endpoint. Clinical trials typically evaluate interventions (treatments) carried out with groups of individuals having some particular medical or behavioral disorder. Usually, the intervention is delivered directly to the individual or another person responsible for that individual (e.g., parent), so uptake of the intervention is usually not an issue. Stakeholder perceptions of the effectiveness or acceptability of the intervention are not often considered, and cost-benefit analysis is rarely included. The focus is on efficacy of the treatment for individuals and the discovery of moderating variables. True experimental and quasi-experiment designs typically are used to rule out, as far as possible, the influence of extraneous variables and allow clear causal inferences to be made.[6,7–10] Clinical-trials methodology was developed primarily to serve the need to evaluate medical/pharmacological interventions, but it has developed in parallel with evaluation research methods in psychology,[11,12] and the principles and practices developed in these two streams have much in common.

Evaluations of psychoeducational interventions tend to follow the program evaluation model if the intervention is something being applied to a community rather than to individuals in a clinical context, whereas the clinical trials model is more typically used for evaluations of interventions applied in the clinical context. However, there are many exceptions to this. For example, in the discipline of applied behavior analysis, interventions applied to communities often follow a model closer to clinical trials than to program evaluation.[5,11–13] In general, which intervention model is followed seems to depend more than anything on the background of the researchers, the theoretical context of the study, and what the re-

searchers consider the main evaluation issues to be. One result of this ad hoc approach is that many evaluations are too narrowly focused to give a picture of the value of an intervention that is broad enough to be generally useful. This is especially true of digital IM interventions, because they raise a wide range of evaluation issues.

Psychoeducational interventions based on digital interactive multimedia

Digital IM interventions are like traditional clinical psychology interventions in that the individual client interacts with an expert system or therapist that responds to each client uniquely. But digital interventions are also like social programs in the sense that the client typically experiences the intervention remotely from the professional who is prescribing it. Just one consequence of this remote usage is the potentially wide variability in the implementation of the intervention with different clients.[14]

The use of digital media introduces a number of special evaluation issues. It is not a straightforward matter to evaluate the "fit" between the theory on which an intervention is based, and the design and actual program code of the intervention software. Traditional psychological interventions have been relatively transparent in this respect. What the therapist says or does in the intervention can be measured and analyzed to determine how well it fits the treatment prescription and underlying theory. In the case of digital media, on the other hand, a lot of what goes on is not transparent to an observer. An evaluator might read about the theory on which the intervention supposedly is based, and even how this theory is translated into the design of the computer program. But is the programming actually faithful to the design, or is there a gap between the specified and actual functions on the program?

A unique advantage of digital media is the ability to record information both about program performance and the interactions with users. This information potentially can resolve the transparency problem outlined above. Provided that data are collected on all the relevant variables, evaluators can review the record for evidence both that the program behaves as intended and also achieves therapeutic goals. Moreover, if specific program elements are designed to change specific client behaviors, relevant records can provide direct evidence of the effectiveness of these elements as therapeutic components. Remarkably, published evaluations of interventions using digital media have not made use of this facility, relying instead on indirect and less-reliable measures of therapeutic effect such as client perceptions and impressions of family members as reported on questionnaires.[15,16]

Common limitations of evaluations of digital interactive multimedia interventions

An informal review of published evaluations of IM interventions shows wide variability in the scope of evaluations, that is, in the number of relevant aspects of the intervention that are evaluated. In some cases, for example in pilot studies, the scope is narrow because the evaluation is intentionally confined to a particular aspect of the intervention, such as the formative process. More often, however, evaluations purporting to test the overall efficacy of an intervention with a particular target group or behaviour contain limitations of scope that seriously compromise the external validity of the evaluation. These limitations include: failure to define adequately the participant inclusion criteria; failure to describe and evaluate the different functions of different intervention components; failure to describe how intervention components are derived from underpinning theory; failure to record and report the usage pattern for each participant; failure to report systematic data on credibility and acceptability of the intervention; failure to collect or report data on interactions between the participant and the intervention; failure to control for misattribution of cause (type III error); failure to address procedural integrity issues; failure to analyze the quality of instructional design components of the intervention; and failure to analyze the cognitive science applications in the user interface designs.

Not all these limitations are applicable to every intervention evaluation, but there are

many published evaluations that would be considerably improved if some of these limitations were addressed. There would seem to be value in setting out a general evaluation model for digital IM interventions that includes all the components that could be considered for inclusion in an evaluation study. This would provide a framework for planning an evaluation. Researchers would consider the need for each component and select those that were indicated by the particular nature of the intervention and the focus of the study. The value of the model would derive from it providing a comprehensive framework and systematic process, which would reduce the likelihood that useful components of the evaluation would be overlooked in the planning process.

The proposed evaluation model is specifically designed for applications of digital IM to psychoeducational interventions. The model combines appropriate components both of program evaluation and clinical trials methodologies. Components are introduced that are specific to digital technology, with the purpose of addressing the putative limitations identified in previous evaluation research on interventions using this technology.

An illustrative digital interactive multimedia intervention: electroencephalographic biofeedback

The value of the proposed model is illustrated by reference to the literature on a widely used psychoeducational intervention, electroencephalographic (EEG) biofeedback for Attention-Deficit/Hyperactivity Disorder (ADHD). This evaluation literature consists mainly of studies of digital IM intervention procedures which use biofeedback training ostensibly to normalize certain frequency ranges of the EEG. Training differentially reinforces the production of 12–15 or 15–18 Hz beta rhythms over 4–8 Hz theta rhythms. Participants, usually children diagnosed with ADHD or a related disorder, are trained at a computer which records EEG data from scalp electrodes then represents certain aspects of these data in a multimedia display. Biofeedback training protocols are used to teach participants to modify the multimedia display by learning cognitive/behavioral strategies that result in changes in relevant parts of the EEG spectrum.[17,18]

The evaluation literature on this particular intervention has many features that suit it to illustrating the issues addressed in this paper. Compared to other digital IM interventions, it has been in use for a long period and has been subjected to many evaluations and even a few reviews of evaluations. Over that period, its application has been progressively broadened to new target populations and disorders. There has been enduring controversy about both the underlying theory and the efficacy of this intervention. Its mode of delivery has moved progressively from contexts providing expert supervision by the intervention developers, through clinical contexts involving non-expert professional supervision, to non-supervised use in private homes (www.eegspectrum.com). New versions are proliferating, including some which purportedly make good use of recent advances in multimedia games technology.[19] Finally, despite continuing uncertainty about its efficacy, EEG biofeedback has attracted wide media publicity as an alternative to well-established behavioral or pharmacological interventions.[20]

COMPONENTS OF THE EVALUATION MODEL

The components of the proposed model are summarized in Figure 1, where they are organized according to their area of focus (columns) and a possible sequence of action (arrows). The rationale for each component is described briefly below, and where relevant, the function of each is illustrated by reference to the evaluation literature on EEG biofeedback for ADHD. The critical questions to be addressed in each component are listed in Table 1.

A. Intervention structure and formative evaluation.

A1. The political context: stakeholder analysis. It is not universally accepted that interventions should be subject to rigorous evaluation. For commercial developers, market acceptance may sometimes have priority over product integrity, so that it may be considered

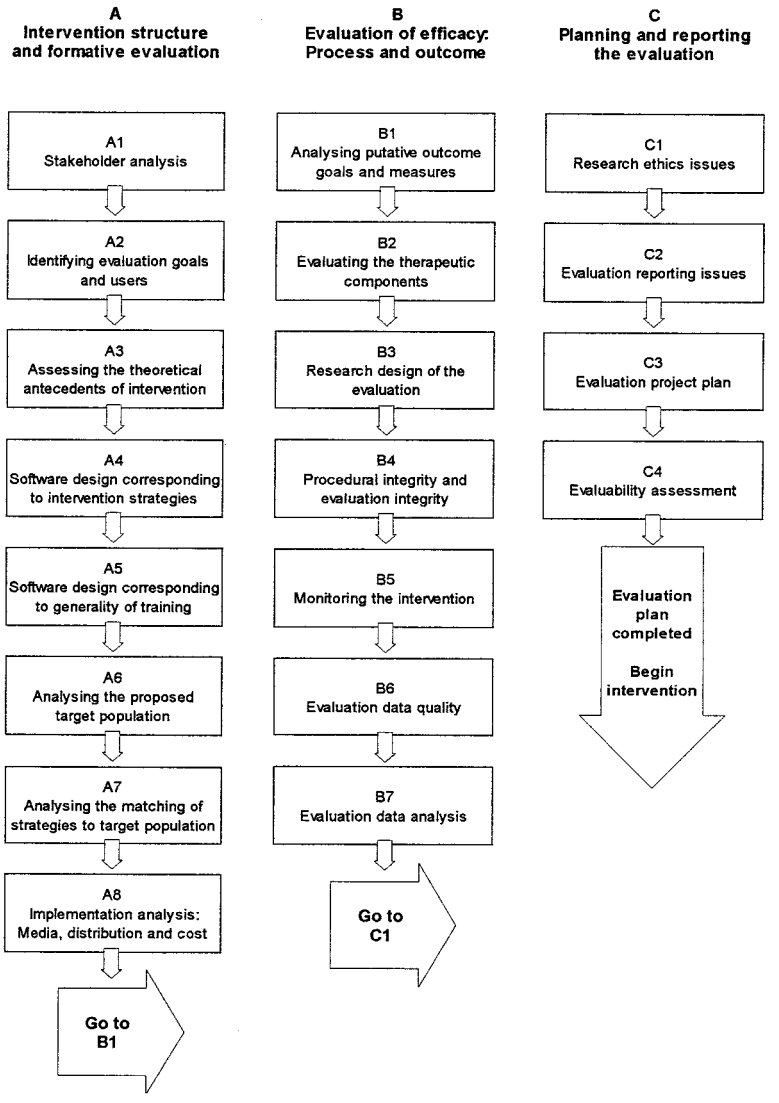

FIG. 1. Components of the evaluation model, showing domains of focus (columns) and possible sequence of action (arrows).

more important that the product sells than whether independent evaluations have *shown* that the product works. In such a context, "evaluation" may simply mean having the product "reviewed" by public relations companies or software and health magazine editorial staff, or seeking endorsements from credible user organizations or "independent experts."

Similar issues may apply in the professional world also, since professional groups and individuals may well have a financial or credibility stake in the efficacy of a product they have designed or published. The profession of psychology has hosted some notable disputes about the effectiveness of various interventions, of which the continuing controversy about the relative merits of alternative therapies to reduce behavioural problems in children with ADHD, including stimulant medication and EEG biofeedback, is a notable example. Participant stakeholders include medical practitioners, psychologists, teachers, pharmaceutical companies, parent advocacy groups, and religious organizations.[21,22]

Evaluators should declare any personal conflict of interest and take stock of other possible stakeholder influences on a proposed evaluation, as this may influence such things as access to needed information about the nature of an intervention or access to suitable research

TABLE 1. CRITICAL QUESTIONS ADDRESSED IN EACH COMPONENT OF THE EVALUATION MODEL

A1. Stakeholder analysis: identification and issues
- Who are the stakeholders?
- What is the nature of their interest?
- Are there actual or potential conflicts of interest that impact on the evaluation?

A2. Identifying evaluation goals and users
- What are the context and goals of the evaluation?
- Who will use the evaluation?
- What influence will the evaluation have?

A3. Assessing the theoretical antecedents of intervention
- Does the research indicate that the targets selected for intervention are valid?
- Is the theoretical basis of the intervention clear and credible?
- What supporting research has been done?
- How are the intervention strategies supported by the research?
- What is the gap between present knowledge and strategies used in the current intervention design?
- Are the strategies credible?

A4. Software design corresponding to intervention strategies
- What is the case for using digital interactive multimedia?
- Are there clear software equivalents for each therapeutic strategy?
- Is the software design for each strategy explicit?
- Is recording and reporting of relevant dependent variables supported in the software?

A5. Software design corresponding to generality of training
- How does the intervention design address generalizability or transfer-of-training issues?
- What specific methods have been used to promote generality of training?
- How credible is the logic supporting generalizability?

A6. Analyzing the proposed target population
- Who is the intervention for?
- What research supports the use of this intervention with this population?
- What are the population size and access issues that will affect uptake and statistical power?

A7. Analysis of the matching of strategies to target population
- What are the characteristics of the targeted population that may affect intervention outcomes?
- How does the intervention design reflect the characteristics of the target population?
- How do intervention strategies address individual variation on key variables in the target population?

A8. Implementation analysis: media, distribution, and cost
- Is the distribution medium appropriate to the potential users and contexts?
- Are there credibility or cost issues?

B1. Analyzing putative outcome goals and measures
- What treatment goals are specified?
- What components are differentiated?
- Are outcome goals specified? Are corresponding measures clear and reliable?
- Is social validity of goals and corresponding measures established, using accepted criteria?
- What are the criteria for evaluating clinical significance of change in outcome variables?

B2. Evaluating the therapeutic components
- What are the therapeutic components?
- Can the components be evaluated separately?
- What DVs are used to index each component?
- Are the measures complete, clear, valid, and reliable?
- Do components interact with one another?
- Is component analysis required?

B3. Research design of the evaluation
- What alternative research designs are feasible?
- Can a single design be used to test hypotheses about both processes and outcomes?
- What control conditions are required?
- Do the control conditions restrict external validity?
- If placebo or nocebo controls are contemplated, might they be compromised by uncontrolled information?
- If "blind" conditions are required for participants, researchers or others, how might "blindness" be compromised by credibility, expectations, or observations?
- What are the potential validity threats?
- Which design best rules out the known validity threats?
- What are the resource implications (cost, personnel, time) for alternative designs?

B4. Procedural integrity and evaluation integrity
- How will procedural integrity be assessed?
- Will procedural integrity be controlled?
- Will controls for procedural integrity compromise evaluation integrity? (This refers to a reactive effect of evalua tion, in which treatment integrity is assessed as better than typically it would be, were it not being evaluated.)

TABLE 1. CONTINUED

B5. Monitoring the intervention
What is the planned duration of the intervention?
What is the plan for data collection during and after the intervention?
Is the plan consistent with good external validity?
Are there maintenance components to the intervention?
Is follow-up required?
Is the design valid for between-group comparisons at follow-up?
B6. Evaluation data quality
What data will be collected?
What data can be collected internally by the digital program?
What additional, external measures will be required?
Are measures valid, accurate, and reliable?
Is there a requirement to assess validity, accuracy, and reliability of the measures?
If so, what methods will be used?
What are the generality of training issues?
Are the DVs targeted by the intervention intended to represent behavior and attitudes outside the intervention context?
Are there measures that will adequately represent transfer of the effects of the intervention to other contexts and times?
B7. Evaluation data analysis
What models and procedures will be used for analyzing data?
Have power, precision, and effect-size issues been resolved?
For group designs, is there a rational basis for selection of covariates?
For single-case designs, can multiple treatment interference be assessed?
If the digital program has collected time-series data for individuals, can this be analyzed for evidence of causality? (Cross-lag correlation, etc.)
C1. Research ethics issues
Have all ethics issues been traversed and ethics approval sought from relevant review boards?
Is the research design compromised by ethics requirements?
Can potentially confounding effects of project information and consent to participate be assessed by measures included in the evaluation?
Have possible conflicts of interest been anticipated?
C2. Evaluation reporting issues
Is the proposed design and style of reports suitable for the intended audiences?
Has timing of outlines, drafts, and final reports been approved?
What reports are required, and who will write them?
Who will own the intellectual property in the reports?
C3. Evaluation project plan
What are the logistics of the evaluation project?
Has a Gantt chart been developed to show the critical path?
What primary personnel are required, and what are their skills and responsibilities?
What secondary personnel are involved, and on what basis is their cooperation sought?
If secondary personnel are not directly funded for their role in this project, what obligation do they have to perform needed functions?
Is it feasible to provide a cost-benefit analysis for the intervention?
What are the costs of the evaluation?
C4. Evaluability assessment
Can the costs and other resource needs be met?
Can cooperation of all staff and participants be guaranteed for the duration of the evaluation?
Is the choice of research design realistic?
Are there contingencies for dealing with loss of personnel, participants, or data?
Are policy and management decisions (of the organisation conducting the evaluation) documented and agreed to?

participants. Stakeholder issues may have to be addressed in the planning stage to prevent the evaluation from possibly being undermined or its results pre-empted in some way. Things that can go wrong include adverse publicity, misinformation campaigns, participant hostility, outside interference with research conduct, non-acceptance of findings, inability to publish findings, and litigation.

A2. Identifying evaluation goals and users. Since all stakeholders are potential users of evaluation results, the evaluation plan should consider how the goals of the evalua-

tion map onto stakeholder interests. Unless the evaluation will throw light on issues considered important by at least some major stakeholders, the results may have little impact. If the evaluation is comprehensive and uses high standards of research practice, the results will be relevant to most potential users and will not be easily dismissed as being flawed. It is notable that, of the 19 evaluation studies of EEG biofeedback for ADHD located for a recent meta-analysis, only six used control groups and only one of these was published in a peer-reviewed journal.[23] Limited evaluations such as uncontrolled case studies and other studies with weak research designs are unlikely to have much influence on professional judgements about the efficacy of an intervention and may actually be a waste of resources.

A3. Assessing the theoretical antecedents of an intervention. Within the professional community in general, the use of approved scientific methodology to evaluate intervention products and procedures is generally considered to be a good thing, although there is ongoing debate about methodology, provenance, and priorities. For example, it is recognized that evaluation is costly of resources, and may also involve evaluation research participants in a process that may have costs for them, such as delaying access to more-effective treatment. It has been argued that proposed new treatments should not be used, or even evaluated, unless their theoretical credibility can first be established. For example, although theoretical arguments supporting the use of biofeedback treatment of ADHD have been advanced by some,[18] others have seriously questioned whether resources should be expended on evaluation of EEG biofeedback for ADHD.[21] Arguably, the intervention should also be based on credible theory that can be used to validate the structure of an intervention and to provide a basis for choosing appropriate measures of effectiveness of the intervention.

A4. Software design corresponding to intervention strategies. Even if the theoretical basis of the intervention strategies seems to be sound at a general level, it also needs to be clear why digital IM is being used as a vehicle for the intervention. The decision to use digital IM should be supported by a credible argument that this medium has advantages over alternative approaches. In the case of EEG biofeedback, it is easy to argue that digital technology is essential for vital components of the intervention, especially EEG spectrum analysis and biofeedback training protocols.[18] In many educational interventions, though, the use of digital technology is predicated less on need than on cost and convenience.

A comprehensive evaluation would include assessment of whether the software behaves as intended by the corresponding intervention theory. Unless the intervention fairly represents the theory on which it is supposedly based, it can be unclear just what the intervention is or what is being evaluated. To assess this adequately, the evaluator may need either to see reports on checks that have already been made by the developers, or have access to the functional specifications for the software. It may also be necessary to go through the specifications both with their author and the programmers who worked from the specifications to produce the source code. It might be thought that the evaluator could more easily assess the software design by observing the behavior of the software during intervention sessions or by examining records of interventions sessions collected by the software. However, neither of these methods allows the evaluator to conclude either that the software design is faithful to the strategy it represents or that the corresponding program behaves as specified in the software design. This is the heart of a digital IM intervention, and unless it is correctly done, then the intervention is not what it purports to be and may even be a waste of resources. It follows that the corresponding part of the evaluation is of primary importance.

This assessment of the veracity of the software design and implementation is equivalent to an assessment of one aspect of the procedural integrity of the intervention, that it, how well the actual intervention matches its designer's functional specifications. The other aspect of procedural integrity is concerned with the frequency and duration of interven-

tion sessions. One of the potential advantages of using digital media as a vehicle for interventions is the potential for high procedural integrity without the requirement of expert supervision, but this clearly depends on the integrity of the software. None of the published evaluations of EEG biofeedback for ADHD report any analysis of the procedural integrity either of the multimedia representations of EEG parameters or the biofeedback training protocols. In many of the reports, it is unclear just what these procedures are, let alone how well they represent the theoretical ideal.

A5. Software design corresponding to generality of training. All therapeutic interventions have to come to terms with a fundamental issue called "generalizability" or "generality of training." If an intervention is given in a specific setting, using a particular agent, its efficacy may be limited to behavior changes in that setting and with that agent. Even these changes may not be enduring over time, unless specific procedures are employed to maintain the changes. In fact, interventions usually seek to change behavior, not in the intervention setting, but in some other setting in which behavior change is desired, the criterion setting. In general, the weaker the functional correspondence between the intervention setting and the criterion setting, the less successful the intervention is likely to be. There are established strategies for dealing with this problem, and a comprehensive evaluation would include an assessment of the use of these strategies in the software design.[12]

Compared with traditional interventions, interventions using digital IM may be able to reduce the impact of some generality problems, but they may also increase the impact of others. Some studies of academic interventions using digital media (e.g., computer-assisted instruction) have found that skills learned on a computer transferred poorly to criterion settings such as pencil and paper. It has been argued that current technology for digital IM allows it to overcome this problem through its ability to create submersive learning environments and to make use of virtual reality, both of which are regarded as promoting generality of training by simulating the criterion setting.[24–26] However, the reality of this claim would need to be assessed for each intervention.

Evaluations of EEG biofeedback for ADHD have not assessed explicitly whether the intervention includes design elements that support generality of training. On the face of it, generality seems to be an important issue, because the specific skills trained on a computer in a clinic setting are very different from the attentional behaviors and settings that the intervention is supposed to change. Most of these evaluation studies do recognize this issue, to the extent that the outcome measures usually include performance measures and behavior ratings at home and school, as opposed to performance on the biofeedback task.

A6. Analyzing the proposed target population. Interventions usually are designed for a specific target population that will be differentiated by the type of problem the intervention addresses (illness diagnosis, symptomatology) and other characteristics such as age or severity of problem. The promoters of the intervention should make clear who it is designed for, and this is the population on whom it should be evaluated. Sometimes interventions developed for one population are subsequently used or advocated for a different population.

This phenomenon seems especially true of interventions using digital IM, possibly because the cost of implementation typically is small; the intervention is self-delivered or requires minimal supervision. Because interventions using digital IM are seen as having a low delivery cost, lay persons and professionals may be tempted to try them just in case they are effective. A notable example is provided by the ever-expanding application of EEG biofeedback training to new disorders. Developed initially for the reduction of seizures,[27] it subsequently was applied to children with "hyperkenesis"[17] and inattention[28] problems, and more recently, also to children with various types of learning disabilities.[29,30] This intervention now is a franchised commercial product being marketed through the Internet directly to parents of children with ADHD or learning disabilities (www.eegspectrum.com), as well as being advocated as a valid intervention for a range of mental health, behavioral,

and academic problems.[21,31–33] From a theoretical perspective, it is questionable whether a single generic intervention should be predicted to be effective with such diverse populations. This view is reflected in a position statement by one group of biofeedback practitioners, which sets out clear inclusionary and exclusionary guidelines,[18] but it seems that these are not widely accepted.

The availability of suitable participants will strongly influence evaluation designers when defining their target population. Especially if a true experimental design is contemplated and there are some potentially strong uncontrolled independent variables, a large number of participants may be needed to ensure adequate statistical power to detect even a moderately sized intervention effect.[34] In these circumstances, it might be tempting to set overly generous inclusionary criteria that could ultimately result in a very heterogeneous sample and compromise the external validity of the evaluation. Evaluations of EEG biofeedback for ADHD have been seriously weakened by the use of vague or informal diagnostic criteria.[23]

It might be argued that digital IM interventions, being accessible, inexpensive and low on supervision requirements, will inevitably be broadly applied to a heterogeneous population. Wouldn't it make more sense to evaluate their use in that context than in a context that is relatively artificial and has little resemblance to the reality? A counter-argument is that the professional community has a responsibility to provide some leadership and promote professional standards, rather than trying to clean up the mess that otherwise occurs. Good evaluation research does require controlled studies with well-defined populations and conditions, at least in the first instance. It is a well-established principle, for clinical trials methodology, that studies of efficacy, using well-controlled conditions, should precede studies of effectiveness that address the question of whether the intervention works under ordinary circumstances.

A7. Analysis of the matching of therapeutic strategies to target population. A comprehensive evaluation would include an assessment of whether the software design takes account of variation between individuals to whom the intervention will be applied. If the same intervention is to be delivered to all recipients, it will need to accommodate differences in relevant variables such as motivation, cognitive profiles, learning aptitude, and levels of prior knowledge and skills. It may be that the intervention is based on a one-size-fits-all assumption, although the ability to provide individually tailored interventions is generally recognized as one of the most valuable characteristics of digital IM. The designer may have analyzed the target population in order to determine what form the intervention should take that would ensure its appeal, and how the form could be varied to reflect both common characteristics and variability in the population. For example, a cognitive-skills rehabilitation intervention for patients with frontal lobe damage would need to recognize the design limitations imposed by the fatigue, memory, and attention problems characteristic of this disorder, while also accommodating the wide variability in symptomatology and severity.

Given that most interventions are intended to change attitudes and behaviors, the instructional strategies might be designed so that the program can identify and accommodate individual differences in cognition and learning.[35,36] Strategies can allow for differences on variables such as impulsivity, learning rate, selective attention, breadth of attention, sustained attention, and working memory. The success of the intervention strategies, at an individual level, will reflect the attention paid to these variables by the intervention designer. Motivation to use the intervention will be influenced by overall appeal on first use, based on the appropriateness of the context, content, level, and style. Motivation to continue with the intervention will be influenced by individual experience with it, reflecting how well it can accommodate individual differences.

These issues are not conceptually different from those thought important for the success of therapy delivered by a "live" therapist. Just as a particular therapy usually is more successful in the hands of its developers than in the hands of others,[37] some digital IM translations of a therapeutic theory or strategy will be more successful than others. The "art" of ther-

apy is about tailoring the conduct of therapy so that the "science" can be made to work at the individual level. Digital IM simply attempts to create a proxy expert therapist by translating the rules of best practice into their digital equivalents. The evaluator might consider whether, for a given intervention design, this translation attempt reaches an adequate standard.

Evaluation studies of EEG biofeedback have had little to say about individualization of intervention programs, although some practitioners see this as a vital component of effective treatment.[18] Aspects that could be individualized include the multimedia environment and the biofeedback training protocol. Since children with ADHD typically may have difficulty sustaining attention during training, provision for the child to choose a preferred multimedia training environment could be very beneficial for learning. Details of the biofeedback training protocols typically are not included in reports of evaluation studies, and it is unclear whether the parameters of these procedures are empirically based (say, on unreported pilot studies) or simply ad hoc. These details should be transparent to the evaluation process, because they can potentially account for large differences in the effectiveness of biofeedback training.

A8. Implementation analysis: media, distribution, and cost. The impact of a potentially effective digital IM intervention can be reduced by cost and other access problems that might restrict its use by the target population. This is an effectiveness issue rather than an efficacy issue, but worth considering as part of a comprehensive evaluation. For digital IM, the choice of platform can potentially restrict use. If the platform is restricted to either Mac or PC, or if the system requirements are too high, a major section of the potential target population will have access problems. Distribution by Internet download or as a program that runs online can raise access problems for those with limited Internet access or narrowband connections. Cost (to the consumer) potentially can be significantly less than other modes on intervention, unless the publisher is seeking to recover development costs quickly. These issues need to be planned to ensure that the target population will have easy and affordable access at home or wherever the intervention is intended to be used.

Credibility of the intervention in the eyes of professional and lay users is another factor that will affect access. Although the public may well have alternative ways of accessing a digital product, many potential users will still seek professional advice before trying something. Credibility with professional gatekeepers is therefore an issue that can impact on implementation both of an intervention and its evaluation. It will be difficult to evaluate an intervention if relevant professional groups distrust it, think it is a scam, see it as a threat to their own status or livelihood, or simply believe that it will not work. In the absence of prior evaluations of efficacy, pilot studies and theoretical analyses will have an important influence on professional opinion.

B. Evaluation of efficacy: process and out-come

The steps described in the first part of this model should establish clearly the underpinning theory and formative research, what the intervention is, what it aims to achieve, who it is for, how it is structured, and who has an interest in it. If this part of the evaluation shows that the intervention reaches acceptable standards, it is appropriate to continue to the second part, the evaluation of efficacy. Otherwise, it may be questionable whether further resources should be expended on evaluation of efficacy if there are serious underlying problems of credibility or acceptability.

B1. Analyzing putative outcome goals and measures. Having already in the first part established what the therapeutic goals of the intervention are, the next step is to establish operational definitions of dependent variables that will be used in the evaluation to represent these goals. Criteria for adequate definitions are found in several texts and articles on evaluation research design.[5,11–13,38–41]

Outcome goals usually are specified as changes in the behaviors and attitudes that are being targeted by the intervention, in the criterion setting. Consideration should be given to

declaring a size of change that would be of practical importance to the individual, a so-called clinically significant change.[39] Suitable indices can be based on statistical methods such as the Reliable Change Index,[42] on the presence or absence of critical behaviors or attitudes, or on cut-off scores for formal diagnosis. For some stakeholders, outcome goals are the sole interest, and it is important that there is clear understanding about what these goals should be and how they will measured.

Some outcome goals may be defined by behaviors or attitudes that are closely related to those directly targeted in the intervention, but others might be defined by entities further along some assumed causal chain. For example, in EEG biofeedback interventions for ADHD, the direct targets of the intervention are EEG patterns, but reduction in behavioral symptoms of ADHD often is the main outcome goal. When a psychoeducational intervention is an adjuvant to another treatment, for example, chemotherapy, the outcome measures of primary interest might be duration of remission or some immune system marker. To some stakeholders, the direct effects of the psychoeducational intervention on health behaviors and attitudes might be only of secondary interest.

B2. Evaluating the therapeutic components. Process goals and the measures corresponding to them usually are closely related to the direct targets of the intervention process. For example, if an EEG biofeedback intervention directly targets the reduction of the Theta/Beta ratio in the EEG spectrum, using differential reinforcement, then this ratio would be a valid index of the therapeutic process. Maintenance of a reduced Theta/Beta ratio, after the biofeedback intervention was discontinued and in a different setting, would be a valid measure of generalizability of this process-induced change. The data on these measures could be collected by the digital program during the intervention sessions. Of course, the process of interest may be more complex than indicated by these measures. For example, biofeedback interventions may use a scaffolding approach to assist the participant to learn the therapeutic task. One version of this is a shaping procedure which initially sets an easy task that is made progressively more difficult as the participant learns. Collecting data on the operation of this process would allow the evaluator to see whether it is working as intended and whether it could be modified to make it more effective.

Process evaluation has two important roles. Firstly, it is valuable for establishing the validity of an intervention process, that is, whether the process does what it supposed to do when it is used with a participant. If the process can be shown to be valid, it will at least be a plausible cause of changes in the outcome measures. Secondly, it can provide insight into mechanisms of change, for example, why the intervention might work better with some participants than others.

When interventions have more than one component, it can be difficult to get information about the efficacy of the different components unless the research design specifically allows for this. Process evaluation can be used to assess how each component operates during the intervention.

B3. Research design of the evaluation. An evaluation needs a research design in order to control for threats to internal and external validity. The sovereign requirement is that it must be possible to arrive at a clear interpretation of the results of the intervention by being able reasonably to rule out alternative interpretations.[11,13,43] A range of suitable designs might be contemplated, and the choice between them is determined jointly by the nature of the research hypotheses to be tested and the practicalities of the evaluation. Factors to be considered include estimated effect sizes, statistical power required, number of participants available, number that can practically be included in the evaluation, whether random allocation to treatments is possible, the number of dependent and independent variables to be studied, and whether or not some dependent variables will be measured more than once.[6,8,11,40,44–48]

In some evaluations, subdesigns might be nested within the overall research design. This is the case, for example, when single-case research designs are used for process evaluation

while randomized between-groups designs are used to evaluate outcomes.

The use of placebo or nocebo controls for participant expectations, and the use of masking or "blinding" of researchers and others involved in the evaluation needs careful consideration. While such controls can be useful for eliminating alternative explanations for putative intervention effects, they can be difficult or impossible to implement effectively in some cases.[7,49–55]

Recent reviews and meta-analyses of the evaluations of EEG biofeedback interventions have pointed out a number of problems with the research designs of these evaluations.[23,31] These include failure to control for basic threats to internal validity such as history and maturation, failure to include appropriate placebo controls, failure to take account of participant drop-out, confounding of treatment components, and failure to collect follow-up data.

B4. Procedural integrity and evaluation integrity. As previously noted, in digital IM interventions, one major aspect of procedural integrity is controlled by the software design. Unlike human-delivered interventions, the putative intervention procedure does not require being translated or interpreted by a therapist, so there is no need to check on how this is done. A software audit stands in for this process. However, it is still necessary to evaluate whether the participants used the intervention in the prescribed manner. Arrangements need to be made to collect valid records of how participants use the intervention. If participants are asked to self-monitor their usage pattern, the evaluator needs to consider how this requirement will influence the pattern of use.[14,56]

B5. Monitoring the intervention The evaluation of an intervention requires a plan for data collection before, during, and after the intervention. The minimal requirements are specified by the research design. In addition, procedures are needed for monitoring participants so that problems that arise can be addressed in a timely fashion, and possible data loss, procedural integrity problems, and defections can be anticipated.[7,57]

B6. Evaluation data quality. Good data quality needs to be assured. Assuming measures have been chosen on the basis that they are capable of meeting acceptable validity and reliability criteria, it still must be insured that standards are met when the measures are used in the current evaluation.[8,12,45,58]

B7. Evaluation data analysis. Data analysis is an important component of an evaluation. Choice of appropriate methods should be made at the time the research design is planned, since the two must be a good fit. Statistical methods for comparing scores of groups of participants exposed to different conditions, or for comparing scores obtained from the same group during different conditions, are readily accessible.[7,8,59–66] Suitable methods for various single-case designs are also available. These are particularly useful for process evaluation at the individual level.[7,11,12,40,67,68]

Inappropriate data analysis is not uncommon, often leading to incorrect conclusions being drawn about the efficacy of the intervention being evaluated. With regard to EEG biofeedback studies, weaknesses of analysis include an almost exclusive use of pre-post *t* tests as evidence for treatment effects in studies with invalid designs, and failure to control for type I errors, especially in multiple comparisons of means.[23]

C. Planning and reporting the evaluation

C1. Research ethics issues. There is no point in planning research that will not be approved by ethics review committees. Relevant issues go well beyond the obvious matters of informed consent and confidentiality. Applicants may be required to produce a valid research design and outline appropriate methods of analysis. Many ethics committees now require applicants to state how they will report their findings, indicating the researchers' responsibility to disseminate findings to appropriate audiences who will receive and use the information. They also require evidence that the project has the approval of relevant stakeholders and that sufficient resources have been committed to make the project viable.[7,69,70]

C2. Evaluation reporting issues. An evaluation might result in more than one report. There might be different reports tailored for different audiences, or separate reports on different aspects of the evaluation. Preliminary reports are not unusual, especially for evaluations of longitudinal interventions. Many factors jointly determine the form, style, content and timing of reports, and these need to be considered before any serious report preparation is done.[5,7,71]

C3. Evaluation project plan. The overall project plan is developed after most of the evaluation components have been considered and decisions made about which will be included and how they will be accomplished. By this stage, the evaluation team will be in a position to make a good estimate of the resources, including time, required for the evaluation.[5,72]

C4. Evaluability assessment. Once the planning process is complete, there is need for a final appraisal of the overall plan, keeping in mind the likely sources of risk that the project might face. Referred to as "evaluability assessment," this process is a check on the feasibility of the proposed evaluation. There are many events that potentially can prevent an evaluation from achieving its goals, but the more usual ones can be anticipated or avoided.[5] The evaluation plan may need to be revised in the light of the evaluability assessment.

CONCLUSION

As indicated in Figure 1, it is good practice to complete the evaluation planning in full before commencing the intervention that will be evaluated. Of course, this does not mean that reports will be written before the intervention is begun, only that the reporting requirements will be planned in advance. Although Figure 1 shows the evaluation process as a fixed sequence, it is understood that the sequence used in practice will vary with circumstances. The sequence shown is intended as a realistic and logical example. In practice, evaluation planning is an iterative process in which components are revisited and revised during the formulation of the evaluation plan. For example, decisions made about components of the third part may result in revision of preliminary design and methodology decisions made in components of the second part.

The model described in this paper combines relevant components of program evaluation and clinical trials to provide a coherent model tailored to the features of digital IM interventions. Using the example provided by evaluations of EEG biofeedback for ADHD, the model is shown to be useful in addressing areas in which previous evaluations of digital IM interventions appear to be deficient or which have been overlooked.

It is not suggested that all evaluation studies should use all the components of this comprehensive model. The evaluation needs of different interventions will vary according to their nature and the context of their use. What is suggested is that the model will provide a useful checklist for evaluation planning, which will lessen the chances that some relevant component will be overlooked.

ACKNOWLEDGMENT

This work was supported by HopeLab, Palo Alto, California.

REFERENCES

1. IsHak, W.W., & Burt, T. (1998). Computer applications in psychiatry: role in patient care, education, research and communication. *CyberPsychology & Behavior* 1:147–150.
2. Wiederhold, B.K., & Wiederhold, M.D. (2000). Lessons learned from 600 virtual reality sessions. *CyberPsychology & Behavior* 3:393–400.
3. Schwartz, M.S., & Olson, P.R. (1995). A historic perspective on the field of biofeedback and applied psychophysiology. In: Schwartz, M.S. (ed.), *Biofeedback: a practitioner's guide*. New York: Guilford Press, pp. 3–18.
4. Shadish, W.R., Cook, T.D., & Leviton, L.C. (1991). *Foundations of program evaluation: theories of practice.* Newbury Park, CA: Sage.
5. Wholey, J.S., Hatry, H.P., & Newcomer, K.E. (1994). *Handbook of practical program evaluation*. San Francisco: Jossey-Bass.

6. Matthews, J.N.S. (2000). *An introduction to randomized controlled clinical trials*. London: Arnold.
7. Institute of Medicine (2001). *Small clinical trials: issues and challenges*. Washington, DC: National Academy Press.
8. Meinert, C.L., & Tonascia, S. (1986). *Clinical trials: design, conduct, and analysis*. New York: Oxford University Press.
9. Pocock, S.J. (1984). *Clinical trials: a practical approach*. Chichester, U.K.: Wiley.
10. Friedman, L.M., Furberg, C.D., & DeMets, D.L. (1998). *Fundamentals of clinical trials*. New York: Springer.
11. Kazdin, A.E. (1994). Methodology, design, and evaluation in psychotherapy research. In: Garfiel, S.C. (ed.), *Handbook of psychotherapy and behavior change*. New York: Wiley, pp. 19–71.
12. Barlow, D.H., & Hersen, M. (1984). *Single-case experimental design: strategies for studying behavior change.* New York: Pergamon.
13. Rosenthal, R., & Rosnow, R.L. (1991). *Essentials of behavioral research: methods and data analysis.* New York: McGraw-Hill.
14. Gresham, F.M. (1996). Treatment integrity in single-subject research. In: Franklin, R.D., Allison, D.B. (ed.), *Design and analysis of single-case research*. Mahwah, NJ: Erlbaum, pp. 93–117.
15. Lieberman, D.A. (2001). Management of chronic pediatric diseases with interactive health games: theory and research findings. *Journal of Ambulatory Care Management* 24:26–38.
16. Brown, S.J., Lieberman, D.A., Gemeny, B.A., et al. (1997). Educational video game for juvenile diabetes: results of a controlled trial. *Medical Informatics* 22:77–89.
17. Lubar, J.F., & Shouse, M.N. (1976). EEG and behavioral changes in a hyperkinetic child concurrent with training of the sensorimotor rhythm (SMR): a preliminary report. *Biofeedback Self Regulation* 1:293–306.
18. Lubar, J.F. (1995). Neurofeedback for the management of Attention-Deficit/Hyperactivity Disorders. In: Schwartz, M.S. (ed.), *Biofeedback: a practitioner's guide*. New York: Guilford Press, pp. 493–522.
19. Othmer, S., & Kaiser, D. (2000). Implementation of virtual reality in EEG biofeedback. *CyberPsychology & Behavior* 3:415–420.
20. B.B.C. (2002). Video games help hyperactive children. news.bbc.co.uk/hi/english/sci/tech/nearsid_894000/894673.stm.
21. Barkley, R.A. (1992). Is EEG biofeedback treatment effective for ADHD children? *CHADDerBox* 5–11.
22. Reid, R. (1996). The three faces of attention-deficit hyperactivity disorder. *Journal of Child and Family Studies* 5:249–265.
23. Heywood, C.E. (2002). An assessment of EEG biofeedback for the remediation of attention-deficit/hyperactivity disorder. *Dissertation Abstracts International* AAT 3029986.
24. Emmelkamp, P.M., Bruynzeel, M., Drost, L., et al. (2001). Virtual reality treatment in acrophobia: a comparison with exposure *in vivo. CyberPsychology & Behavior* 4:335–339.
25. McComas, J., Pivik, J., & Laflamme, M. (1998). Children's transfer of spatial learning from virtual reality to real environments. *CyberPsychology & Behavior* 1:121–128.
26. Wiederhold, B.K., & Wiederhold, M.D. (1998). A review of virtual reality as a psychotherapeutic tool. *CyberPsychology & Behavior* 1:45–52.
27. Kaplan, B.J. (1975). Biofeedback in epileptics: equivocal relationship of reinforced EEG frequency to seizure reduction. *Epilepsia* 16:477–485.
28. Shouse, M.N., & Lubar, J.F. (1979). Operant conditioning of EEG rhythms and ritalin in the treatment of hyperkinesis. *Biofeedback and Self Regulation* 4:299–312.
29. Tansey, M.A., & Bruner, R.L. (1983). EMG and EEG biofeedback training in the treatment of a 10-year-old hyperactive boy with a developmental reading disorder. *Biofeedback & Self Regulation* 8:25–37.
30. Scheinbaum, S., Zecker, S., Newton, C.J., et al. (1995). A controlled study of EEG biofeedback as a treatment for attention-deficit disorders. In: *Proceedings of the 26th Annual Meeting of the Association for Applied Psychophysiology and Biofeedback.* 131–134.
31. Arnold, L.E. (1999). Treatment alternatives for attention-deficit/hyperactivity disorder (ADHD). *Journal of Attention Disorders* 3:30–48.
32. Robbins, J. (1998). *Wired for miracles? Psychology Today* 8.
33. Othmer, S., Othmer, S.F., & Kaiser, D.A. (2002). EEG biofeedback: a generalized approach to neuroregulation. www.eegspectrum.com/applications/intro/generalizedapproach
34. Machin, D. (1997). *Sample size tables for clinical studies*. Oxford: Blackwell Science.
35. Beale, I.L. (2000). *Reading basics (disc 2): integrated neural learning.* Arrowtown, N.Z.: Brainwaves.
36. Sampson, D., Karagiannidis, C., & Kinshuk (2002). Personalised learning: educational, technological and standardisation perspective. *Interactive Educational Multimedia* 4:24–39.
37. Kavale, K. (1990). Variances and verities in learning disability interventions. In: Wong, B. Y. L. (ed.), *Intervention research in learning disabilities.* New York: Springer-Verlag, pp. 3–33.
38. Cronbach, L.J. (1980). *Toward reform of program evaluation.* San Francisco: Jossey-Bass.
39. Kazdin, A.E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification* 1:427–452.
40. Kratochwill, T.R. (1978). *Single-subject research: strategies for evaluating change.* (New York: Academic Press.
41. Scriven, M. (1993). *Hard-won lessons in program evaluation.* San Francisco: Jossey-Bass.
42. Jacobson, N.S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy. *Journal of Consulting and Clinical Psychology* 59:12–19.

43. Winer, B.J., Brown, D.R., & Michels, K.M. (1991). *Statistical principles in experimental design*. New York: McGraw-Hill.
44. Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
45. Chow, S.-C., & Liu, J.-P. (1998). *Design and analysis of clinical trails: concept and methodologies.* New York: Wiley.
46. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
47. Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: design and analysis issues for field settings.* Chicago: Rand McNally.
48. Jadad, A.R. (1998). *Randomised controlled trials: a user's guide. London: BMJ Books.*
49. Critelli, J.W., & Neumann, K.F. (1984). The placebo: conceptual analysis of a construct in transition. *American Psychologist* 39:32–39.
50. Horvath, P. (1988). Placebos and common factors in two decades of psychotherapy research. *Psychological Bulletin* 104:214–225.
51. Peek, C.J. (1977). A critical look at the theory of placebo. *Biofeedback and Self-Regulation* 2:327–335.
52. Shapiro, A.K., & Shapiro, E. (1997). *The powerful placebo: from ancient priest to modern physician*. Baltimore: Johns Hopkins University Press.
53. Shepherd, M., & Sartorius, N. (1989). *Non-specific aspects of treatment*. Toronto: WHO/H. Huber.
54. Wilkins, W. (1985). Placebo controls and concepts in chemotherapy and psychotherapy research. In: Schwartz, G.E. (ed.), *Placebo: theory, research and mechanisms*. New York: Guilford, pp. 83–109.
55. Wilkins, W. (1986). Placebo problems in psychotherapy research: social-psychological alternatives to chemotherapy concepts. *American Psychologist* 41:551–556.
56. Cramer, J.A., & Spilker, B. (1991). *Patient compliance in medical practice and clinical trials*. New York: Raven Press.
57. Klein, J.P., & Moeschbereger, M.L. (1997). *Survival analysis: techniques for censored and truncated data.* New York: Springer.
58. Piantadosi, S. (1997). *Clinical trials: a methodologic perspective*. New York: Wiley.
59. Cleophas, T.J.M., Zwinderman, A.H., & Cleophas, T.F. (2000). *Statistics applied to clinical trials*. Dordrecht: Kluwer Academic.
60. Edgington, E.S. (1987). *Randomization tests*. New York: Dekker.
61. Haase, R.F., & Ellis, M.V. (1987). Multivariate analysis of variance. *Journal of Counseling Psychology* 34: 404–413.
62. Howell, D.C. (1992). *Statistical methods for psychology.* Belmont, CA: Duxbury.
63. Marascuilo, L.A., & Serlin, R.C. (1990). *Statistical methods for the social and behavioral sciences*. New York: Freeman.
64. McFadden, E. (1998). *Management of data in clinical trials.* New York: Wiley.
65. Redmond, C.K., & Colton, T. (2001). *Biostatistics in clinical trials*. Chichester, U.K.: Wiley.
66. Siegel, S., & Castellan, J. (1988). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.
67. Kazdin, A.E. (1984). Statistical analyses for single-case experimental designs. In: Hersen, M. (ed.), *Single case experimental designs: strategies for studying behavior change*. New York: Pergamon, pp. 285–324.
68. Onghena, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: a rejoinder. *Behavioral Assessment* 14: 153–171.
69. Smith, T. (1999). *Ethics in medical research: a handbook of good practice.* Cambridge, U.K.: Cambridge University Press.
70. Spece, R.G., Shimm, D.S., & Buchanan, A.E. (1996). *Conflicts of interest in clinical practice and research.* New York: Oxford University Press.
71. Browner, W.S. (1999). *Publishing and presenting clinical research.* Baltimore: Williams & Wilkins.
72. Palumbo, D.J. (1987). *The politics of program evaluation.* Newbury Park, CA: Sage Publications.

Address reprint requests to:
Ivan Beale, Ph.D.
HopeLab
101 University Ave., Ste 220
Palo Alto, CA 94301

E-mail: ibeale@hopelab.org