methodological issues in investigations of massage/bodywork therapy

Claire M. Cassidy, Ph.D.

AMTA foundation
American Massage Therapy Association Foundation

This paper was commissioned by the AMTA Foundation and written by Claire M. Cassidy, Ph.D. It has not been edited.

AMTA foundation
American Massage Therapy Association Foundation

Mission Statement

The AMTA Foundation advances the knowledge and practice of massage therapy by supporting scientific research, education and community outreach.

Goals

Massage therapy is accessible to the broadest spectrum of society.

Members of the general public, healthcare professionals and the wellness community value massage therapy.

There is more well-designed research about massage therapy.

Research reflects massage therapy as practiced.

The massage therapy profession is research literate.

Current research findings are integrated into massage therapy practice.

The Foundation is creative, energetic, and organizationally effective.

820 Davis Street, Suite 100, Evanston, IL 60201
847.869.5019 • fax 847.864.1178
www.amtamassage.org/foundation/home.htm

foundation
AMTA American Massage Therapy Association Foundation

Board of Trustees

John Balletto, President
Providence, RI

Sharon Marden Johnson, Vice President
China Village, ME

Randa Cherry
Cedar Rapids, IA

Debra Curties
Toronto, ON, Canada

Jacqueline A. Hart, M.D.
Newton, MA

Tim Herbert
Eugene, OR

Brian M. Marcotte, Ph.D.
Providence, RI

Maureen Moon
Boulder, CO

Steve Olson
Fargo, ND

Clarence E. Smith, M.D.
Miami Beach, FL

Diana L. Thompson
Seattle, WA

Staff

Elizabeth M. Lucas
Executive Director

Gini S. Ohlson
Director of Development and Foundation Manager

Debbie Scanlon
Development Coordinator

Marcus Banks
Research and Programming Associate

CONTENTS

INTRODUCTION

Massage and Bodywork Therapies (MBT) are known worldwide and have an ancient history. Until recently knowledge in this health care system (as in all health care systems) was primarily gained through clinical observation and experience. Today, the MBT profession wishes to scientifically test and explore its clinical knowledge. The task this profession faces is to do science in such a way that the resulting data are of high quality and accurately represent massage and bodywork to the larger world.

This task is not insignificant. Scientific method demands that bias be controlled and minimized to ensure that the data produced is as accurate and credible as possible. Though it is easy to write these words, in practice it is quite difficult to minimize bias and maximize the usefulness of data. This is because bias appears in many forms, often quite subtle. One of the subtlest is the bias introduced by sociocultural expectation and habit. The practice of science, and the practice of health care, are both deeply affected by such expectation. Thus, doing research on MBT, which has been socioculturally defined as "alternative," demands special vigilance to ensure that the factors that make this profession distinctive are accounted for in the design of research. Without such vigilance, results may not accurately represent what massage and bodywork are capable of doing, and why.

This point is developed in this paper by introducing the concept of "model fit validity" and showing its use in designing research. The paper will not survey scientific method; instead it is arranged as a series of loosely linked discussions of "issues" in MBT research. An understanding of these issues will help readers to accurately apply existing methodological literature to research in MBT.

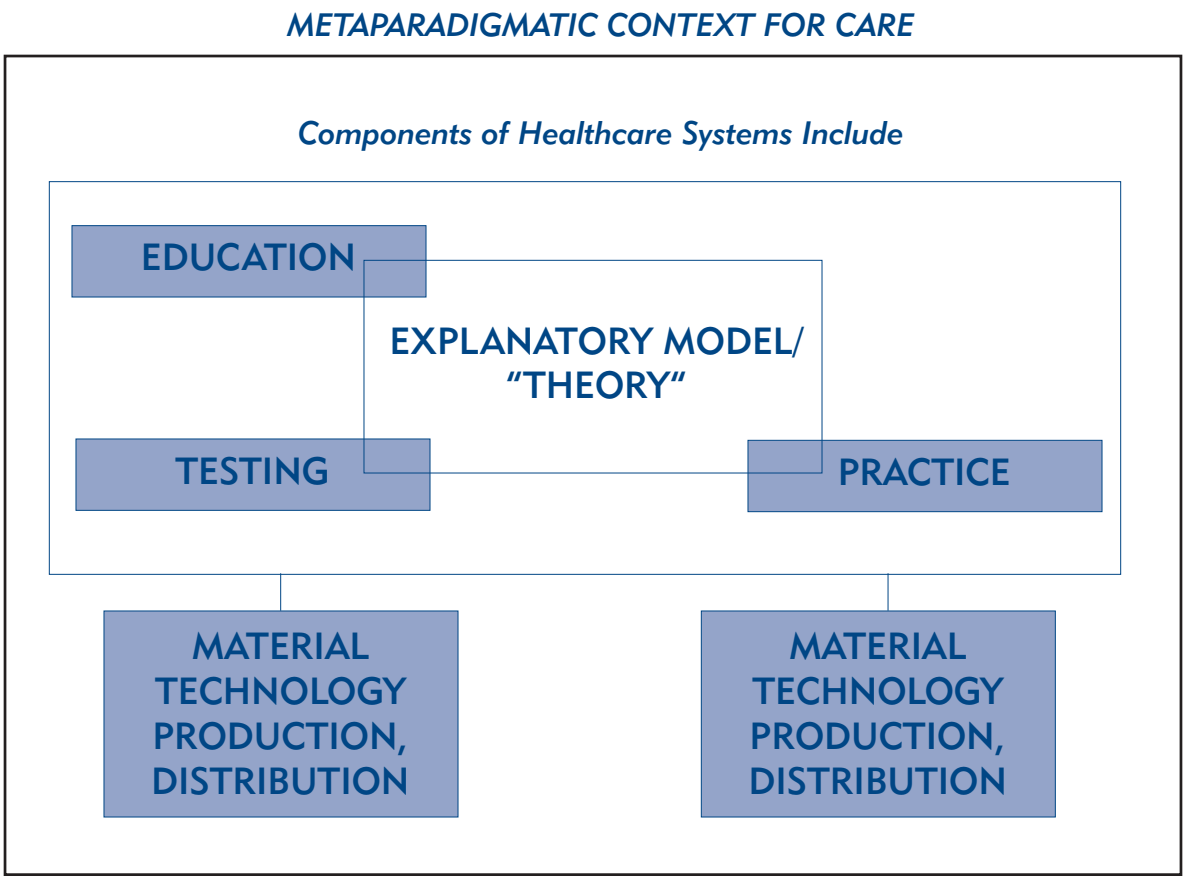
---

[1]Term coined by Arthur Kleinman: "the notions about an episode of sickness and its treatment that are employed by all those engaged in the clinical process. The interaction between the EMs of patients and practitioners is a central component of health care. The study of practitioner EMs tells us something about how practitioners understand and treat sickness. The study of patient and family EMs tells us how they make sense of given episodes of illness, and how they choose and evaluate particular treatments. The study of the interaction between practitioner EMs and patient EMs offers a more precise analysis of problems in clinical communication. Most importantly, investigating EMs in relation to the sectors and subsectors of health care systems discloses one of the chief mechanisms by which cultural and social structural context affects patient-practitioner and other health care relationships (Kleinman 1980:105).

ISSUE 1: WHAT KIND OF MEDICINE IS MBT?

Every form of medicine can be analyzed in terms of how it trains practitioners, delivers care, creates the technology that supports its therapies, is supported by the members of society, and, finally, by how it understands itself (Figure 1). Knowing how a practice understands itself is crucial to the design of research. The part of self-understanding that is obvious to users, that is distilled into textbooks and is taught overtly, is called theory . Theory, however, represents only a part of the whole explanatory model of any practice: much of what practitioners know and teach, and what patients expect, is out-of-awareness. These aspects—assumptions, beliefs, untaught practices—are nevertheless just as important as the overt practices. Indeed, it is what one is unaware of that can most easily mislead when planning research.

Figure 1: The Structure of Cultural Health Care Systems



Based on practitioner interviews and published definitions, I propose the following partial description of the explanatory model of MBT for use in discussing model fit validity and the utility of popular research designs.

MBT consists of a group of **hands-on** interventive strategies that have come to be grouped together, often in different ways by different observers. Though all share the desire to promote patients' well-being, how they perceive themselves doing this varies quite a lot. An obvious continuum exists between those who argue for a strictly materialistic model—they physically touch patients and aim to create structural change—to those who employ an energetic or even spiritual model, often not touching the physical body, but working just above it on the "energy body." A similar continuum exists between those who give materialistic explanations for what is happening "inside" their patients—the release of fascia, the promotion of circulation, the production of endorphins—and those who give psychosocial explanations focusing on self-perception and self-actualization.

MBT practitioners commonly define their practices as "**holistic**" meaning that their *intent* is to treat the "whole person, mind, body, and spirit." The delivery environment also encourages other components of holism[2]: an interactive and relatively egalitarian therapeutic relationship in which the practitioner draws upon the patient's knowledge and perception of his or her own body to guide intervention; and other forms of closeness such as informal conversational styles and direct and immediate payments to practitioner.

Some specific goals of MBT include:

- to touch the patient/client (touch may not involve physical contact)
- to manipulate soft tissues—especially skin, muscles, fascia
- to move bodily fluids—blood, lymph, cerebrospinal fluid
- to increase range of motion in joints
- to promote structural change
- to balance and remove blockages in the flow of energy
- to enhance immune function
- to ease pain
- to promote relaxation
- to provide comfort—both physical and emotional
- to stimulate the body's natural tendency to heal itself
- to support self-realization.

Touch is the *sine qua non* of MBT. It is defined as *intentional* in contrast to casual or occupational. The character of this touch, say practitioners, is what makes or breaks the therapeutic encounter. Good touch is rewarding to patients; it is also rewarding to practitioners. Touch is powerful: it activates the deepest recesses of being and can have profound healing (or harming) effects. Touch can perform structural tasks; it can perform emotional, energetic and spiritual tasks. In the latter case it can help people become more attuned to their bodies, "feel" them better, thus be in a better position both physically and intellectually to make changes, such as to stop addictive or harmful habitual behaviors. Touch raises consciousness. This kind of touch is *skilled*; there is no accident about the benefits that accrue to patients.

MBT is delivered in particular settings and in particular ways: a) it is immediate (it does not occur later and at home as does self-medication); b) it is intimate—the practice is one-on-one,

[2]"Holism" is a popular term used in many ways, yet rarely operationalized. For an effort to define operationally, linked to research data, see Cassidy 1998.

and the practitioner's hands, words, energy touch the patient's often partially unclothed body creating a vulnerability that demands a high degree of trust between patient and practitioner; c) it is direct—in private practices there are often no intermediaries between patient and practitioner—together they make appointments, pay bills, create treatment plans; d) it takes time, typically 30 - 60 minutes for a whole-body session.

Though, ultimately, the MBT profession itself must study and identify the overt and hidden components of its own explanatory model, from this preliminary analysis I conclude that essential characteristics of MBT include:

- it's a hands-on practice
- touch is core
- communication between patient and practitioner is core
- practitioners hope to address the "whole" patient.

These are factors that must be accounted for in research designs. The reasons why will become clearer later in this paper. At the moment, notice how different these core characteristics are from the core characteristics of biomedicine, the medicine whose explanatory model assumptions dominate medical scientific research. Biomedicine emphasizes pharmaceuticals, not touch; patient and practitioner communication is nowadays commonly stylized and reduced to brief contacts of ten minutes or less; sickness and diseases are the focus of care, not the person.

ISSUE 2: FINDING RESEARCH QUESTIONS: THE FOUR FIELDS MODEL

MBT has, to date, developed a relatively small amount of scientific data concerning its character and effectiveness. A tremendous "backlog" of research need exists, but how is one to choose which sorts of questions to address first (second, third...)? The following model may help clarify this issue for MBT practitioners.

The question that guides most research on health care is "Does it work?" This question about effectiveness can be asked in a myriad of ways and at a myriad of scales. Another equally cogent question, though much more rarely asked, is "Does it serve?" This second question applies when something has been shown to "work" but its usefulness in daily life remains in question. For example, an intervention could "work," but could be so painful, so complex, so alienating, or so costly that it is unlikely to "serve" people. The issue concerning service tends to make more sense to holistically oriented scientists (and practitioners) than to reductionistic thinkers because the former are deeply interested in and concerned with systems and relationships.

The various ways of gathering scientific data, and the various venues in which data can be collected to answer the questions "Does it work?" and "Does it serve?" can be understood to devolve into sub-questions about mechanisms:

What makes a practice effective in a social sense?

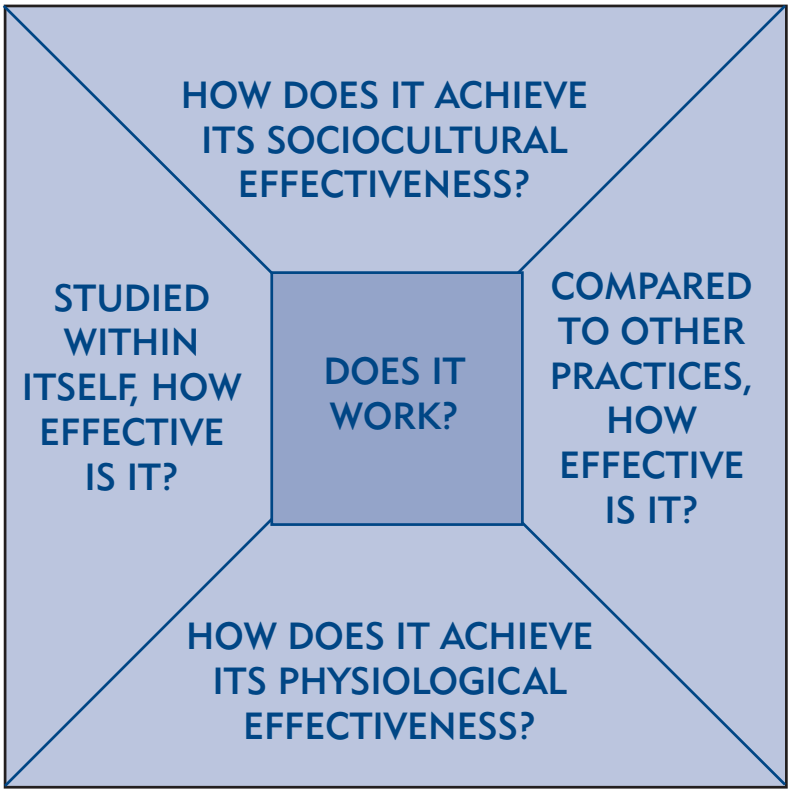
What makes it effective in a clinical sense?

What makes it effective in a physiological sense?

The second question can be viewed from two angles, depending on whether one is trying to understand the deep structure of MBT itself, and study practices that fall within the purview of MBT, or if one is trying to compare MBT interventions with other interventions, such as those of biomedicine. Figure 2 summarizes all this in a simple map.

As the Figure shows, there are four domains of research, in the sections numbered 1 to 4. Methods appropriate to Domain 1 include archival, qualitative and quantitative survey research (Table 1). Methods appropriate to domains 2 and 3 are similar, and include archival, survey, clinical outcomes and clinical trials research. Methods appropriate to domain 4 include mainly laboratory experiment approaches.

Figure 2:



Domain 1 includes all questions that deal with demographics, costs of care, epidemiology, education and outreach (schooling, public information), law and politics, history of the field, and anything that has to do with how people perceive and evaluate the field, explain it to themselves or others, and become motivated to study in the field or receive MBT therapy.

Domain 2 includes all questions that deal with issues about the practice of MBT—comparisons of intervention techniques (Swedish massage vs. deep tissue massage vs healing touch...), appropriate time spent in each session, appropriate frequency of session, how long the effects of a session should last and how this can be measured, the effect of the practitioner on result of care, and many other such questions.

Domain 3 includes all questions that aim to compare the effectiveness of MBT at relieving symptoms or maintaining functionality with the effectiveness of other practices at doing the same thing. It is in domain 3 that MBT researchers will compare their interventions with those of biomedicine, psychology, Chinese medicine, and so on.

Domain 4 includes all research on underlying physiological mechanisms. It is often performed on animal models, such as seeking evidence of how MBT therapy releases endorphins, modifies serotonin release, changes the rate of production of white blood cells, and similar questions.

Note that the content of these domains requires different research approaches and is also likely to attract different audiences. Domain 1 contains questions that will attract social and behavioral scientists, while and Domains 2 and 3 will attract clinicians, and Domain 4 will attract bench scientists. Others will also be interested: the descriptive and explanatory material that emerges from Domain 1 and the overlap issues between Domain 1 and Domains 2 and 3 (e.g., cost-effectiveness data) will be of interest to healthcare planners and insurance/HMO managers. The comparative data from Domains 2 and 3 will attract the attention of insurance and managed care organizations, plus the interest of practitioners of medicines other than MBT. The public should be interested in all social data (Domain 1) and in comparative data from Domains 2 and 3, plus somewhat in the data from Domain 4. Funding agencies and policy makers should be interested in data from all four domains.

Table 2 offers examples of questions that typify each domain. The list is certainly not exhaustive. Some questions fit neatly into one domain, while others tend to link two domains. For example, if one wishes to ask a question about the cost-effectiveness of the care while comparing the effectiveness of a biomedical intervention and a bodywork intervention, the project would apply to both domains 1 and 3. Similarly, if one wanted to track the success of MBT practitioners from the time they were admitted to school through the fifth year of their independent careers, this project would overlap domains 1 and 2.

Because of the prominence of the biomedical and reductionistic models of medical research in our society, there is a strong tendency for people to honor research in domains 3 and 4 over research in domains 1 and 2. These forms of research are also much better funded. Ironically, in order to do high quality research in domains 3 and 4 one actually needs to know a great deal about issues that fall into domains 1 and 2. It is my belief that the first research task for MBT is not to compare their practices with biomedical practices, nor yet to involve themselves in

Table 1: Domains of Research

Domains of Effectiveness[3]	Main Research Methods[4]	Topics of Study	Audiences for Information
1: SOCIOCULTURAL EFFECTIVENESS	• Archival Survey: Qualitative Quantitative • In depth study Qualitative	• History • Demographics • Epidemiology • Satisfaction • Cost-effectiveness • Practioner & delivery characteristics • Patient-practitioner relationships • Practice politics, economics, law • Educational practices • Philosophy of practice	• Patients/Users • Practitioners • Researchers • Third Party Payers • Policy Makers • Behavioral Scientists • General Public
2: WITHIN PRACTICE EFFECTIVENESS	• Archival Survey: Qualitative Quantitative • Outcomes trials • Clinical trials	• Comparitive effectiveness of techniques in specified conditions • Time to achieve specified results • Time effects last • Practitioner effects	• Practitioners • Researchers • Third Party Payers • Policy Makers • Medical Scientists • General Public
3: COMPARATIVE EFFECTIVENESS	• Survey: Qualitative Quantitative • Outcomes trials • Clinical trials	• Comparitive effectiveness of MBT vs other medicines	• Practitioners • Researchers • Third Party Payers • Policy Makers • Medical Scientists • General Public
4: PHYSIOLOGICAL EFFECTIVENESS	• Laboratory research • Animal Models	• Biological mechanisms underlying clinical effectiveness and practices	• Practitioners • Researchers • Third Party Payers • Policy Makers • Bioscientists • General Public

[3]Types of researchers for Domain 1: anthropologists, demographers, economists, epidemiologists, historians, philosophers, psychologists, sociologists...; for Domains 2 & 3: practitioners and supporting social and bio-science researchers; for Domain 4: physiologists and similar bioscientists.

[4]For definitions of research types, see Appendix 1.

animal research models, but to get to work doing basic research about features of MBT practice that make it distinctive (domain 2) and motivate people to study it or seek it for healthcare (domain 1).

The reason for this counter-cultural (and therefore counter-intuitive) recommendation is simply this: if MBT researchers wish to design research that meets the many criteria of high quality science, they must know exactly how to answer basic issues that emerge at the outset of research:

- how to identify an appropriate sample and sample size,
- how to define an intervention that is measurably different from another intervention,
- how to know if the time given to the intervention is appropriate,
- how to know which intervention is likely to stack up well against a pharmaceutical intervention
- what kinds of people seek MBT care
- what kinds of people deliver MBT care
- whether non-MBT trained people can take the place of MBT-trained practitioners (this has been suggested as a 'control' feature)
- what features of MBT practice absolutely cannot be left out of a test and still claim it is a true measure of MBT practice.

These issues are fundamental to research design. Thus, at present, in the absence of trustworthy answers to these issues, it is difficult to create valid research designs. Most "alternative" practices lack such data, and alternative medicine researchers frequently run into problems caused by its lack. Another way to put this is: MBT needs to know more about itself scientifically before it can accurately design research to compare itself to others. The issue of validity is taken up in more detail in the next section.

[5]A similar approach to indentifying research questions, but without the logic offered here, can be found in Foundation for Integrated Medicine 1997.

Table 2: Four Fields of Research Questions: Examples of Important Researchable Topics

Domain 1: Sociocultural Mechanisms of Effectiveness

1. Who uses MBT? (= demographics of use)
2. What benefits do they say they receive from it?
3. What features of practice keep them coming back for more? (= marketing issue)
4. How much do they pay for care? Are they satisfied with this? Want what to change?
5. How satisfied are they with their practitioners? The care setting?
6. For what complaints and conditions do people seek MBT? (= epidemiology)
7. What practices (types of massage, bodywork) within MBT do practitioners employ, how often, why?
8. Where do practitioners think the field is going?
9. What is the range of opinion about third party payments, fee for service, etc.?
10. What public outreach approaches have been most effective in raising awareness in a positive way about MBT, which have not been effective?
11. What is the professionalization process in MBT?
12. What is the history of the MBT field, especially with regard to the development of particular intervention techniques, and to professionalization?

Domain 2: MBT Practice Mechanisms of Effectiveness

1. How much time must a person experience a particular intervention to experience relief of symptoms? 15 minutes? 30 minutes? 45 or 60 minutes? Once a week, twice a week, twice a month? Relief that lasts one day, two days... two weeks or more?
2. How does the application of an intervention by one practitioner differ from the same intervention applied by a different practitioner? How much inter-practitioner reliability is there in the delivery of care? What does "good" practice look like, feel like; how can it be measured?
3. Which MBT approaches are most effective for which conditions? Are there any popular interventions which should actually be contra-indicated?

Overlap Issues in Domains 1 and 2

1. What is the explanatory model of MBT... especially as the practices that fall under this rubric are so variable?
2. What attracts people to practice MBT, and what keeps them practicing?
3. What features of the education serve practitioners well, not so well?
4. What clinical perceptions do experienced practitioners have and how do they apply them?
5. What sorts of people respond best to particular interventions?
6. What sorts of people do best as practitioners?

Domain 3: Comparing the Effectiveness of MBT to other medical practices

1. For which conditions is MBT therapy as effective or more effective than the standard biomedical intervention for that condition?
2. The same question with MBT compared to other medical systems.

Overlap Domains 1 and 3

1. What is the comparative cost-effectiveness of biomedical and MBT interventions for a particular condition?

Overlap Domains 2 and 3: Working Toward the Level Playing Field

1. What distinctive medical insights and intervention techniques has MBT to offer biomedicine?
2. Reverse of Q1
3. How can the two medical practices combine forces so as to serve patients better?
4. The same set of questions applies to the relations of MBT with any other medical system.

Domain 4: Physiological Mechanisms of Effectiveness in MBT

1. What are the physiological features underlying the effectiveness of MBT practices? Of the practices that involve soft tissue manipulation? Of those that involve bone manipulation? Of those that enter the energy fields and change them?
2. Are current observations about mechanism accurate? Sufficient?
3. Does MBT have the same or similar effects in non-human animals? Why?

ISSUE 3: MAKING RESEARCH DESIGNS & DATA CREDIBLE: ON CRITERIA OF SOUNDNESS, WITH A SPECIAL FOCUS ON MODEL FIT VALIDITY

Science is concerned with the credibility or believability of data. In contrast to other ways of gathering information (more details in Appendix 2) science wants to ensure that what it learns is trustworthy. It does this by gathering data very carefully. It tries to identify sources of bias (error) and control or minimize these. It measures things time and again, both testing the quality of the measurement instruments and the quality of its conclusions. It insists that what it knows is relative and approximate—that at any time a new technique, technology, or idea may make preceding data obsolete. It also insists that the effects of the personal—desire, preference—be minimized; interpretations are to arise from evidence, not from faith or fiat. Out of all this effort should come useful data, defined as data that is likely to apply in many different situations, and answer to many challenges.

The effort to minimize bias demands, of course, that bias first be recognized. This task is on-going—it applies to every researcher and to every research project. Fortunately, however, scientific research has been practiced long enough that many sources of bias are well known, and there are well established ways to control their effects. Table 3 lists some familiar criteria of soundness (more detail in Appendix 3). Note: Quantitative and qualitative sciences often use slightly different terminology—in the list I have more or less combined sources. To learn more about all of these except model fit validity, consult tests such as Bernard 1998, Brink & Wood 1988, Carmines & Zeller 1979, Creswell 1994, Kirk & Miller 1986, Marshall & Rossman 1989, Morse & Field 1995. The concept of model fit validity is discussed below.

Table 3: Criteria of Soundness

Precision: Does the measurement instrument measure at a scale appropriate to the question asked?
Reliability: Does the measurement instrument get much the same type of answer every time it is used?
Transferability: Does the measurement instrument function equally well in a locale different from the one in which it was developed?
Credibility/Validity: Does the research design and/or the measurement instrument actually measure what it intends to measure?

- **Face Validity**:
  Do operational indicators of a concept "make sense?"
- **Internal Validity**:
  Is the proposed experimental intervention capable of or likely to create or measure an observable or significant difference?
- **Construct/Content Validity**:
  Do measurement instruments measure what they are intended to measure?
- **Statistical Conclusion Validity**:
  Are the statistical tests applied appropriate and are the statistical conclusions credible?
- **External Validity**:
  Do items measured translate to issues that matter in the "real world"?—Are they representative, are they generalizable?

MODEL FIT VALIDITY: Are the assumptions underlying the design well understood and factored in to the design so that the resultant data accurately represent the people or practice or intervention being tested?

Model Fit Validity: All the criteria of soundness listed except the last are applied during the process of doing and analyzing research. But well before researchers get to the point of doing research, they must design the research. To do that demands making decisions concerning reality and "what matters" which help one know which questions are "worth asking". In short, as MBT researchers examine and critique existing research designs and existing research instruments (most of which have been developed according to the criteria characteristic of biomedical philosophy), they should be asking "Will these serve our purposes?; Do these designs fit the delivery characteristics of MBT? Did the maker of the instrument mean what we mean by these questions? Was something essential to our understanding left out? Has something essential to MBT been misrepresented?"

To get at what is "worth asking" and if existing designs and instruments "serve our purposes" researchers must know a lot about the philosophical, theoretical or paradigmatic underpinnings of their profession and their science. Being able to mesh the profession's model of health care reality (explanatory model) with the research design so that the two dovetail is the essence of achieving model fit validity.

As noted in the discussion of Issue 1, all health care systems are guided by distinctive health care models. Science is also guided by culturally-mediated ideas and ideals. The basic message is that the task of ensuring model fit validity is of special concern for MBT because it is popularly defined as "alternative," and self-defines as "holistic."

The usual sequence of behaving in the world goes something like this:

assumptions —> logical strings —> behaviors —> outcomes

To gather scientific knowledge:

assumptions—> logical strings—>research designs—>data

The implication of the flow from left to right is that everything to the right is dependent upon the quality and character of what happens to the left: indeed, the data that emerges from scientific research is credible only so long as the preceding steps are coherent and account for bias.[6]

Most criteria of soundness address error as it may emerge at the research design step in the sequence, that is, well down the path from the explanatory model. Model fit validity addresses errors that may emerge at the beginning of the sequence:

assumptions—> logical strings—>research designs—>data

model fit validity other criteria of soundness

One of the tricky aspects of identifying and accounting for assumptions is that very often we simply do not know that they are operating: as normative aspects of our cultural and occupational background, they are out of our awareness. To take an example: a common

[6]The folk summation for this well-known situation is GIGO—garbage in-garbage out.

assumption in massage and bodywork therapy is that touch is "good" for people. Of course any MBT practitioner can cite instances when touch was aversive—but that is not the point. The point is that deep within the practice of MBT is an accepted idea, an idea, indeed, that it would be difficult for MBT practitioners to give up without at the same time giving up their profession. This idea must be factored in to all research designs... yet, in all probability, this idea is taken so much for granted by MBT practitioners that it would hardly occur to researchers to consider it an issue in design.

As the word equation shows, out of assumptions come strings of logic which lead to an outcome—action, data, and so forth. We can summarize such strings as "if-then" sequences. Continuing with the example above: If touch is good for people, (and I like people), then I shall touch them. In writing down this sentence I have had to add another assumption in parentheses to explain motivation. This assumption is shared by virtually all health care practitioners all over the world: they want to help people. Again, one might hardly think that this would need saying, but of course there are professions in which helping is either irrelevant or counterproductive.

The point for our purposes is that **in the practice of scientific research, the helping function of health care has sometimes been factored out of research designs. In the design of MBT research, the skill in touching that is developed by the professional practitioner has sometimes been factored out of the research design.** These are errors of **model fit** that could not have occurred had researchers paid sufficient attention to identifying assumptive patterns and to factoring these **in** to the research design.

On Science, Models and Paradigms: Errors of model fit are common in cross-cultural or cross-paradigmatic research. MBT practitioners are involved in that kind of research because, for better or worse, the profession is socioculturally defined as "alternative."

An important implication of the earlier statement that scientific knowledge is relative is that (all) knowledge is constructed. Indeed, the implication of the concept of culture is that one's reality is constructed. Using the word equation above, we could show that each set of assumptions yields a somewhat (or markedly) different set of logical strings and fuels different behaviors; and each set is a construct that represents a selection from all possible assumptions and behaviors. Our way, whatever it may be, is not the way, but is simply one way among many, a model of reality. In short, science deals in models (not in truths).

Here are some further implications:

- If an assumption changes, the subsequent logic, design and data will also change.
- Since it represents a selection from all possibilities, one person's or occupation's model may not make sense to another, or may not suit another.
- If a model can be constructed, it can also be deconstructed.

These points have profound significance not only for the practice of science, but also for the practice of medicine.

Metaparadigm: The models by which we live our lives are sometimes given special names. The largest scale of model is the metaparadigm. This term refers to sets of overarching assumptions that guide the deep (often unconscious) structures of whole societies, over centuries-long periods of time. In U.S. society we are guided by two metaparadigms, the

reductionistic and the *holistic* or *relational*.[7] Because the reductionistic metaparadigm is dominant, those who organize their behaviors according to the premises of the second metaparadigm are—by definition—"alternative" ...and puzzling to those who reason exclusively in terms of the dominant metaparadigm.

Every profession is also guided by a smaller scale paradigm, called an occupational paradigm. As mentioned in Issue 1, if someone who identifies with a particular profession tries to explain what really matters to that profession, what makes it distinct from other professions, and how it "works" in a philosophical or applied sense, then that person is explicating the explanatory model of that profession; from within the profession the part of the explanatory model that is in-the-awareness of users is called theory.

Occupational paradigms take their cues from metaparadigm: there are "reductionistic" practices of medicine and "holistic" practices of medicine, and there are also "reductionistic" and "holistic" ways of practicing science. Since the reductionistic model is dominant in our society, so also are reductionistic ways of practicing medicine and science. Table 4 lays out some of the premises that underlie the two ways of practicing science. Which approach may better represent MBT?

Although in the "real world" the distinctions between reductionism and holism are not as absolute as is suggested by Table 4, and even though it is reductionistic even to create a listing of this apparently oppositional nature, it is worth doing as a shorthand for exploration of the difference metaparadigm makes. Please note, however, that judgement does not play a role here: neither metaparadigm is "better" or "worse;" rather, they exist and they must be taken account of in the careful practice of science.

Table 4: Two Ways of Practicing Science Contrasted

Reductionistic Tendencies	Holistic Tendencies
focus on singularity	focus on complexity
boundaries	connections
entities, objects, categories	context, relationships, networks
either/or thinking; opposition	both/and thinking; complementarity
singular expertise concept	collaborative expertise concept
progress concept	balance concept
activism & future orientation	presentism & contextual decision making
interest in quantities	interest in qualities
analysis (excluding)	synthesis (including)

Reductionism is fueled by the tendency to analyze, or separate things into component parts, an energy that lends itself to an interest in numbers and quantification. The links that hold the parts together are of less interest than the parts, so that as the process of reductionism proceeds

[7]There are many names for these metaparadigms. Other names for the reductionistic include: deterministic, categorical, hierarchic, patriarchical, ontological. Other names for the holistic include: relational, feminist, synthetic, ecological, physiological, and so forth. For more on metaparadigm and medicine, see Cassidy 1995 and the references quoted therein.

there seem to arise a larger and larger number of objects, entities and categories, and the intellectual task is one of recognizing the boundaries that distinguish one thing from another. In the absence of a focus on connectivity, the resulting entities and categories come to seem more and more to exist on their own, relatively unchanging and materially real. These many objects and entities can be sorted and grouped; sorts are normally arranged vertically as in the systematic table of species. However, though hierarchies can serve mainly descriptive purposes, many actually represent hierarchies of value, thus reductionistic systems commonly concern themselves with issues of truth and heresy. Emphasis on distinction leads to a kind of competitive and judgemental opposition in which the participants try to determine which categories are "best" (by some measure); those that don't "measure up" are excluded from the mainstream and fall to the periphery where they may be forgotten (or safely ridiculed) by the orthodox. In such an hierarchic system, those who know more about what is "right" are defined as experts whose opinion should count more than the opinions of non-experts (for example, "doctors" know more than about what "counts" in health care than their "patients"). Finally, since things can nearly always be further subdivided (reductionism), and each such division and resulting product must be evaluated (judgementalism), applying the reductionistic metaparadigm yields an unending supply of issues to assess and always the potential of learning something new: this in turn leads to a focus on the future (with its promise of "progress") and to a desire to be activist about seeking novelty and "improvement."

Holism, in contrast, is fueled by a fascination with the linkages and relationships among the apparently disparate—here the goal is to pull together the parts and observe "the whole that is greater than the sum of its parts" (synthesis), or, simply not to separate things in the beginning but to try to understand them as they are, whole and complex. This "whole" is very often not really explicable or graspable—it seems to hover just out of reach, a characteristic which does not make it any the less real to its perceivers, but does emphasize the importance of point of view, and thus the dynamic nature of the whole. Holism is far more relativistic—deep in its bones—than reductionism. Rarely are constructs, observations or objects deleted from an holistic system—instead, space is made for them, little judgmentalism is applied to making decisions (there cannot be heresy in an holistic system), and holistic thinkers ask utilitarian questions such as "under which circumstances will this idea/approach work?" The fact that it doesn't work under all circumstances doesn't matter—doesn't even seem relevant, since holism assumes that reality is infinitely complex and the task is to find points of balance or efficiency rather than "truth". Indeed, holistic thinkers are encouraged to enjoy ambiguity, to be flexible, or to put it proactively, to familiarize themselves with many models of reality and determine when to apply them to the situation at hand ("surf paradigms"). This point also means these thinkers tend to focus on the present and on the quality of events or actions, for they argue first that the future is not here yet, and second, that it must be much like the past since complex systems, though dynamic, rarely change much in their essential character. Thus rather than emphasize measurement, holistic thinkers emphasize connectivity; this translates to a tendency to prefer qualitative over quantitative data, and to prefer horizontal teamwork designs—"everyone has their own expertise"—over hierarchies of value.

Metaparadigm ultimately guides the design of research. Research fueled by the reductionistic metaparadigm focuses on analysis and quantification, and on asking questions such as "which does x best?". Holistic research focuses on studying linkages and networks, the flow of information, and on asking "what maximizes the responsiveness of x?"[8] Both approaches are

[8]This is a version of the question "Does it serve?"

useful; t**he tasks for the researcher are to know which approach one is using, and which better serves the current purpose.** *To use paradigm so as to represent one's practice accurately is to achieve good model fit.*

Returning now to Issue 1 for a moment: which of the two metaparadigms is more characteristic of MBT? The answer is (probably) *holism*: MBT combines many disparate practices under one roof, and most practitioners know several different forms of practice; MBT creates close connections to patients when practitioners "read" the bodies of their patients as they put their hands on/near them; MBT practitioners tend to create "teams" or (otherwise stated) establish fairly horizontal and interactive relationships with their patients as people, assuming that patients have a part to play in the origin of their malfunction, in the response to treatment, and in prevention of future malfunction.

The bottom line in this discussion is this: MBT researchers must pay special attention to ensuring that the unique character of MBT is secured within each and every research design so that these designs achieve high model fit validity. Assuming that the research is otherwise well-designed, this will in turn ensure that the resulting data accurately reflect the capabilities of MBT. To the extent that MBT is an holistic practice rather than a reductionistic one, it must accept biomedically-promoted designs only with care and attention to their actual utility for MBT. The next two sections consider the usefulness of popular research designs with the issue of model fit validity in mind.

ISSUE 4: QUALITATIVE AND QUANTITATIVE DESIGNS AND MBT

As noted in Issue 2, the basic questions "Does it work?" and "Does it serve?" are easier to grasp if subdivided into four domains which use different approaches to assess effectiveness. For planning the future of MBT research, my prioritizing recommendation is:

First: **Focus on developing data on sociocultural effectiveness and within-practice effectiveness** (Domains 1 and 2) because a) these provide data essential to achieving high model fit when working in Domains 3 (between-practice effectiveness) and 4 (physiological effectiveness); b) will enhance the sense of internal coherence of the profession, and c) will produce data of interest to the widest audience including users and providers, funders, educators, and third party payers.

Second: **Develop knowledge of between-practice or comparative effectiveness** (Domain 3) because this will help "level the playing field" with biomedicine and improve opportunities for patient care collaborative relationships.

Third: Develop knowledge of physiological effectiveness (Domain 4) because knowing "how" it works may open up some new intervention possibilities; for some it will also enhance the credibility of the MBT profession.

Domain 1 or sociocultural data is commonly collected using in-depth or survey designs; survey techniques are also commonly built in to clinical outcomes and trials designs. This section reviews a few relevant points concerning these methods; for detail see texts such as Bernard 1998, Denzin & Lincoln 1994, DeVellis 1991, Feldman 1995, McCracken 1988, Miles & Huberman 1994, Schuman & Presser 1996.

In-depth qualitative research: The focus of this form of research is on ***detail*** and on ***explanation***. Interestingly, only qualitative and laboratory research can address the question of "why", that is, provide explanation. Other forms of research are descriptive or predictive, but not explanatory (more detail in Appendix 4). Another crucial reason for engaging in qualitative research is to construct quantitative questionnaires with high model fit validity.

Qualitative research consists of a series of techniques that gather data from individuals and allow them to reveal their thoughts, beliefs, models, assumptions, motivations, passions, reasons and reasoning; there are even techniques that allow people to talk about their unconscious knowledge, something that *cannot* be done with quantitative techniques. Most of these techniques involve either interviews or written responses, or the use of various "games" that let people reveal knowledge in novel ways. Today there are a variety of excellent software programs to ease the analysis of qualitative data.

A major use to which MBT could put qualitative research is to find out about practitioner and patient *motivation*. For example, why do people study MBT, what satisfies them about its practice, what compels people to seek out MBT, and what value do they receive that causes them to keep coming back? These topics all concern perception, experience and satisfaction; results of such research can be used for marketing, but equally, they can be used to improve student assessment, teaching, delivery environments, and patient satisfaction. Additionally, the "stories" people tell about their lives as practitioners or their experiences as patients are powerful motivators for many purposes: for funders, for people considering MBT, and for practitioners.

There is a long habit of denigrating qualitative research as "anecdotal." This usage merely reveals ignorance of research methods. An "anecdote" is, formally, a single story, usually without context, which is told to create some sort of sensation or make a single point—it does not fulfill scientific criteria, though it may galvanize an audience. In contrast, qualitative researchers are interested in samples collected with appropriate attention to sample frame, and carry out formal and systematic analyses of their data. Qualitative researchers commonly seek not only to know the range of opinion or experience on a subject, but also the relationship of theme to outcome (for example, do people who claim a degree of "responsibility" for their health have better outcomes, compared to those who do not make this claim?). Qualitative researchers also pay attention to factors such as the language people use to express their perceptions, and the ways in which the public view of a medical topic differs from the professional view.

Many qualitative researchers also create survey or quantitative questionnaires. Starting from an assumption that what they know (as researchers) is not necessarily in sync with what their target audience knows, they do a small scale qualitative assay of the target audience to find out what matters. Out of these data they develop survey questions that "make sense" by reflecting the audience's issues and language into the questionnaire. This process creates a questionnaire with high model fit validity: such a questionnaire commonly has higher response rates and (assuming it is otherwise well designed) should yield valid and credible data.

Unfortunately, very often questionnaires are designed without the preceding qualitative step, a practice that can easily yield an instrument with low model fit validity. Consider a case from my experience: A national mental health survey instrument used by the U.S. government was producing untrustworthy results. A group of anthropologists was asked to use qualitative

research techniques to find out what might be wrong. To make a long story short, what had happened is that somewhere along the line quantitative researchers had constructed questions that made no cultural sense to their audience, questions like "In the last two weeks, how often have you felt blue or depressed?" My qualitative research showed that "blue" is a term that is both age and race specific and refers to a mild degree of emotional discomfort, while "depressed" is a relatively frightening term that refers to a degree of discomfort serious enough to demand visiting a health specialist. Putting the two words together in one forced choice format was confusing, and required people to respond to one or the other word. Since out of thousands of massed responses there was no way to know to which word any given person had responded, the statistical results were uninterpretable. This "just-so" story for questionnaire makers is included to emphasize this point: *creating culturally astute surveys (surveys with high model fit validity) virtually demands an initial qualitative step*. That is a good point with which to leave the subject of qualitative research: it is best done at the initiation of research, and it is also best done whenever the goal is to understand "*why*".

Survey Design Issues. Survey designs are popular and can provide very useful descriptive data. One issue in the construction of new survey instruments is noted above. Below I discuss two points: the way questions are asked, and the utility of existing standardized survey instruments for the use of MBT researchers.

1. One can ask questions on surveys in many ways including
    - open-ended question... or writing in one's own answer (a qualitative technique)
    - free listing or forced listing
    - forced choice
    - scales such as the Likert pattern or the Visual Analog Scale.

    Each of these approaches has its advantages and its disadvantages. The advantages of open responses—the open-ended questions, the free listing—is that respondents can say what most matters to them, and can do so in their own language; the disadvantages include longer time to analysis, and receiving "top of the mind" responses which may or may not reflect the respondent's whole attitude to or knowledge of the subject.

    Forced answer questions are popular because they are easy to analyze and produce numerical answers, but they do not allow respondents freedom in expressing themselves and again, one cannot know if the respondent is saying all s/he knows or feels about the subject. More importantly, the utility of forced choice questions is extremely dependent on the relevance of the question being asked. In other words, if questions don't have high model fit validity, the results won't be trustworthy. Questions become more useful to the extent that the content has been gathered via qualitative research, and to the extent that the language reflects that of the target audience. There are, of course, many other issues to consider in designing survey questions—sequencing, internal accuracy checks, etc.—far too much for a short paper, so once again I refer interested readers to the literature.

2. There are a vast number of standardized medical survey instruments already on the market (MdDowell & Newell 1996, Ogles 1996). Many survey researchers are tempted to use them because it's simpler (short-cutting the process of creating a questionnaire) and because one's results can be directly compared with others' published results. These are excellent reasons whenever the survey instrument is actually appropriate for the intended use. But MBT researchers should use caution in accepting the utility of existing survey instruments,

always asking themselves, "Do they ask questions that make sense within the explanatory model of MBT? Are their issues our issues?"

For example, many existing instruments are intended to assess responsiveness to care for a specified condition, such as HIV or arthritis or migraines. Though familiar, all three of these are actually biomedically-defined disease entities—that is, they are cultural constructs. They may not "make sense" to MBT, and (more than that) treating these conditions may not fall within the occupational or legal purview of MBT, since MBT practitioners are not biomedical practitioners. Therefore, the first thing an MBT researcher must do when assessing the utility of a survey instrument tied to a biomedical disease label is back up and examine not the label, but the associated symptoms—what MBT therapists may actually be treating. What now do these conditions become? Do they become fatigue? Low mood? Joint pain? Head pain? Ocular oddities? Dizziness? And if they are so transformed, what has happened to the utility of the survey instrument for MBT?

This case of an instrument guided by a biomedically-defined condition is fairly transparent. More subtle are cases in which the instrument purports to measure, say, quality of life, yet either fails to measure issues of importance to MBT (e.g. responsiveness to touch), or asks questions that reveal a different take on a subject than is typical of MBT. Suppose an MBT researcher wants to know if an existing instrument to measure QOL and "pain" is useful for MBT. Besides the usual focus on time, duration, location, and severity, MBT researchers might wish to ask questions such as: Does the instrument measure the movement of the breath? Groundedness? The ability of the patient to be present in the body or in specific parts of the body? Adaptation to pain?

MBT researchers must locate the points of strain within existing instruments on a case by case basis. However, to illustrate the point further I offer an example cogent for acupuncture. Acupuncture practice is based on a perception that imbalanced energy flow is the root cause of malfunction—this energy, though not formally defined as yet, is certainly more than the popular meaning that lies within such phrases as "I had lots of pep today" or "I am not as tired as I used to be." Yet, these are the only measures of "energy" offered on a highly popular standardized quality of life survey questionnaire. The fact that "energy" is measured in so limited a way on this questionnaire is not surprising: it was written by biomedically-astute researchers, not by Chinese medicine-astute researchers, and it reflects the priorities and understandings of its makers. But *every* acupuncturist considering the use of this questionnaire would need to decide if the rest of the questionnaire was sufficiently useful to excuse its inadequacy on the subject of energy.

In this example, the acupuncturist could solve the problem of communicating with the scientific world by using the familiar standardized survey instrument, while at the same time achieving higher model fit by constructing a questionnaire to measure energy as perceived by Chinese medicine and running it alongside the standardized questionnaire. This technique is one MBT practitioners may also wish to use.

ISSUE 5: CLINICAL OUTCOMES & TRIALS DESIGNS AND MBT

The technical difference between a clinical outcome study and a clinical trial study is simply this: the outcomes study endeavors to *measure the **effectiveness** of care in naturalistic settings*

such as a clinic or office, and thus allows for bias introduced by the "normal and usual" behaviors of practitioners and patients; the trials study *defines very strict limits to measure the* ***efficacy*** *of an intervention*. For example, a clinical trials study of a weight loss protocol for malignant obesity might confine participants to a hospital ward for the duration of the project, while a clinical outcomes study of the same weight loss protocol would ask participants to come to an office or other setting to receive care. In the first, the participants would be prevented from carrying out behaviors (such as snacking) that might interfere with interpretation of the specific effects of the intervention; in the second participants would be asked not to break the protocol but everyone would know that they might, so researchers would also gather data on "how difficult it was" to follow the protocol. The trials design is useful to measure highly specific effects; the outcome design is useful to measure effects in a lifelike setting where people make their own decisions. Currently the outcomes approach is more popular both because it is somewhat easier to carry out, and because its results more closely reflect probable usage of an intervention in the "real world." The outcomes approach is the one most likely to serve MBT practitioners.

Components of Experimental Research Designs. A technical language guides the design of experimental research. Table 5 (next page) provides a list of terms selected for their relevance to the issue of model fit validity. More information is given in Appendix 5, and for detail, consult methodology sources such as Cassidy, Cassileth, Jonas et al 1995, Hammerschlag 1998.

Non-specific Effects: The term placebo has been used and misused in many ways. When it is used as a noun—a placebo—it generally refers to a pharmaceutical or other intervention that is intentionally pharmaceutically inactive. Placebo drugs—made to resemble the test drug—are used in pharmaceutical research, and in this case the goal is to measure the difference between the placebo response and response to the test drug. Placebo substances may also be offered to a patient in a clinical setting if the clinician feels that the patient will respond with improved symptomatology and "does not need" an active substance. This use—to "please" patients—has won a reputation for the term placebo that is muddied both by the image of the obstreperous patient and the authoritarian practitioner.

In fact, however, *every patient-practitioner relationship* is affected by something commonly called "the placebo effect", and which anthropologists call expectation. When patients expect to get better, practitioners expect to benefit them, communication is effective and / or the health care setting is supportive, then a degree of healing occurs even if nothing else "active" is offered. Efforts to measure the placebo effect have shown it to account for 30-80% of response in various studies (Moerman 1996, 1998).

Although it is common to talk of placebo as "inactive"—a term borrowed from pharmaceutical research—it is much more accurate to speak of placebo effect as "nonspecific" since it is clearly active. Thus we may see the response to care as consisting of several additive parts: the nonspecific effects of expectation that one will receive (deliver) healing, the semispecific effects of simply offering hands-on care, and the specific effects of particular interventions. The actual goal of clinical research is to segregate these effects so as to measure the specific effect of particular interventions.

In most scientific research—that fueled by the reductionistic approach—designers have tried to *minimize* the effects of expectation and preference on the experiment, in hopes of being able to measure the specific effects of the medical intervention (most of the design mechanisms in

Table 5 represent that effort). Holistic researchers, with their preference for naturalistic or "real world" designs, are more likely to suspect that it isn't really possible to delete the effects of expectation and caring from research designs, and therefore suggest it might be equally appropriate to attempt to *maximize* the effects of placebo; differences among specific interventions would, in that case, still represent real differences in the utility of the interventions.

Table 5:
Common Design Features in Clinical Experimental Research Designs

Nonspecific Effects

Placebo: Positive effects of expectation and belief, e.g., expectation that the health practitioner's intervention will be beneficial

Nocebo: Negative effects of expectation and belief, e.g., distress from believing that one is receiving or delivering "worthless" care

***Controls:* any design feature intended to minimize the effects of chance or preference upon the test measures**

Assignment: process of allotting potential participants to a research project

Randomization: potential participants are allotted randomly to the test groups

Matched Pairs: participants are matched (sex, age, complaint) and then (randomly) assigned to different test groups

Choice: potential participants are allowed to choose which test intervention they would prefer to receive

Blinding Control: Participants are purposefully kept from knowing what part of the research they are participating in.

Single Blind: 1: only the patient is unaware if they are receiving the active, placebo, standard or test intervention 2: both patient and practitioner know what the patient is receiving, but the research analysts do not know (= "blind assessor" design feature)

Double Blind: Neither patient nor practitioner know if they are receiving/delivering the active, placebo, standard or test intervention

Placebo Control: an intervention that looks (tastes...) much like the "active" intervention but that is intended/assumed to be inactive (nonspecific)

Sham Control: in hands-on practices, a physical intervention that is intended/assumed to be inactive (nonspecific)

Protocol Control: an intervention design is predetermined and practitioners are not permitted to alter it (use their clinical judgement) during the research.

Standard Control: an intervention that is recognized as the standard or usual-and-appropriate way to treat a condition.

Standard Care: practitioner uses their own clinical judgement and delivers their standard care to the patient

These points represent one aspect of the real philosophical "continental divide" between the two ways of practicing science. Since holistic research is still in its infancy these remarks are also "cutting edge." Nevertheless, achieving high model fit validity implies that the problems

of doing holistic research must be faced by holistic medical practices: MBT researchers will—at least part of the time—be working at the growing edge of science.

The term **nocebo** refers to the negative effects of expectation—if a patient fears an intervention, or a practitioner doesn't wish to deliver an intervention, then nocebo can act, and a less-than-expected healing effect can occur. Subtly, but extremely important for the design of research, nocebo effects occur *if a practitioner is asked to do something that works against his/her desire to offer help and healing*. For example, if as part of a research design a practitioner is asked to deliver an inactive, inadequate, or inaccurate intervention, that practitioner may find it "difficult" to do, or may "feel bad" about doing it. This is unfortunate enough, but because a *relationship* exists between the practitioner and his or her patient, the patient also suffers from nocebo—trust and the healing response both become muted.

The nocebo issue is rarely discussed in traditional research sources. Many resist the implications of nocebo—which loom large as problems in placebo—and sham-controlled research designs. The issue is, however, usually clear to holistic/relational researchers who therefore critique and avoid research designs that ask practitioners to behave in ways which compromise their desire to offer help and healing.

Controls: A major concern of outcomes and trials research is to establish controls to help minimize bias. There are many types of controls, some of which are listed below. They vary in their utility and ease of application, and of course, they also vary in their attractiveness to reductionistic and holistic researchers.

The first control enacted in most clinical research involves ***assignment***, or the process of sorting participants into the two or more "arms" of the research design. For example, if one is comparing two forms of MBT intervention on low back pain, which participants shall be assigned to receive TuiNa and which shall be assigned to receive deep tissue massage? The actual answer to such a question is guided by the research design one has selected. However, there are three basic possibilities: randomization, matched pairs, and choice. In the case of randomization, the participant has no say in which treatment s/he will receive. Based on identification number, sequence of intake, or some other mechanism, the participant is simply assigned to receive one or the other. The practitioner also should have no say in the randomization process—assignment is strictly by a pre-designed randomization mechanism. The goal is to minimize the bias introduced by perception and expectation by minimizing the ability of the participant to express preference and the practitioner to express clinical insight (such as opinion as to likelihood of this person responding to this intervention), so that differences that emerge later on can be attributed as much as possible to the intervention instead of to human emotion.

In the case of *matched pairs* participants are matched according to predetermined markers (sex, age, race, complaint, chronicity...) and then one is assigned to receive one intervention and the other is assigned to receive the other intervention. The fact that these are similar people (by the measures selected) is intended to minimize bias introduced by difference; then, if one set of participants responds markedly differently, it is taken to indicate a significant difference in the effects or utility of the two interventions.

Choice involves allowing participants in the research to select for themselves which sort of intervention they want to try. This design doesn't minimize bias from perception, but on the

contrary, allows it to act—as it does in the real world. The argument is that if expectation (placebo effect) is a large part of the healing process anyway, why not factor it in to design? If difference in response still emerges, then that difference can be attributed to the intervention itself.

The issue of randomization and choice is another region of divide between reductionistic and holistic thinking in research. Reductionistic thinkers, seeking material results, and mistrusting imponderables like "opinion," "perception," and "clinical judgement" want to minimize the ability of these factors to act during measurement. Holistic thinkers, convinced that the actions of preference cannot be factored out of any interaction, prefer to create "naturalistic" experiments.

A ***blind*** control means that some participant to the research does not know whether s/he is receiving, delivering or assessing "active" or "placebo," "test" or "standard" care. The more famous blind design is called "double blind". Here neither patient nor practitioner know which intervention the patient is receiving. This design works best where the intervention is something that can be easily disguised, such as a pill. It is assumed that if neither patient nor practitioner know which substance the patient is taking, then expectation will work equally in all test subjects, and differences in response can be attributed to differences in the effectiveness of the test substances. It is virtually impossible to use double blind designs to assess hands-on medical practices.

A "single blind" experiment can take two forms. In one, the patient does not know which intervention s/he is receiving, though the practitioner does. This design is difficult to carry out because the practitioner's enthusiasm (placebo effect) or distrust (nocebo effect) for the test procedure is commonly "telegraphed" to the patient, obviating the "blind" status of the patient. In the second, both patient and practitioner know which intervention is in use, but a third party, the data assessor, does not know. In this case the data assessor receives data that has been coded to hide who has received what intervention. The assessor simply analyzes the data. Subsequently the code is broken, the data sets segregated and compared. If a significant difference emerges, it can be attributed to a real difference in the effectiveness of the interventions. *The single blind assessor design feature is a highly useful one for hands-on researchers.*

A ***placebo control*** consists in creating an "inactive" intervention—the placebo—the effects of which are then compared with the effects of an hypothetically "active" intervention. Placebo control works best when the intervention is something that can easily be disguised, such as a pill. The task of creating a truly "inactive" placebo is considerable—some substances used in the past such as sucrose and lactose have had untoward active effects in persons sensitive to sugar or to milk. Also, study participants talk—either directly or indirectly it is common for a large proportion of participants in placebo controlled studies to quickly discover if they are receiving placebo. Efforts to create "placebos" in hands-on practices have proven uniquely difficult since it is hard to imagine a "hands-on" activity that isn't active, even if it is less active than the therapeutic intervention. Finally, there are real ethical questions concerning the appropriateness of treating presumably ill persons with substances thought to have no formal effect on their complaint.

The ***sham-control*** represents one effort to overcome the limitations of the placebo design for hands-on practices. Here researchers set up a hands-on physical intervention that is intended to be inactive and compare its effects with those from an intentionally active intervention. An

example from acupuncture is needling points on the body that are not traditional acupuncture points, and contrasting the effects of this needling with needling traditionally known points. The same MBT use-limitations listed for placebo control tend to cling to sham-control: can any touch be inactive? How inactive? Is it ethical to offer what one assumes is inactive? Finally, both placebo-control and sham-control designs run a high risk of producing nocebo effects.

A ***standard intervention*** is one which the larger society has determined to be normative and standard care for some complaint, and is contrasted to a test intervention. ***Standard care control*** is, however, a design feature in which the participating practitioners are allowed to offer their patients exactly what they would offer them were they not participating in research. In other words, in a standard care design, the practitioner uses his or her clinical judgement in selecting interventions. This is contrasted with the ***protocol-controlled*** design in which a particular intervention protocol is predetermined and the practitioner must deliver that, and just that, during the period of the research.

Popular experimental research designs. The design features discussed above can be combined in various ways to produce discrete research designs. Table 6 lists several popular experimental designs; readers should remember that there are many other designs that apply when doing archival, qualitative, survey or laboratory research.

In each of the designs that follow, it is assumed that

a. Participants are carefully selected. They meet both inclusion and exclusion criteria which are prestated and fit the rationale of the research.
b. It is possible to employ random, matched pair, or choice assignment; to use longitudinal or pretest-posttest designs (cross-sectional designs are less common in experimental research; definitions of these terms in Appendix 5).
c. Length of test period, frequency of intervention, frequency of data collection, types of data to be collected, format of data collection, follow-up and so forth, are all predetermined.
d. Blinding can be utilized to minimize the effects of expectation. Blind patient and practitioner apply primarily to the first two designs, while blind assessor designs apply to the remainder.
e. Each design predetermines against what the test intervention will be compared—no treatment, placebo intervention, sham intervention, self-as-control, or standard treatment.
f. Each design clearly defines what the test intervention will be and the ways to measure treatment effects.[9]

Table 6: A Selection of Clinical Experimental Research Designs

Blinded Placebo Control
Blinded Sham Control
Cross-over Control
Wait-list Control
Standard Care Control
Adjunct Standard Care Control
Choice Comparison

[9]For guidance on the sequence of research, see Appendix 6.

The most famous experimental research design, developed for use in testing pharmaceutical drugs, is the double blind placebo-controlled experiment. In this design a test intervention (e.g., a new drug) is compared with a placebo intervention, and neither practitioner nor patient know which they are delivering/receiving. Statistical analysis is applied to determine if the test and placebo results are *significantly* different. Note that it is assumed that people will respond to both interventions, because the placebo response is activated in all healing situations. If the test intervention scores significantly higher than the placebo, the difference is taken to represent the relative success of the test intervention.

Can double blind designs be employed in research on hands-on medical practices? The general opinion on this question is "no." However, there have been efforts to do so, for example, to have a skilled therapist tell an unskilled one "what to do" and then withdraw while the care is being delivered. It should be clear that this approach has many limitations of precision, specification, and nocebo effects.

It is also possible to do a single blind placebo-controlled experiment. Recalling the discussion of single blinding above, note that in this process one must utilize a *blind assessor* design. As always, the placebo must be very carefully defined and designed so as to maximize its acceptability and the probability that it is really a placebo.

In the single blind sham control design, a sham *hands-on* intervention is employed and its effects compared with those of a test intervention. The sham could be a touch intervention delivered by the practitioner but which the practitioner does not think (for reasons that can be convincingly stated) will have the intended therapeutic effect, or it could involve using an electrical instrument such as a TENS device, allowing it to seem to be working but in fact, assuring that it does not deliver an electrical stimulus. Expectation is presumably activated as much by the machinery as by the hands-on manipulation so it is fair to compare patient responses. In this design the patient is blinded; it would be possible to superimpose a blind assessor as well, but clearly, the practitioner is not blind. Given the difficulties in defining a sufficiently "inactive" sham, and the ethical aspects of asking a practitioner to deliver false care, this design has not actually won many converts.

The cross-over design consists of testing two interventions sequentially in the same people. One half of the group is assigned to receive one intervention first, while the other half receives the other intervention. After a specified test period, the groups switch to receiving the other intervention. This design has the advantage of having people serve as their own controls—since they receive both interventions they have a direct opportunity to experience the difference of the two in their own bodies.

Between the two intervention sequences is a rest period that is called a wash-out period. The intention of this period is to wait long enough for the effects of the first intervention to "wash-out" so that when the new intervention is tried, it is tried on a clear field. This model works well for many pharmaceuticals where the rate of loss of a metabolite can be measured directly. It is more problematic for hands-on interventions. Exactly how do you determine how long the effects of massage last?

A wait-list control design consists of assigning some participants to receive the intervention now, while others wait a specified amount of time before receiving the same intervention. The research task is to learn how much the apparent activity of the test intervention is affected by

the tendency of much illness to resolve on its own. Researchers compare the change in symptomatology between those who are waiting for the intervention and those who are receiving it. If, say, 40% of those receiving care report disappearance of symptoms while 38% of those not receiving the care also report disappearance of symptoms, then the researchers conclude that the intervention is not having a significant therapeutic effect.

Notice that in the wait-list design no groups receive "no" intervention or interventions that are presumed to be nonspecific. Instead, those who are waiting for care serve as a "no treatment" control group until they begin to receive care. Often, those who began the treatment first and thus end the intervention first, continue to provide information to the researchers in a post-treatment period, helping to learn how long the effects of the intervention last once the intervention is stopped.

In standard care control design patients receive standard care from their practitioner, and the results of care among patients of different practitioners is compared. For example, patients receiving standard MBT care of lower back pain might be compared with patients receiving standard biomedical care of lower back pain. Standard care is defined as whatever is normatively done for patients with a given complaint. This design is extremely naturalistic: few strictures are put upon the practitioners, and patients know they are receiving normative care. The same sorts of measurement data must be collected from both groups, and blind assessor design used for analysis. Placebo and nocebo effects work as they do normally, as does clinical judgement.

One advantage of the standard care design is that one does not have to try to match the character of the interventions—that is, one can compare a pharmaceutical intervention with a hands-on intervention. If the response of the patients to the test intervention is as good as or better than to the standard, one can accept the test intervention as being as good as the standard. Thus the standard care control design lends itself to the comparison of "alternative" interventions with "standard" (biomedical) interventions. Another advantage is that the practitioner does not have to modify his or her normal care patterns, thus the potential for a nocebo effect is markedly reduced. The sociopolitical advantage is that when the outcome data indicate equal or better results for the nonbiomedical intervention, the "alternative" has taken a clear and measurable step toward establishing its effectiveness.

In the adjunct standard care control design, the same rules apply as above, with the difference that all participants receive the standard care, while only some receive the test care, or some receive the test care now and others later (wait-list feature). In this case the ethical difficulties are markedly minimized because everyone receives the standard care. The effects of the test intervention are then seen as additive—that is, as an adjunct to the effects of standard care. This design therefore does not result in a measure of the stand-alone effectiveness of a test intervention, but (sociopolitically) it does allow tests to take place that otherwise might not happen.

In choice designs, participants are allowed to choose which test group they would like to enter. Their progress in that group is then tracked exactly as if they had been randomly assigned. Choice can be used in any design that does not require blinding of the participant. It has so rarely been used that little is known of "what would happen;" however, the choice design has the advantages of mirroring the real world of personal decision much more closely than does

random assignment, and the effects of expectation can be charted as easily by their presence as by their absence.

One recent pilot study did employ the choice feature (D Eisenberg, personal communication 1999). A sample of people complaining of acute lower back pain was randomly divided into two groups. Part of the group was assigned to receive "usual" (standard) care, that is, biomedical care. The other part was offered their choice of biomedical, chiropractic, acupuncture, or massage care. All the groups were followed and assessed for clinical outcomes, satisfaction, and the total cost and utilization of services. The purpose of this research was not to compare the efficacy of the four practices as medical practices, but rather, to examine the effectiveness of an insurance eligibility intervention.

Experimental Designs Suited to MBT Research. Of the experimental designs listed in Table 6, 4 are particularly applicable to MBT research: **wait-list design, standard care comparison, adjunct standard care comparison, and choice comparison.**

Since MBT is a hands-on approach, it will always be difficult to create credible placebo or sham interventions, and all will suffer from ethical limitations. Similarly, double blinding is virtually impossible if a skilled practitioner is to deliver the care. Thus placebo and sham control designs are unlikely to serve MBT well. Cross-over designs may serve if researchers can credibly measure wash-out.

The other four designs are well suited to testing interventions that are hands-on and may not be viewed as normative. All allow the effectiveness of interventions to be measured (and sometimes contrasted) without putting practitioners in ethically difficult positions. All allow for single blind assessor designs. Randomization of assignment, or choice, can be employed; in fact both could be employed in a single design so that the additive effect of placebo could be directly measured.

SUMMARY RECOMMENDATIONS

In writing this white paper I was asked to include recommendations along with data and interpretations. These have been scattered throughout the paper; here I have drawn them together in summary form.

1. Science is powerful because it recognizes that knowledge and understanding are relative and changeable, and that bias is inherent. Scientific research cannot avoid bias—there are no perfect studies—it can only partially control for bias. The more sensitive researchers are to the subtle expressions of bias, the better they can minimize it, in turn improving the credibility of their data and, one hopes, their interpretations of that data.

   A succinct way of making the same point is this: **Science does not deal in truth or in proof; it deals in *evidence*.** The scientific task is to collect credible evidence and to interpret it equally credibly.

2. **Model fit validity** is a particular issue in the avoidance of bias that consists in being sure that research designs accurately reflect the explanatory or assumptive models of the practices or peoples that the research wants to understand. This form of validity has

rarely been discussed in the medical literature because most research, to date, has derived from the values of the reductionistic metaparadigm. However, as holistic practices of medicine increasingly seek to explore their practices via scientific research, the problem of achieving high model fit validity has emerged as central. To make the point using a daily-life analogy: to understand the game of baseball requires that one study the rules of baseball and assess the quality of the game according to those rules; this is an example of achieving high model fit validity. An example of low model fit validity would be if one tried to assess baseball by applying the rules of football.

MBT has been socioculturally defined as "alternative" and often self-defines as "holistic." For these reasons issues of model fit loom large for MBT and demand that its researchers proactively tend the problem of designing research that truly measures their practices' therapeutic capacity. In short, **MBT researchers are challenged to make themselves understood and to perform research *that meets the paradigmatic standards of their own profession***. Effectively engaged in cross-paradigmatic research, they must "go the extra mile" in taking care to develop the rationale of their research. However, by taking this extra care, they can avoid a potent source of bias, that of not accounting for paradigm in the design of research.

As noted above, several existing research designs will serve MBT research well. In addition, as a holistic practice, MBT researchers have the opportunity to help pioneer the development of new designs that function from the assumptions of the holistic metaparadigm, e.g., choice-based designs.

3. Doing research with high model fit validity demands that researchers know the "rules" of the practice they wish to understand, the "why" of each intervention, that is, in this case, the explanatory model of MBT. In addition, researchers must know a good deal about the shape of the practice out in the real world: who uses or delivers the practice and why, how many use it, for what conditions, and so on. **Currently there are shortfalls of knowledge in both of these areas for MBT**. These data are needed in order to build experimental, especially comparative designs (MBT vs biomedicine or physical therapy or...). Thus, **of the four effectiveness research domains identified in Issue 2, it is currently most critical to do research on the first two—measurement of sociocultural effectiveness and measurement of effectiveness components from within the profession.**

REFERENCES

Bernard, HR 1998 Research Methods in Cultural Anthropology, 3rd ed. Sage Publications, Thousand Oaks, CA

Brink, PJ and MJ Wood 1988 Basic Steps in Planning Nursing Research, From Question to Proposal, 3rd ed. Jones & Bartlett Publ, Boston.

Carmines EG and RA Zeller 1979 Reliability and Validity Assessment. Sage Publications, Thousand Oaks CA

Cassidy, CM 1998 Chinese Medicine Users in the United States, Parts 1 & 2. Journal of Alternative and Complementary Medicine 4(1): 17-27 and 4(2):189-202.

Cassidy CM 1994 Unraveling the Ball of String: Reality, Paradigms, and the study of alternative medicine. Advances, Journal of Mind-Body Medicine 10:5-31.

Cassidy, CM, B Cassileth, W Jonas, R Pavek, L Silversmith 1995 Appendix F: A Guide for the Alternative Researcher. In: Alternative Medicine, Expanding Medical Horizons, Workshop on Alternative Medicine (eds), US Government Printing Office, Washington DC

Creswell, JW 1994 Research Design, Qualitative and Quantitative Approaches. Sage Publications, Thousand Oaks CA

Denzin, NK and YS Lincoln, eds. 1994 Handbook of Qualitative Research. Sage Publications, Thousand Oaks CA

DeVellis, RF 1991 Scale Development, Theory and Applications. Sage Publications, Thousand Oaks, CA

Feldman, MS 1995 Strategies for Interpreting Qualitative Data. Sage Publications, Thousand Oaks, CA

Foundation for Integrated Medicine 1997 Integrated Healthcare, a Way Forward for the Next Five Years? London

Hammerschlag, R 1998 Methodological and Ethical issues in Clinical Trials of Acupuncture. Journal of Alternative and Complementary Medicine 4(2): 159-172.

Kirk, J and ML Miller 1986 Reliability and Validity in Qualitative Research. Sage Publications, Thousand Oaks, CA

Kleinman, A. 1980 Patients and Healers in the Context of Culture, An Exploration of the Borderland between Anthropology, Medicine, and Psychiatry. University of California Press, Berkeley

Lorig, K, A Stewart, P Ritter, V Gonzalez, D Laurent, J Lynch 1996 Outcome Measures for Health Education and other Health Care Interventions. Sage Publications, Thousand Oaks, CA

McCracken, G 1988 The Long Interview. Sage Publications, Thousand Oaks CA

McDowell, I and C Newell 1996 Measuring Health, A Guide to Rating Scales and Questionnaires. Oxford University Press, Oxford

Marshall, C and GB Rossman 1989 Designing Qualitative Research. Sage Publications, Thousand Oaks, CA

Miles, MB, A M Huberman 1994 Qualitative Data Analysis, 2nd ed. Sage Publications, Thousand Oaks, CA

Moerman, D. 1998 Medical Romanticism and the Sources of Medical Practice. Complementary Therapies in Medicine 6:198-202.

Moerman, D, et al. 1996 Placebo Effects and Research in Alternative and Conventional Medicine. Chinese Journal of Integrated Traditional and Western Medicine 2(2): 141-148.

Morse, JM and PA Field 1995 Qualitative Research Methods for Health Professionals, 2nd ed. Sage Publications, Thousand Oaks, CA

Ogles, BM, MJ Lambert, KS Masters 1996 Assessing Outcome in Clinical Practice. Allyn & Bacon, Boston.

Schuman H and S Presser 1996 Questions and Answers in Attitude Surveys.Sage Publications, Thousand Oaks, CA.

Stewart, AL and JE Ware, Jr., eds 1992 Measuring Funcitoning and Well-Being, The Medical Outcomes Study Approach. Duke University Press, Durham NC

Streiner, DL and GR Norman 1989 Health Measurement Scales, A Practical Guide to their Development and Use. Oxford University Press, Oxford.

Yin, RK 1994 Case Study Research, Design and Methods, 2nd ed. Sage Publications, Thousand Oaks CA.

APPENDICES

The following additional material is offered to develop some issues for readers who have relatively little background in science or in research.

APPENDIX 1: TYPES OF RESEARCH

Table 7 lists the main types of research. Any of these is appropriate to use in MBT research; the selection depends upon the questions that the researcher wants to answer. Each approach to research has certain advantages, certain limitations. Details about these methods can be found in any of a myriad of research texts.

Table 7: Types of Research

Archival-Historical:

Use documentary or interview resources to understand and/or reconstruct events of the past

Qualitative:

Use open-ended interviews or written materials to gather explanatory information about perception, opinion, motivation, preference, satisfaction

Quantitative:

Surveys: use forced-choice approach to gather pre-determined information about distributions and frequencies

Clinical Outcomes: use qualitative, survey, &/or experimental approaches to test hypotheses in a real-world setting such as a clinic

Clinical Trials: use qualitative, survey &/or experimental approaches to test hypotheses in a *controlled* setting

Laboratory Trials: use experimental approaches to test hypotheses about the mechanisms that underlie observed phenomena

Mixed Quantitative-Qualitative:

Qualitative methods can be integrated into quantitative research to aid the provision of context and the gathering of explanatory data

Meta-analysis:

compare published outcomes or trials tests of an hypothesis to maximize sample size and reach a larger sense of the "best" answer to a question

Qualitative, quantitative, and trials research are discussed briefly in the body of the paper. Also see information in Appendix 5. Other forms of research include:

Archival/historical research consists of locating and analyzing documents concerning the past of a practice, profession, technique, or interpretation. For example, a researcher might want to compare the origins of two bodywork techniques which appear to represent similar ideas

about the body, yet arose separately. To do this the researcher might trace back written materials on both practices, compare them, study the lives of their originators (if known) and try to find how much one had influenced the other.

Laboratory Trials. The goal of this research is usually to find the mechanisms underlying some behavior or condition or disorder. Sometimes animals are used instead of humans to test interventions. Laboratory research demands special equipment and special training. The results of laboratory research must often be tested by clinical trials or outcomes research to find out if they are safe and effective in humans and in real-world settings.

Meta-analysis. This is a relatively new form of comparative documentary research in which researchers set up criteria of soundness to assess the quality of published reports of clinical trials and clinical outcomes. Typically a topic is selected, the criteria of soundness determined, and then as many articles on the topic as possible are subjected to the analysis. The effect of the process is to combine data and by implication sample sizes from numerous sources. By doing this one can a) assess the quality of research that has been performed on a given intervention and draw out recommendations for future appropriate research designs; b) identify specific limitations of existing research that can be addressed in future research; c) discover if stable results or even a consensus opinion is beginning to emerge concerning use of a particular intervention. Meta-analysis is a research approach that can only be effectively applied once a considerable amount of published data exists on a particular topic or intervention; the ability to do meta-analysis implies a degree of maturity of research.

APPENDIX 2: WHY SCIENCE IS DIFFERENT... AND POWERFUL

Science differs from other major ways of gathering information. There are many ways to assemble and assess information. Some of these include experience ("I know this because I can do it"), revelation ("God showed me"), authority ("the classics say so," "the powerful say so"), and precedence ("it worked this way in the past" "this is the way we do it"). Each of these ways of assembling information is dependent on either personal experience or on cultural and historical norms.

Science is another way to collect and assess information. **Science demands that personal, culturo-historical, perceptual and even technological bias be controlled or minimized so that the credibility and generalizability of the information collected is as high as possible.**

Simple as it is to lay these words out on paper, this demand is, historically speaking, both unique and remarkable. It is because scientific method emphasizes testing claims and measuring the quality of data, recognizes the existence of bias and tries to control or minimize it, and wants to achieve researcher neutrality, that is widely respected as a method that produces trustworthy data. By emphasizing precision, accuracy, and generalizability, and by trying to control the effect of the personal and the singular on theory building, science has made it possible to assess claims by measurement rather than by faith or fiat.

Almost needless to say, the method does not always succeed. Culture and history guide our behavior in ways we often cannot trace, and the practice of science, like all other human endeavors, is not culture-free. Culture and history affect what questions get asked and how, as well as what technology exists to test the questions. Because of these and many other sources

of bias, the scientist knows that no research design can be perfect, and accepts that any "fact" (or theory) may succumb to the arrival of new information.

Nevertheless, scientific method is powerful. The following list emphasizes high points in the rules for practicing science:

1. The concern is with data that can be gathered in the everyday or material world.
2. Scientific knowledge is understood to be relative, approximate, and confoundable.
3. Because of #2, quality control in data collection must be a foremost concern. To minimize bias demands both that information be collected carefully, and that those who collect it "play fair." Scientific method demands that
   a. Data be collected systematically
   b. Data be assessed for quality
   c. Researchers be aware of bias and minimize its effects as far as possible.[10]

Point 1: Science deals with data in the everyday or material world. Some scientists work only with material objects, but as technology has increasingly allowed for the measurement of the smaller and less material, scientific method has expanded to include these. In addition, scientific method can be applied to the study of perception and opinion. In short, science can deal in either direct or indirect measurement. However, scientific method cannot be applied to assessing the quality of, say, a theory about God, or to the assessment of mathematical abstractions.

Point 2: Science recognizes that knowledge is relative and approximate. This powerful statement, one which dismisses the concept of absolute Truth, is an important part of what makes scientific method powerful.

One implication of relativity is that "facts" (and theories) are worthy only so long as they satisfactorily explain the issue under study. If they have been gathered inaccurately, or if their credibility slips in the presence of a new idea, a new technology, or a new test, then the scientific thinker is expected to release any attachment to those facts (or theories), and seek to clarify the situation until credibility once again emerges. Meanwhile, it's appropriate to say "I don't know."

It is however, *never* appropriate to claim that science has "proved" something: science collects evidence and weighs it; it establishes the likelihood of events but does not deal in absolutes such as "truth" or "proof." I emphasize this point because popular languaging about science frequently uses these terms; their use represents a misapprehension of what science is about.

A second important implication of relativity is that the practice of science demands that assumptive patterns be understood and accounted for, since they represent a form of bias. Scientific research proceeds in a series of steps:

assumptions —> derived logic —> derived designs—> data

Though many researchers consciously enter this equation at the third step, it should be clear that the third step is dependent upon the preceding steps. Indeed, the quality of the eventual data is dependent upon the coherence of the entire preceding string. Thus it is as important to know one's assumptions and logic as to know about popular research designs. In fact, in order

[10]Point 3 is developed throughout the paper and appendices.

to know which research designs will serve best, one must know one's assumptions. This point is developed in the body of the paper in the discussion of model fit validity.

Scientific observation differs from clinical observation. Scientific assessment differs from clinical observation (which is a subtype of the method of *experience*) not so much in its greater accuracy—for clinical observation is often extremely accurate—but in wishing to *test* observations for *wider applicability*, in wanting to know *why* or *how* the clinical results are achieved (what is going on behind the scenes), and, finally, in wanting to know all this with as little reference to the original observer as possible.

The process of accurate clinical observation, which is still the basis of all health care systems in the world, is sometimes deemed "prescientific". That is, based on the patients in their own practice and sometimes those in colleagues' practices, clinicians do evolve theories and test them. This approach is not fully scientific because tests are a) typically small scale, and b) usually do not include efforts to avoid bias, including that of the preference of the practitioner. Small scale testing means that variability among or limitations of observations may not appear for some time. Meanwhile, the original observer has integrated his observations into his existing theoretical structure, possibly evolved a new theory to explain the data, and may be loathe to consider new information which seems like a "challenge" to existing theory—especially his personal (cultural, historical) take on the issue. Some health care systems have historically dealt with this common situation by simply absorbing nearly every idea ever hatched (synthetic or holistic approaches) while others want to minimize the number of ideas available and thus argue over "heresy" and "normativity," sorting practices among these two possibilities (reductionistic approach). In short, rather than returning to the data and examining it further (which would be a scientific approach), people often refer back to authority and power, to precedence and to fame, and use these as selectors of their preferred position.

Another important difference is that clinical observation depends upon the skills of the clinician, and her observational skills in turn depend upon the ideas she may be carrying around in her head concerning how the body works and what happens or could happen when an intervention is offered. These ideas are inevitably somewhat personal, cultural, historical—they are not neutral, they are limited, and, in short, they are biased. This does not mean they are intentionally biased; most observers do not wish to misperceive or mislead. But the fact is, we can only reason from within the models that we have been taught, whether as children being socialized, or as adults learning our occupations. This limits us: *one of the rarely discussed tasks of researchers is to broaden their knowledge of models, which translates as broadening their knowledge of potentiality in research design*. Meanwhile, science has recognized the inevitability of bias, and has developed many techniques to minimize its effects.

It is impossible to completely avoid bias. Distinct upbringings and experiences, the effects of culture and the time and place we live in, the technology we employ, even the language we speak, all affect how and what we can think. No one has access to All Knowledge—thus bias is inherent. These points mean that there is no such thing as a "perfect" experiment. Indeed, the task of science is not to remove bias, but to minimize it. To minimize it demands that scientists be aware of its presence, and consciously build bias-protective elements into every step of their research. The overarching researchers' term for this concern is *validity*.

To illustrate the change from the prescientific or clinical way of observing to the scientific one, let's look again at the clinician briefly introduced above. She notices something unusual in her practice, awakens to it, and makes careful observations. Her first case, if formally developed, would be called a *case history*. Case histories are always interesting, but such data is not generally considered scientific because it describes a unique event. However, this clinician actively seeks more examples of the particular condition, so she can test the intervention that was successful in the first patient on several more. She may look back over the years of her clinical notes (retrospective). She may decide to collect data carefully on every new patient who arrives with a similar complaint (prospective). Either way, if she develops her data formally, she has created a *case series*.[11] The case series is a pivot point between clinical observation research and formal scientific research. The case series is an excellent place to begin research *when very little is known about the therapeutic issue*.

Case series often serve as preliminary data series to inform the design of experimental scientific research. If this clinician wishes to go on, she must now develop *hypotheses* that express her ideas about the relationships among the factors she has observed, and then formally *test the hypotheses*, beginning with a *pilot test*.

This example emphasizes clinical research, the testing of therapeutic interventions A clinician may, of course, wish to do other kinds of research. He may wish to survey his patients about the kinds of health care they use, or their satisfaction levels with his care. He may wish to survey colleagues about their experience of their profession, or the frequency with which they treat the issue that interests him and how they treat it. He may want to *compare and contrast* two different ways of delivering massage to treat the same complaint. He may wish to *read or translate older texts*, or *interview* elder practitioners, to understand where current practices came from. He may want to *experiment* with the effects of an intervention on non-humans to seek a physiological mechanism. Whatever his goals, as long as he stays within his own group of patients, or does not try to control the accuracy of his observations by minimizing bias, he is working prescientifically. To move to scientific research requires that the researcher put into place formAL MECHANISMS TO MINIMIZE BIAS AND MAXIMIZE ACCURACY.

APPENDIX 3: CRITERIA OF SOUNDNESS

Material in this Appendix provides more detail on criteria listed in Table 3 in the body of the paper.

Precision helps ensure accuracy, if the level of precision is appropriate to the subject; it can mislead if it is excessive or insufficient. Some decisions are easy—in planning a wall, knowing the weight of one brick to the half pound is sufficiently accurate; knowing it to one hundredth of a pound would serve no useful purpose. Some things are hard to measure—such as the pressure of a hand delivering massage. Is it important to know exactly how much pressure is applied, or will a generally understood measure such as "superficial," "moderate," "deep" serve the purpose more accurately?

Precision also applies to questions on surveys. Each question must subdivide "reality" enough to produce useful data but not so much as to confuse respondents. Suppose one wanted to

[11]Details on doing case series research can be found in Yin 1994.

know if people "improved" while receiving bodywork for frozen shoulder. A question is developed which allows them to say that the pain a) disappeared, b) improved, c) stayed the same, d) got worse. This question is not sufficiently precise, because "improved" is too large compared to the other categories. This question will provide more useful data if "improved" is subdivided as "improved a great deal," "improved somewhat," "improved a little." However, it would probably not improve precision to subdivide "improved" even more because "improvement" is subjective and thus not given to extreme precision.

Reliability is a measure of whether an instrument—be it a device, machine, or questionnaire—gets much the same answer each time it is used in similar circumstances. For example, will an electronic thermometer report the same temperature each time it is inserted into the same child's ear, when this is done 5 times in 30 minutes? Will an electrocardiograph record a person's heartbeat steadily and similarly for all the time that the person is hooked up, assuming that the person lies quietly? Will a survey question receive much the same kind of answer each time it is used, within reason?

Transferability applies especially to questions and questionnaires. It is a measure of whether a question or format will transfer to a new locale and still "make sense" and "be relevant" in the new locale. Suppose some researchers wanted to know what high school students think of football. They develop a survey questionnaire in Bethesda MD. They find it transfers easily to San Antonio TX but only peripherally to Winnipeg Canada and not at all to Bordeaux France. This questionnaire is only partially transferable because the content applies in only some locales; in others the topic is not important to high schoolers or is a complete mystery to them. Whenever one develops a questionnaire, or decides to use an existing—even a "standardized" instrument—it is wise to test it for transferability. Of course, there are instances where transferability is not an issue, for example, when assessing a practice that one already knows is used in only a limited locale.

Credibility/Validity. The essential issue covered by credibility or validity measures is: does the research design or the analytical instruments measure what they intend to measure? For example, if researchers want to know if MBT "increases immune sufficiency," one can judge the quality of the design and data by asking, initially, *have they asked a question that MBT can approach?*, and secondarily, *have they selected ways to measure immune sufficiency that actually measures it?* There are many sub-types of validity: the first five listed in Table 3 are defined below; model fit validity is discussed in the body of the paper.

Face Validity: "on the face of it" does this measure make sense? Recently I asked a series of questions on a survey that let people report what kinds of professional health care they used, and what role they thought acupuncture care played in the improvement in their health (91% reported that their health had improved). The results showed that "most of the time" people used more than one system of health care, yet they reported that acupuncture was the most important factor in improving their health. This was a surprising result—one would have expected acupuncture to be listed as contributory in the whole spectrum—and I had to decide if I would accept the result or assume that something about the content or order of the questions made the statistical results untrustworthy. Having no other published data to compare mine to (as in MBT, there has been little social research on acupuncture), I had to apply the criterion of *face validity*. I searched for reasons to distrust the data. I looked elsewhere in the data for supportive information... and found it in the handwritten responses

to a qualitative question. These remarks were so positive regarding acupuncture that I decided to accept the surprising results at face value.

As the example implies, face validity is a "weak" form of validity which applies particularly at the outset of research on a topic. It's valuable because it often helps identify issues that should be examined in later research projects. Face validity is likely to prove important for MBT research.

Internal Validity makes the researchers ask themselves if the intervention they plan to use is likely to make a measurable difference. For example, suppose an MBT researcher wanted to know if seated massage of the neck and shoulders was equally relaxing if it was delivered for 10 minutes or 15 minutes during coffee breaks at a corporation. Such a time difference is precise, can be reliably measured, is transferable... but is it large enough to make a valid difference in degree of relaxation?

Construct and Content Validity. Do the measurement instruments measure what they are intended to measure? In the previous example, are the measures of "relaxation" chosen to test the time differential actually able to test relaxation? In the example before that, were the questions about types of healthcare used framed in such a way that the respondents knew what was being asked for and recognized the terms offered?

Statistical Conclusion Validity. Were appropriate statistical tests applied to the data, and were they interpreted correctly? Every statistical test is based on certain assumptions about distribution (e.g., the "normal curve"). If data is collected in such a way that it does not conform to the assumptions, then it is not appropriate to apply those tests to that data. This is probably the least arguable of all the forms of validity; it is also technical, which is why it is always wise to bring a statistical consultant in at the beginning of a quantitative research design process.

External Validity is more subtle because it adds a second time dimension to data collection. It asks, basically, if the results of the research apply in the "real" world. Some research, especially research that intends to examine a very minute issue and therefore attempts to control all other impinging issues, easily lacks external validity. In this case, though the data collected "works" (is valid) in its original limited setting, when it enters the real world of accident, flux and uncertainty, it may fail. Many pharmaceutical drugs have suffered from poor external validity; the famous case of the "new coke" which everyone liked in the lab but no one would buy in stores is also an example of poor external validity.

The concern with external validity often applies to efforts to identify, say, people's aptitude for an occupation—will a person who shows up on a questionnaire as having "what it takes" to be a massage therapist actually be likely to go to MBT school, do well in school, establish herself as a practitioner and succeed as a practitioner? Another example: if a person signifies on a questionnaire that they miss human touch and would like more, will this hunger translate to seeking MBT if it is made available to them?

APPENDIX 4: DESCRIPTION, EXPLANATION, PREDICTION, INTERPRETATION

The data that comes out of scientific research can be *descriptive, explanatory, or predictive*. The vast majority is descriptive, including virtually all statistically assessed data. ***Descriptive* questions include who, what, when, where, how many, how much?** Most quantitative research produces descriptive data.

Explanation is much harder to come by, requiring first that a good deal of descriptive data already exist, and second, that special designs be employed. **Explanatory questions include why, and how?** Qualitative and laboratory research can address explanation.

Prediction emerges when the answers to the basic questions are so well known for an issue that the appearance of a particular pattern is known to reliably precede or imply a future pattern.

When a researcher gathers information, whether it be descriptive, explanatory or predictive, she or he must also interpret the data. To interpret means to make sense of information beyond the mere existence of the data. It is to put them into some sort of context. The context is created by one's expectations, the model/paradigm one brings to analysis, thus ***interpretation is always personal, cultural, historical***. Note then, that while science demands that data collection—of data used to describe, explain, or predict—be as free of bias as the designer can make them, interpretation of these data is always embedded in the cultural and historical.

I emphasize this point because it is not uncommon for people to confuse description and interpretation, description and explanation, and explanation and interpretation; such confusions are a potent source of arguments and misapprehension of data.

To ground these remarks in an example: influenza is an infectious disease that has been described in European sources for several hundreds of years. Its symptoms, physiological effects, demographic & climatic patterns, and sequelae are well established. The associated organism has been identified and formally shown to cause influenza. Its genetic patterns are known, as well as its propensity to mutate. The kinds of people most likely to "catch" influenza, and the kinds of environments in which it is most likely to be spread, are well known. In short, influenza has been well described. The how of infection is also well understood, so much so that specialists can make effective vaccines relatively easily. Specialists are also able to predict with some certitude when the danger of influenza illness or epidemics will rise—i.e., when a new mutant organism, crowds and the right season coincide. One thing missing from this assessment is, however, data on "why"—because people seem to have understood and accepted the general linkage between infection, season, and crowds and there has been little cultural disagreement to be accounted for in designing public health interventions for the control of influenza. As a contrast, consider the case of the HIV virus. Here there has been a great need to understand lay explanatory models because people have proven remarkably unwilling to protect themselves even when they knew how the disease was spread and how dangerous it was. It was only when lay perceptions were well understood—primarily as a result of qualitative research—that outreach could effectively motivate behavioral change.

Where does interpretation fit into the influenza example? In a way, it hardly feels present, since the majority of observers accept the remarks made above as "factual." However, there are some who argue from a different position, thus interpret these data differently. For example, there are scientists who say that the presence of an infectious organism is not sufficient to explain the development of symptoms. Others remark on the dangers associated with vaccinations. Laypeople often present arguments from outside the scientific method: some might object to interfering with Nature by creating vaccines; others might claim that epidemics occur when God wishes to punish human misdeeds. In each case the same set of data is used to reach different conclusions: the logic that precedes the conclusions is based in a variety of models about reality.

There is, in fact, no way to "prove" than any interpretation is "right" or "wrong." Science is never about "proof." What makes the scientific method powerful, however, and part of what makes people erroneously use the word "proof" when discussing scientific findings, is that it tries *to raise the likelihood of an interpretation being accurate and accepted as credible by collecting data systematically and with due regard for bias*.

APPENDIX 5: DOING RESEARCH (ADDITIONS TO TABLE 5)

Deleted from Table 5 in the body of the paper are some terms and design components that are important but don't relate as directly to the discussion of model fit validity as the terms discussed there. These are briefly discussed here, and are summarized as Table 8.

Sampling: Since usually one cannot test an intervention or gather data from *every* person to whom the issue might apply, researchers must make a selection out of the whole group, and this selection is called a *sample*. The sample must be a fair image of the larger group—it must contain a similar proportion of people by sex, age, race, occupation... complaint or perceptual category, as the whole population. It must be big enough to allow for analytic credibility, yet it must not be so big that waste is introduced into the research task. Ensuring ***representation*** (*stratified* demographically to reflect the local situation) and ***statistical power*** (adequate size for the research task at hand) is often technical and can be quite difficult—there are whole texts on the problems of sampling—and if one is planning a large scale study, it is wise to consult a sampling specialist to ensure that the ***sampling frame*** is appropriate to the research task.

Table 8: Components and Terminology of Experimental Research Designs

Sample:
sub-set of all people to whom the research question may apply

Cross-sectional Design:
data is collected just once from each participant,yields a "snapshot" image of the issue at one point in time

Pre-test/Post-test Design:
data is collected twice, once before and once after a particular test or intervention is applied.

Longitudinal Design:
data is collected repeatedly from each participant, who are tracked as they move through an event, the trial of an intervention

Follow-up Period:
data is collected past the time of the intervention to measure how long the effects of the intervention last

Pilot Test:
preliminary test of the research design on a small sample

The issue of sampling for quantitative and qualitative research is different. In ***quantitative*** research the usual task is to achieve a "large" sample not only so as to represent all who might belong in the test category (e.g, all those who might someday suffer from low back pain), but also so as to minimize bias, for the effect of unusual cases (***outliers***) on the whole distribution is statistically muted when the sample is larger. For example, if one were collecting data on the costs of MBT, and most people were receiving 10 treatments a year but one person in the sample received 52 treatments, the biasing effect of this one outlier on the average cost would be enormous if only ten people were sampled, but would become progressively less as more and more people were sampled who fit the 10/year model.[12]

In ***qualitative*** research there is much less emphasis on "large" sample size. Instead, researchers want to identify the range of variants of use or perception on a particular topic. Sampling takes a different form: the researcher continues to interview people until s/he is convinced that nearly all opinion on a particular subject has been expressed. For example, suppose one wanted to know what the public thinks "bodywork" is. The first few interviews produce a number of different definitions... as interviews continue it becomes clear that some of these are common explanations, while others are less common or possibly unique to one respondent. Eventually the rate of finding new variants falls so low that the researcher decides that s/he has tapped out this perceptual issue—s/he knows the range of ideas about bodywork present in this population of people. Note that in qualitative research, the sample size is not known at the outset of the research, but emerges as the research proceeds. The value of this procedure is two-fold: first, the researcher knows the range of opinion, which in turn helps identify, for example, how large the task of outreach may be. Suppose research shows that very few people use the term "bodywork" as the profession does, then the communications task is large. Second, if the researcher now wants to turn to doing quantitative research, s/he knows not only what perceptions are common (i.e., must go into questionnaires), but also has clues to what size and shape of sampling frame s/he must create in order to tap a representative large sample.

Cross-sectional and longitudinal designs. A first issue in research design concerns *how often the data will be collected from the same participants*.

If it is collected just once—as in a survey—it is called a ***cross-sectional design***. The cross-sectional approach is appropriate to use when one wants an overview of a topic, such as to gain an idea about who uses MBT and for what purposes, or the general attitude among MBT users toward third party coverage. The advantages of cross-sectional design are that it is rapid and can gather data on large numbers of people at once. It has another advantage for many circumstances: most cross-sectional designs are anonymous, meaning that the researchers do not know who completed the survey. Anonymity means a) the task of ensuring security for the data is reduced, and, b) there is (generally) no need to use Informed Consent forms preceding the survey. The disadvantages of cross-sectional design include that the data collected are descriptive "snapshot" data without much context—while one can *count* how many times people report something, it's hard to know if a similar count would be obtained if a different group or time frame had been selected. Because they lack context cross-sectional data are difficult to use for prediction and they cannot be used for explanation.

There is one important exception to this rule: One can create a ***mixed qualitative-quantitative*** design by including appropriate qualitative questions in a cross-sectional quantitative survey

[12]In practice, extreme outliers are commonly deleted from central tendency analyes in quantitative research.

questionnaire. The qualitative results can then be used to provide context for the quantitative results. This markedly strengthens the survey, which now provides not only descriptive data but explanatory data, thus sometimes allowing for prediction to emerge as well. This approach can be accomplished, for example, by including "white space" for handwritten answers in paper survey instruments, or by interviewing respondents as soon as they have completed a written quantitative questionnaire.

The ***pre-test post-test*** design is a simple 2-step design that is popular to use when a single well-defined intervention is being tested over a relatively short period of time. In this design, baseline data are collected from participants, the intervention is performed, and then data are again collected from participants.

The time frame can be as short as an hour or as long as a few weeks. The advantages of this design are: it is easy to understand and administer, and produces two descriptive snapshots of status at a specified interval. The disadvantage is that it does not allow for detailed tracking of "what happened" and usually does not include a follow-up period to assess whether the intervention has had lasting effects. Pre-test post-test designs are not anonymous, so researchers must take steps to protect participants by securing the data and ensuring that participation was voluntary and informed by using Informed Consent forms. This design might be appropriate to use if one were comparing (for example) relief of chronic neck pain over a four-week period with a) daily use of NSAIDS alone, and b) use of NSAIDS plus twice-a-week neck massage for 30 minutes.

Longitudinal research is a powerful design which also requires a bigger input of design, tracking, and analysis time. The reward is that longitudinal data can provide insight on the process of response to an intervention—rate or components of change—something which is difficult to know from using the previous two design components. In this case, participants record response to treatment, and/or researchers collect data from participants at repeated pre-specified intervals throughout the life of a research project, and usually, for some time after the end of the test intervention (= ***follow-up period***). One advantage of this design is that it allows change in individuals to be charted. Individual patterns of change can be correlated to identify characteristic patterns, or pivot points during intervention, or pivotal intervention strategies. For example, supposing the goal was to track the effects of deep tissue massage on a chronic and variable condition such as fibromyalgia. Using a longitudinal survey design, the researcher would first establish a baseline pre-treatment status for each participant. Subsequently, the researcher would repeat tests and/or add new ones, in some defined conjunction with massage sessions. A quality design would continue to track participants even after the massage component of the project had ended, in order to see how long the changes obtained lasted after the intervention ended. Another advantage of longitudinal design is that it provides a large quantity of precise data and allows development of insight into both disease and intervention process. Potential disadvantages include that researchers must be prepared (computers, statistics) to deal with a large quantity of details. And, since such research cannot be anonymous, researchers must secure the data and protect the participants' privacy with human subject protection formalities such as Informed Consent forms.

Two improvements in design are now considered essential. First, every largescale or quantitative research project should have a preliminary test of all of its design components via a ***pilot test***. Such a test assesses the quality and acceptability of every measurement instrument, the ease and appropriateness of the protocol (step-by-step plan of research), the

feasibility of achieving the goaled sample size, collecting the desired samples, analyzing the samples, tracking the research components, motivating the participants... plus the adequacy of the computer input plan and the data analysis plan. Problems thus exposed can be corrected before running the fullscale project, saving money and frustration, and helping to ensure the credibility of the eventual data.

Second, experimental (as vs survey) designs should almost always include a ***follow-up period***. This is a period of time after the test intervention has ceased to be given, but during which researchers continue to collect data from participants. The value of a follow-up period is that it helps establish how long the effects of an intervention continue to "work" for the participants. Generally, an intervention that works longer is considered to have better effectiveness (including cost-effectiveness).

APPENDIX 6: THE SEQUENCE OF DOING RESEARCH

Table 9 summarizes basic steps in doing research. Most of these apply equally to doing qualitative or quantitative (or mixed qualitative-quantitative) research. Details can be found in research texts.

Table 9: Review of Steps in Research

1. Identify the research question
2. Do background research
3. Clarify the research question in light of the background research
4. Create an appropriate research design
    - Identify goals and design that will serve those goals = Research Rationale
    - Identify team members
    - Designers/interpreters; practitioners; administrators; computer consultants; statisticians; participants
    - Create informed consent form, if needed
    - Identify appropriate venues, appropriate sample size
    - Identify appropriate measurement instruments; create, test, as needed
    - Create time plan, budget, and protocol
    - Apply for grant if needed
    - Identify permits required and get them
5. Pilot Test research design
    - Test every aspect of the design
6. Modify research design in light of results of pilot test
7. Run the test
    - Track every step of the design, participants
    - Make sure all equipment is available and working
    - Ensure incoming data is secure
8. Analyze the results
    - Clean the data
    - Enter data into computer (if large sample size)
    - Analyze data using appropriate qualitative and quantitative software or other equipment
    - Create charts and graphs
9. Interpret the results
10. Publicize the results