Journal of Methods and Measurement in the Social Sciences
Vol. 2, No. 1, 45-62, 2011

Thinking Inside the Box: Simple Methods to Evaluate Complex Treatments

J. Michael Menke
Critical Path Institute, Tucson, AZ

We risk ignoring cheaper and safer medical treatments because they cannot be patented, lack profit potential, require too much patient-contact time, or do not have scientific results. Novel medical treatments may be difficult to evaluate for a variety of reasons such as patient selection bias, the effect of the package of care, or the lack of identifying the active elements of treatment. Whole Systems Research (WSR) is an approach designed to assess the performance of complete packages of clinical management. While the WSR method is compelling, there is no standard procedure for WSR, and its implementation may be intimidating. The truth is that WSR methodological tools are neither new nor complicated. There are two sequential steps, or boxes, that guide WSR methodology: establishing system predictability, followed by an audit of system element effectiveness. We describe the implementation of WSR with a particular attention to threats to validity (Shadish, Cook, & Campbell, 2002; Shadish & Heinsman, 1997).

Keywords: comparative effectiveness, whole systems research, therapy innovation, health care reform

The study of whole health systems is nothing new, nor does it require special tools or advanced mathematics. Indeed, contemporary whole system methods frequently monitor the US health care system and compare national health care systems (Valentine et al., 2004; Hooven et al., 2009). In the 1960's, Avedis Donabedian developed a schema for health system evaluation by dividing health care systems into structure, process, and outcome (Donabedian, 1966). Outcomes from any whole health system are directly attributable to process, which in turn is shaped by system structure. Evaluating whole health systems requires tools for evaluating structure, process, and outcome.

Also emerging in the 1960's was the new field of program evaluation for investigating the effectiveness and behavior of complex whole systems (Rossi, Lipsey, & Freeman, 2004). Subsequently, health technology assessment methods were developed for the expressed purpose of comparing complex interventions introduced through whole systems of delivery (Chalkidou et al., 2009).

Step one in a whole system evaluation is to specify the research perspective. Practitioners, patients, insurance subscribers, payers, and society-at-large all have distinctive stakes and value different outcomes. Patients want to get well; insurance payers want low costs of care; employers want employees back to work; society generally wants lower disability costs and more productive lives; and practitioners want more

income. Disagreements among these stakeholder perspectives are inevitable. For instance, averting back injuries lowers costs to society, but decreases surgeon incomes and hospital revenue, and shifts services to prevention, which is not normally reimbursable.

Step two in Whole Systems Research (WSR) requires choosing a reference system treating the same health condition. Without a referent, clinical successes spontaneous improvements may be misattributed to a system of care, when non-treatment factors are actually responsible. As aspirin and rest are cost-effective treatments in the primary care of musculoskeletal pain, a "novel" whole system must improve upon the "aspirin system's" cost, effectiveness, and risks.

The main threat to WSR has been and always will be selection bias, since the personality differences of those who prefer or accept a novel treatment system may be the same difference that influences their rate of healing (Preference Collaborative Review Group, 2008; Swift & Callahan, 2009; Weinstein et al., 2008; Weinstein et al., 2006). Thus, WSR research must always acknowledge and address this basic challenge to good science.

Finally, WSR must serve as a tool for overcoming the common biases of clinical observation. Practitioners and even patients are not the best judges of their clinical successes (Fischhoff, 2000; Gaeth & Shanteau, 2000; Hammond, 2000; Kleinmuntz, 2000). Further, number and kind dramatic clinical successes under one system may be just as common in all other systems, where natural history, spontaneous remission, or resolution occurs with or without treatment. However, cure of serious disease unexpectedly, as under an novel system of care, is more newsworthy due to contrast biases and framing effects (Connolly, Arkes, & Hammond, 2003).

One Attempt to Innovate in Health Care

Some of the most controversial "medical" innovations of today are complementary and alternative medicines (CAMs), which include Ayurvedic, Traditional Chinese, naturopathic, and chiropractic medicine. CAMs developed independently of scientific or allopathic medicine. CAMs are not newly discovered medical treatments seeking inclusion in health care delivery as some precede modern medicine by many centuries.. CAM practitioners and proponents are existing systems of care seeking legitimacy, acceptance, and parity with allopathic medicine.

However, CAM research results have thus far been modest, leading some purveyors of CAM to be suspicious of the tight experimental controls that are today's standard in allopathic research. Rather than accepting that negative results point to a need to change, to redirect, or refine treatments, many CAM practitioners have claimed their observed clinical successes are attributable to ever more subtle energies and esoteric theories.

Unfortunately, justification more increasingly complex explanations is a signal that a paradigm is about to shift to a more parsimonious explanation (Kuhn, 1962). Among defensive "non-explanations" of *how CAM works* include chaos theory, non-linear systems, free-scale networks, quantum mechanics and quantum entanglement, and the misattribution to the emergent construct of vitalism (Riekeman & Bolles, 2010). Meta-physical explanations thus excuse rather than explain why CAM effectiveness may only be observed in a "whole system" context (Plsek & Greenhalgh, 2001) – and justify that "real" cause and effects exist outside of normal anatomy and physiology (Menke, 2003; Reiser, 1995; Thagard, 2007). Unfortunately for CAM proponents, defending themselves in this manner defines CAM systems as lacking in predictability and thus increasing uncertainty in the health care system, if eventually fully accepted and integrated.

The fear that "conventional research methods" might miss real effects has become so widely shared among CAM providers and researchers as to prompt a call for "whole systems research" (WSR) from NIH (National Center for Complementary and Alternative Medicine, 2007). WSR suggests CAM treatments might only be evaluated fairly within their whole delivery context of history, diagnosis, examination, idiosyncratic explanation of health issues, unique clinical setting, and carefully managing patient expectations.

Aspirations to Legitimacy and Acceptance

Over the 20th Century, allopathic medicine developed systems of diagnoses, specialized into various domains with improved predictability in who might benefit from treatments based on anatomical and physiological sciences. CAMs also have systems of diagnosis and care, with some dating back for centuries. In the professional domain, however, most CAMs have yet to standardize and codify their structure and process formally in a transparent and uniform manner for each CAM profession.

The diversity of CAM perspectives and treatment protocols between and within CAMs suggests to skeptics that system effectiveness may be due to factors other than treatment. Non-specific treatment factors are also present in the practice and delivery of allopathic medical care, but these elements are believed to be understood and the extent of their contributions effectively isolated by randomized controlled trials. Nearly all clinical encounters include emotionally-charged personal conscious or unconscious life and death questions, family and peer pressure (Asch, 1951), experience of powerlessness when confronting pain or disease (Bem, 1967), the authority of physicians and hospitals (Goffman, 1959), automatic social behavior (Cesario, Plaks, & Higgins, 2006), convergence towards agreement with others (Sherif, 1936; Haney, Banks, & Zimbardo,

1973; Milgram, 1963), stereotype threat to self-image (Aronson, Wilson, & Akert, 2005), illness- imposed cognitive dissonance (Festinger & Carlsmith, 1959), too much or too little self-efficacy or self-esteem (Bandura, 1997; Bandura, Freeman, & Lightsey, 1999), cultural perception and preference (Nisbett, 2004), priming and persuasion by practitioner (Cesario & Higgins, 2008), auto kinetic effects (Sherif, 1935), and the complexities of the therapeutic alliance (Iacoviello et al., 2007; Kazdin, Whitley, & Marciano, 2006; Meissner, 2007). In addition, there are psychosocial clinical factors of meaning and belief (Becker, 1973), mortality (Goldenberg et al., 2001), and anxieties piqued by a health crisis (Landau et al., 2004; Maxfield et al., 2007). All systems of care are subject to non-specific effects such as these, but through controlled experiments, the net effect of treatment may be measured.

The possibility that CAM whole systems help just some patients is inconsequential to the question of an effective whole system. The most basic question is whether any CAM whole system can compare favorably with an allopathic standard of care for the same condition. This is the fundamental premise of comparative effectiveness research. Direct comparisons between CAM and conventional medical systems address this question (Assendelft, Koes, van der Heijden, & Bouter, 1992; Koes et al., 1992; Meade & Frank, 1990). So far, CAM evidence has been checked for or assumed to converge towards stable and acceptable effect sizes, possibly because systematic review narratives (Lawrence et al., 2008), consensus processes (Globe, Morris, Whalen, Farabaugh, & Hawk, 2008), and meta-analyses do not compare CAMs directly to other treatments or standards of care.

Structure, Process, and Outcome of Health Systems

Systems are a collection of elements relating and interacting with each other interdependently, operating together as a functional unit for a common or shared purpose. Energy, material, and information flow between system elements and in and out of system boundaries that have various degrees of permeability (Miller, 1965). System boundaries and elements are conceptual and as such require making choices for their inclusion (Cabrera, 2008). Donabedian's three levels of a health care system, structure, process, and outcome may be applied to a CAM or allopathic system (Donabedian, 1966). Allopathic structures include hospitals and clinics that support the practice of medicine. CAMs have different educational and treatment structures.

A Whole Systems Research (WSR) Methodology

As whole systems behaviors are emergent (i.e., unexpected), their outputs can be difficult to predict even with thorough knowledge of the deterministic actions of their elements (Bell & Koithan, 2006; Elder et al., 2006; Koithan et al., 2007; Ritenbaugh, Verhoef, Fleishman, Boon, & Leis, 2003). The "whole systems" research initiative is intended to include all system elements that might interact with or influence whole system outcomes. In other words, the unit of study for WSR is the whole system. Reductionist science's inclination to isolate and evaluate system elements may miss the real potential of a specific herb or treatment (Greenfield, Kravitz, Duan, & Kaplan, 2007; Kravitz, Duan, & Braslow, 2004; Senn, 2004; Senn & Harrell, 1998). The inability to isolate active treatment components is also an acknowledged problem in allopathic medicine (Kadane, 1996; Kline, 2004; Senn & Harrell, 1998).

In Whole Systems Research, it is incumbent upon practitioners to identify all essential elements in the system prior to research. In this schema, CAMs are considered as "black boxes," with patient health measured before and after treatment. From system output, are structure and function thus revealed. For instance, if output variance is high, there might be a responsive subset of patients identifiable by homogeneous class structures or a diversity in provider skill (Richters, 1997).

WSR may provide better external validity than clinical trials, by better characterizing daily clinical practice (Miller, 1974). There are solid methodological concerns to prefer the whole systems perspective as a first step in countering reductionistic science that attempts to distill treatments to "active ingredients." In reductionism, observable emergent properties of living systems dissipate as the unit of life under study decreases in size from nation to community to organism to organ to cell to proteins.

Indeed, first focusing on whole system output avoids getting prematurely bogged down in non-existent mechanisms of actions and fabulous theories used to explain an observable phenomenon that may or may not be seen. Quite simply, unless health is measurably improved either in numbers of people affected, duration, or severity, an innovative treatment holds no practical value as health care delivery innovation. Likewise, a single individual patient success is meaningful if it (1) never occurs in any other system of care, and (2) occurs in patients with identifiable patient structures (i.e., genetic, sex, or race).

Short of the ideal randomized controlled clinical trial, other quasi-experimental methods exist that may compensate sufficiently for lack of true experimental control. The discussion below addresses the methodological issues that require thought when undertaking WSR.

I. Design

A. Ask Only Questions That Can Be Answered

A priori evidence for the deployment of research resources should be preceded by an *ex ante* evaluation (Drummond, Sculpher, Torrance, O'Brien, & Stoddart, 2005; Larson & Kaplan, 1981). Questions of value of the information generated from a study precede and direct experimental design and data collection. The first error to be avoided is in posing a nonsensical research question – called a Type III error (Crabtree & Miller, 1992). "How much chi is required to reduce a fever?" is likely to be a Type III error.

Yet, if a research question makes sense, it still may be unimportant. For instance, lessening humanity's affliction with a common, debilitating, and non-catastrophic condition has no evident value if there are already cost effective protocols, or the condition is of sufficiently low prevalence or low illness burden. A last step is to assess whether the question was worth asking and is worth asking again in a more refined form to further reduce uncertainty. Finally, estimates *of how much* information asymmetry was reduced by a study can be estimated, and a value attached to the need for additional research. These are the value of information analyses.

B. Choose Outcome Measures

In spite of a clarion call for WSR for CAM research, proponents and skeptics alike are still tempted to posit mechanisms of action before first establishing system effectiveness. In reversing the order of inquiry to "how does it work?" before answering "does it work?" only wastes valuable resources.

Care must be taken to not miss contributory placebos or other non-treatment elements as sources of clinical success. The CAM focus on vitalism and self-generated health, may lead CAM practitioners to ignore the psychosocial and environmental aspects of that promote and maintain certain maladies. Back pain disability, for instance, has a very large psychosocial component (Gatchel, Polatin, & Mayer, 1995; Papageorgiou et al., 1997). Back pain disability perpetuated by psychosocial causes is not likely to improve while under any system of care. An inability to get a stable estimate of lesion location and degree suggests an unknown causal factor or the presence of an emergent construct (Little, Lindenberger, & Nesselroade, 1999). Unknown and unacknowledged factors might include therapeutic alliances (mediating factors) or stresses at home or work (moderating factors) (Elvins & Green, 2008; Knaevelsrud & Maercker, 2007). Emergent constructs do not suggest a deeper cause, but the marvelous brain of the observer to classify and interpret phenomena according to his or her training and habit.

C. Acknowledge Selection Bias and Control It

The primary threat to CAM research is selection bias. That is, people who choose CAM could be fundamentally different from those who seek allopathic care in belief and lifestyle choices. Large-scale randomized controlled clinical trials for comparing medicine and CAM would be the ideal for comparing the effectiveness of various systems of care. However, randomization may be subverted by post-randomization events, such as attrition. That is, patients favoring CAM are probably different from those not seeking CAM system care (Preference Collaborative Review Group, 2008; Swift & Callahan, 2009).

II. Interpreting Results

A. The Distribution of Effect Sizes in a Population

Could CAM have identifiable subgroups or individuals who respond to care? One charge by CAM proponents is that results of patients who benefit from CAM care may get lost in common medical research methodology (Senn, 2004). Others (Wennberg, 1996) point to basic questions of treatment results: Does a 20% improvement mean 20% of the patient population improved, symptoms improved by an average 20% in everyone, or a subgroup was improved by 100%, but represented 20% of the population? Most likely, clinical improvement refers to the improvement distributed throughout a population of interest. Population improvement is difficult to interpret for highly idiosyncratic and rare patient problems.

An allopathic medicine core assumption is "universality of pathophysiologic and pharmacologic mechanisms" (Greenfield, et al., 2007). Studies are carefully designed to isolate treatment from non-treatment influences in outcome. Changes in patient health are observed after applying different types or degrees of treatment, and cause inferred. If a treatment effect is assumed to be a stable entity, variance within treatments is considered to treatment interactions and measurement error (Greenfield, et al., 2007; Jonas, Beckner, & Coulter, 2006). Health changes after treatment may be random (i.e., not attributable to treatment) or attributable to treatment in two ways. The treated individuals may experience a general benefit widely distributed in the group, or an identifiable sub-group may benefit as a heterogeneous treatment effect (HTE) (Kravitz, et al., 2004). Heterogeneous treatment effects may occur from sex, age, genes, psychosocial context, or habits.

B. Take Care to not Misattribute Cause

Finally, most people are poor at estimating probabilities. "Gut instincts" and intuition can easily mislead even the most intelligent and educated (Ellsberg, 1961; Oliver, 2003; Rieger & Wang, 2006; Samuelson, 1960). As a result, statistical methods regularly outperform clinical intuition (Kleinmuntz, 2000; Meehl, 1954). It may seem paradoxical at first thought: the *probability* of a spontaneous cancer remission in a specific patient is nearly zero, but the *likelihood* that someone in the world will experience spontaneous remission of his or her cancer sometime is nearly certain. The conflation of probability (events that happened) with likelihood (events that might happen) elicits awe and adds an emotional component to causal misattribution (Jaynes, 1986). "I saw it with my own eyes" can keep, and has kept, real medical and scientific progress stalled for decades.

Well-established biases in human judgment explain miracles. Did the patient really have the disease? Framing effects, salience biases (Kahneman, 2003; Tversky & Kahneman, 1974), emergent constructs (Cohen, Cohen, Teresi, Marchi, & Velez, 1990), misinterpretation of probability (Menke & Skrepnek, 2009), and causal misattribution (Earman, 1992; Jaynes, 1986) are especially active when fears of illness and mortality are provoked (Waddell, 1996; Greenberg, Pyszczynski, Solomon, Simon, & Breus, 1994).

Mistaking causal explanations with obtuse references to non-linear systems, complexity theory, complex systems non-local actions, free-scale networks and quantum mechanics unnecessarily detour or stall a legitimate inquiry (Menke & Skrepnek, 2009; Barabási, A.-L. & Bonabeau, 2003).

C. Consider Threats to WSR Internal Validity

Internal validity refers to the ability to assert that treatments and not some other factor are responsible for different outcomes. The *history* threat to internal validity refers to non-treatment events that may occur between the first and second system observation that affect measurement. Some patients enter treatments as a signal of their commitment to behavioral and lifestyle change. Choosing Traditional Chinese Medicine for back pain may be accompanied by a resolve – perhaps also encouraged by their practitioner to lose weight, reduce stress, drink less, and exercise. After six months of management, which aspects of clinical change were from TCM specific treatment factors and which from ancillary behavioral changes?

Maturation effects are seen in the long-term treatment of conditions that resolve with the passage of time. Multi-year treatment of mild scoliosis, enuresis, back pain (60 to 70% complete resolution in 6 weeks) (Shen, Samartzis, & Andersson, 2006), and even psychosis, may yield successes, but some proportion of these conditions improve with time and attention, or are episodic. Attribution to treatment may be unwarranted.

The act of repeated *testing* (or data collection) may introduce bias through its setting, administration, how the test is built (*instrumentation bias*), and patient learning or practice effect. Thus, the test itself can easily pick up influences unrelated to treatment, such as practitioner and patient expectation. CAM testing and instrumentation bias might be especially active if the test does not calibrate with patient behaviors at work and home.

Statistical regression is when patients with abnormal – usually acute - health conditions normalize towards the group's mean over time. Patient symptoms when tracked hour by hour will fluctuate. Patients with an acute injury tend to get better over time, even without treatment. When acute severe pain (or other malady) patients enter treatment, their natural improvement may be mistakenly attributed to the CAM treatment system.

As discussed previously, *selection bias* is a major issue in CAM research because CAM patients may differ fundamentally from allopathic ones. *Experimental mortality* also is one issue discussed above in the "Compelling nature" section. Unresponsive patients dropping out can make the treatment group appear to have better outcomes.

D. Identify Threats to External Validity

External validity assures the practitioner that what was found in one study is generalizable to other circumstances and patients. Selection biases may interact with the type of CAM system. Interaction between system and selection is likely with CAM and allopathic comparisons, where up to 65% of CAM effectiveness arises from the practitioner-patient bond, and 16% of effectiveness is accounted for by preference (Hyland, 2005; Preference Collaborative Review Group, 2008; Swift & Callahan, 2009). With a s*election bias by group interaction*, we might expect patients getting the treatment they want to improve more or more rapidly.

Patients under different systems of care may also experience different treatment contexts. Homeopathic patients meeting at a lay homeopath's home is a different set of demand characteristics from medical patients reporting to a hospital. Generalizability of homeopathic treatment to homes and hospitals is thus difficult.

One of the reasons for CAM use may be a failure of medicine to address some patient problems. Some chiropractic patients have had multiple surgeries, some of which failed; such patients are then often addicted to

narcotics for their pain. In the other direction, they may have had many previous medical or CAM successes by the time they enter a CAM study. As Campbell and Stanley noted, "the effects of prior treatments are not usually erasable" (Campbell & Stanley, 1963). Variability introduced by many concurrent interventions and history can obscure current treatment effects, weakening the attribution of results to the treatment.

III. Design Refinements

The Comparative Effectiveness Decision Framework

Comparative effectiveness summarizes the current knowledge of a CAM system with respect to the current allopathic standard of care. Preliminary findings can emerge from a one-shot case study (consistent improvement of people after an intervention), a one group pretest-posttest design (group improvement after treatment), or static group comparison (two groups, one treated and one not). Preliminary evidence may also include meta-analyses or expert opinion. These non-robust bits of evidence, and any published studies, and meta-analyses can be blended into a *"virtual experiment" to compare* CAM and medical treatments to estimate the relative value in effectiveness and cost for a specific health condition (Neal, 1992).

A fundamental principle in WSR is to compare one whole system to another. Practitioners of CAM systems observe clinical improvements, perhaps even dramatic ones, especially among large numbers of patients. However, medical systems also produce clinical improvements and more than a few "miracles." Observed in isolation, the output from a CAM system might mislead observers to ask and pursue *how* a CAM system works even before it has been demonstrated that the CAM system performs better than base rates of natural or spontaneous remissions present in all universes. For the CAM quest, the standard of comparison is the allopathic system treating the same condition. Direct and indirect costs and risks are to be included in the analysis. A WSR analysis should yield a *relative* CAM effect size compared to the allopathic standard, to inform decisions as to whether or not CAM should be recommended to patients and payers.

Adaptive Trials

Possibly, patients who did not succeed or only slightly succeeded with CAM treatment may obscure some notable successes in some individuals in the treatment group. Further, health care is generally practiced on individuals according to patient progress. Thus, CAM WSR can include individual responses to whole system care by including adaptive

treatments in treatment and measuring clinical progress (Bierman, Nix, Maples, & Murphy, 2006; Murphy & McKay, 2003; Murphy, 2005; Murphy, Collins, & Rush, 2007; Murphy, Lynch, Oslin, McKay, & TenHave, 2007; Murphy, Oslin, Rush, & Zhu, 2007).

Instead of assuming that all patients are effectively identical in responses and tolerances to treatment, adaptive treatment designs monitor and adjust dosages and treatments frequency during the course of care to meet the personal clinical objectives for each patient.

Preference Trials

The preferences that patients have about CAMs and allopathic medicine may be a factor in system outcome. A patient assigned to a non-preferred treatment might differ in adherence, confidence in, and response to care. A meta-analysis of the patient preference effect on 26 studies of 2300 patients gave a small effect size of r = 0.15, 95% CI of 0.09 to 0.21 in favor of patient preferences (Swift & Callahan, 2009). A Cochrane Review of eight musculoskeletal trials (n = 1594) found a similar effect size of 0.16 (95% CI of 0.11 to 0.31) with those getting their preferred treatment doing better (effect size = 0.15, 95% CI of -0.04 to 0.34) (Preference Collaborative Review Group, 2008). The Cochrane paper also found lower dropouts in preferred treatment groups.

A preference trial works by eliciting patient preferences before assignment to a treatment condition. Patients with a preference for allopathic or CAM whole systems are given their choice of treatment. Patients who have no preference are randomly assigned to a treatment condition. After assignment, each treatment condition has some patients who preferred the treatment received and others who were indifferent. Thus, preference is then a control factor in the analysis and may be analyzed for its effect on outcome. In the large multi-centered SPORT trial to compare surgical versus non-operative approaches to lumbar disk herniation (Birkmeyer et al., 2002; Weinstein, Lurie, Olson, Bronner, & Fisher, 2006; Weinstein, Tosteson, et al., 2006), the 501 patients who expressed no preference were randomized to either surgery or no surgery groups, while those choosing either surgery (n = 521) or non-surgery (n = 222) were assigned to their preference. Surgical patients did better in all groups, but there were a number of patients who sought more treatments than their assigned one, and so the effect may have been less certain.

Quasi-experimental Designs

Quasi-experiments are experiments that lack random assignment to the treatment groups. Quasi-experiments have “similar purposes and structural attributes to randomized experiments” (Shadish, et al., 2002).

Causal inference from quasi-experiments must meet three basic requirements: cause precedes effect, cause covaries with the effect, and alternative explanations for the causal relationship are implausible. Confirming that cause precedes effect can be addressed by controlling threats to internal validity. Demonstrating that cause and effect covary is done with well thought-out experimental design and adjusted statistically to some degree in the analysis phase. Demonstrating implausibility of alterative explanations is accomplished through *coherent pattern matching* where complex predictions are made on the basis of CAM theory. The more complex the predicted outcome and the better the match, the more likely the treatment is responsible. These three principles can be used in a variety of quasi-experimental designs.

Innovative treatments may hold elaborate theories and mechanisms of action and some degree of clinical evidence, but systems must first be evaluated for whole system effectiveness, cost, and risks against a current standard of care for treatment of a specific condition in question. Finally, given reliable whole system performance, system elements must be inspected to confirm their actual contribution to system performance.

Acknowledgements

This paper was sponsored in part by: National Center for Complementary and Alternative Medicines (NCCAM) of the National Institutes of Health (NIH), Grant Number: T32 Fellowship AT01287, the National University of Health Sciences, and the National Chiropractic Mutual Insurance Corporation. Corresponding author: J. Michael Menke; Critical Path Institute, 1730 East River Road, Tucson, Arizona 85718-5893, email: michael.menke@gmail.com.

References

Aronson, E., Wilson, T. D., & Akert, A. M. (2005). *Social Psychology, 5th edition*. Upper Saddle River, NJ: Prentice Hall.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men* (pp. 177-190). Pittsburgh, PA: Carnegie Press.

Assendelft, W. J., Koes, B. W., van der Heijden, G. J., & Bouter, L. M. (1992). The efficacy of chiropractic manipulation for back pain: Blinded review of relevant randomized clinical trials. *Journal of Manipulative & Physiological Therapeutics, 15*, 487-494.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Bandura, A., Freeman, W. H., & Lightsey, R. (1999). Self-efficacy: The exercise of control. *Journal of Cognitive Psychotherapy, 13*, 158-166.

Barabási, A.-L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American, 288*(5), 60-69.

Becker, E. (1973). *The Denial of Death*. New York: The Free Press.
Bell, I. R., & Koithan, M. (2006). Models for the study of whole systems. *Integrative Cancer Therapies, 5*, 293-307.
Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review, 74*, 183-200.
Bierman, K. L., Nix, R. L., Maples, J. J., & Murphy, S. A. (2006). Examining clinical judgment in an adaptive intervention design: The fast track program. *Journal of Consulting and Clinical Psycholology, 74*, 468-481.
Birkmeyer, N. J., Weinstein, J. N., Tosteson, A. N., Tosteson, T. D., Skinner, J. S., Lurie, J. D., et al. (2002). Design of the Spine Patient outcomes Research Trial (SPORT). *Spine, 27*, 1361-1372.
Cabrera, D. (2008). *Distinctions, Systems, Relationships, Perspectives: The Simple Rules of Complex Conceptual Systems*. Unpublished manuscript.
Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally College Publishing Company.
Cesario, J., & Higgins, E. T. (2008). Making message recipients "feel right": How nonverbal cues can increase persuasion *Psychological Science, 19*, 415-420.
Cesario, J., Plaks, J. E., & Higgins, E. T. (2006). Automatic social behavior as motivated preparation to interact. *Journal of Personality and Social Psychology, 90*, 893-910.
Chalkidou, K., Tunis, S., Lopert, R., Rochaix, L., Sawicki, P. T., Nasser, M., et al. (2009). Comparative effectiveness research and evidence-based health policy: Experience from four countries. *The Milbank Quarterly, 87*, 339-367.
Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement, 14*, 183-196.
Connolly, T., Arkes, H. R., & Hammond, K. R. (2003). *Judgment and Decision Making: An Interdisciplinary Reader, 2nd Edition*. New York: Cambridge University Press.
Crabtree, B., & Miller, W. (1992). *Doing qualitative research*. Newbury Park, CA: Sage Publications.
Donabedian, A. (1966). Evaluating the quality of medical care. *The Milbank Quarterly, 44*(Suppl.), 166-206.
Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J., & Stoddart, G. L. (2005). *Methods for the economic evaluation of healh care programmes*. Oxford: Oxford University Press.
Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.
Elder, C., Aickin, M., Bell, I. R., Fonnebo, V., Lewith, G. T., Ritenbaugh, C., et al. (2006). Methodological challenges in whole systems research. *Journal of Alternative and Complementary Medicine, 12*, 843-850.
Ellsberg, D. (1961). Risk, ambguity, and the Savage axioms. *Quarterly Journal of Economics, 75*, 643-669.
Elvins, R., & Green, J. (2008). The conceptualization and measurement of therapeutic alliance: An empirical review. *Clinical Psychology Review,* 28, 1167-1187.
Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58*, 203-210.

Fischhoff, B. (2000). Value elicitation: Is there anything in there? In T. Connolly, H. R. Arkes & K. R. Hammond (Eds.), *Judgment and Decision Making: an Interdisciplinary Reader, 2nd Edition* (pp. 517-543). New York: Cambridge University Press.

Gaeth, G. J., & Shanteau, J. (2000). Reducing the influence of irrelevant information on experienced decision makers. In T. Connolly, H. R. Arkes & K. R. Hammond (Eds.), *Judgment and Decision Making: an Interdisciplinary Reader, 2nd Edition* (pp. 305-323). New York: Cambridge University Press.

Gatchel, R. J., Polatin, P. B., & Mayer, T. G. (1995). The dominant role of psychosocial risk factors in the development of chronic low back pain disability. *Spine, 20*, 2702-2709.

Globe, G. A., Morris, C. E., Whalen, W. M., Farabaugh, R. J., & Hawk, C. (2008). Chiropractic management of low back disorders: Report from a consensus process. *Journal of Manipulative & Physiological Therapeutics, 31*, 651-658.

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. New York: Doubleday Anchor.

Goldenberg, J. L., Pyszczynski, T., Greenberg, J., Solomon, S., Kluck, B., & Cornwell, R. (2001). I am not an animal: Mortality salience, disgust, and the denial of human creatureliness. *Journal of Experimental Psychology: General, 130*, 427-435.

Greenberg, J., Pyszczynski, T., Solomon, S., Simon, L., & Breus, M. (1994). Role of consciousness and accessibility of death-related thoughts in mortality salience effects. *Journal of Personality and Social Psychololology, 67*, 627-637.

Greenfield, S., Kravitz, R., Duan, N., & Kaplan, S. H. (2007). Heterogeneity of treatment effects: Implications for guidelines, payment, and quality assessment. *American Journal of Medicine, 120*(Suppl. 1), 3-9.

Preference Collaborative Review Group. (2008). Patients' preferences within randomised trials: Systematic review and patient level meta-analysis. *British Medical Journal, 337*, a1864.

Hammond, K. R. (2000). Coherence and correspondence theories in judgment and decision making. In T. Connolly & H. R. Arkes (Eds.), *Judgment and Decision Making: An Interdisciplinary Reader, 2nd Edition*, (pp. 53-65). Cambridge: Cambridge University Press.

Haney, C., Banks, C., & Zimbardo, P. G. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminolology and Penology, 1*, 69-97.

Hooven, F. H., Adachi, J. D., Adami, S., Boonen, S., Compston, J., Cooper, C., et al. (2009). The Global Longitudinal Study of Osteoporosis in Women (GLOW): Rationale and study design. *Osteoporosis International, 20*, 1107-1116.

Hyland, M. E. (2005). A tale of two therapies: Psychotherapy and complementary and alternative medicine (CAM) and the human effect. *Clinical Medicine, 5*, 361-367.

Iacoviello, B. M., McCarthy, K. S., Barrett, M. S., Rynn, M., Gallop, R., & Barber, J. P. (2007). Treatment Preferences Affect the Therapeutic Alliance: Implications for Randomized Controlled Trials. *Journal of Consulting and Clinical Psychology, 75*, 194-198.

Jaynes, E. T. (1986). Bayesian methods: General background. In J. H. Justice (Ed.), *Maximum Entropy and Bayesian Methods in Geophysical Inverse Problems*. Cambridge: Cambridge University Press.

Jonas, W. B., Beckner, W., & Coulter, I. (2006). Proposal for an integrated evaluation model for the study of whole systems health care in cancer. *Integrative Cancer Therapies, 5*, 315-319.

Kadane, J., Ed. (1996). *Bayesian methods and ethics in a clinical trial design*. New York: John Wiley and Sons.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*, 697-720.

Kazdin, A. E., Whitley, M., & Marciano, P. L. (2006). Child-therapist and parent-therapist alliance and therapeutic change in the treatment of children referred for oppositional, aggressive, and antisocial behavior. *Journal of Child Psychology and Psychiatry, 47*, 436-445.

Kleinmuntz, B. (2000). Why we still use our heads instead of formulas: Toward an integrative approach. In T. Connolly, H. R. Arkes & K. R. Hammond (Eds.), *Judgment and Decision Making: An Interdisciplinary Reader, 2nd Edition* (pp. 681-711). New York: Cambridge University Press.

Kline, R. (2004). *Beyond significance testing*. Washington, DC: American Psychological Association.

Knaevelsrud, C., & Maercker, A. (2007). Internet-based treatment for PTSD reduces distress and facilitates the development of a strong therapeutic alliance: A randomized controlled clinical trial. *Biomed Central Psychiatry, 7*, 13. http://www.biomedcentral.com/1471-244X/7/13

Koes, B. W., Bouter, L. M., van Mameren, H., Essers, A. H., Verstegen, G. M., Hofhuizen, D. M., et al. (1992). A blinded randomized clinical trial of manual therapy and physiotherapy for chronic back and neck complaints: Physical outcome measures. *Journal of Manipulative & Physiological Therapeutics, 15*, 16-23.

Koithan, M., Verhoef, M., Bell, I. R., White, M., Mulkins, A., & Ritenbaugh, C. (2007). The process of whole person healing: "Unstuckness" and beyond. *Journal of Alternative and Complementary Medicine, 13*, 659-668.

Kravitz, R. L., Duan, N., & Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly, 82*, 661-687.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Landau, M. J., Johns, M., Greenberg, J., Pyszczynski, T., Martens, A., Goldenberg, J. L., et al. (2004). A function of form: Terror management and structuring the social world. *Journal of Personality and Social Psychology, 87*, 190-210.

Larson, R. C., & Kaplan, E. H. (1981). Decision-oriented approaches to program evaluation. *New Directions for Program Evaluation: Evaluation of Complex Systems*, 49-68.

Lawrence, D. J., Meeker, W., Branson, R., Bronfort, G., Cates, J. R., Haas, M., et al. (2008). Chiropractic management of low back pain and low back-related leg complaints: A literature synthesis. *Journal of Manipulative & Physiological Therapeutics, 31*, 659-674.

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4*, 192-211.

Maxfield, M., Pyszczynski, T., Kluck, B., Cox, C. R., Greenberg, J., Solomon, S., et al. (2007). Age-related differences in responses to thoughts of one's own death: Mortality salience and judgments of moral transgressions. *Psychology and Aging, 22*, 341-353.

Meade, T. W., & Frank, A. O. (1990). Low back pain: Comparison of chiropractic and hospital outpatient treatment. *British Medical Journal, 300*, 1723.

Meehl, P. E. (1954). *Clinical versus Statistical Prediction*. Minneapolis: University of Minnesota Press.

Meissner, W. W. (2007). Therapeutic alliance: Theme and variations. *Psychoanalytic Psychology, 24*, 231-254.

Menke, J. M. (2003). Principles in integrative chiropractic. *Journal of Manipulative & Physiological Therapeutics, 26*, 254-272.

Menke, J. M., & Skrepnek, G. H. (2009). Complexity. In M. Kattan (Ed.), *Encyclopedia of Medical Decision Making* (pp. 144-149). Newbury Park, California: Sage Publications.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371-378.

Miller, D. (1974). Popper's qualitative theory of verisimilitude. *The British Journal for the Philosophy of Science, 25*, 166-177.

Miller, J. G. (1965). Living systems: Basic concepts. *Behavioral Science, 10*, 193-235.

Murphy, S. A., & McKay, J. (2003). *Adaptive Treatment Strategies: An Emerging Approach for Improving Treatment Effectiveness*. Unpublished manuscript.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics and Medicine, 24*, 1455-1481.

Murphy, S. A., Collins, L. M., & Rush, A. J. (2007). Customizing treatment to the patient: Adaptive treatment strategies. *Drug and Alcohol Dependence, 88(*Suppl. 2), 1-3.

Murphy, S. A., Lynch, K. G., Oslin, D., McKay, J. R., & TenHave, T. (2007). Developing adaptive treatment strategies in substance abuse research. *Drug and Alcohol Dependence, 88(*Suppl. 2), 24-30.

Murphy, S. A., Oslin, D. W., Rush, A. J., & Zhu, J. (2007). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology, 32*, 257-262.

National Center for Complementary and Alternative Medicine. (2007). Whole Medical Systems: An Overview. *BackGrounder, National Center for Complementary and Alternative Medicine, D236*. Retrieved from http://nccam.nih.gov/health/backgrounds/wholemed.htm

Neal, R. M. (1992). Bayesian mixture modeling by Monte Carlo simulation. In C. R. Smith, G. J. Erickson & P. O. Neudorfer (Eds.), *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis, Seattle, 1991* (pp. 197-211). Dordrecht: Kluwer Academic Publishers.

Nisbett, R. E. (2004). *The Geography of Thought*. New York: Free Press.
Oliver, A. (2003). A quantitative and qualitative test of the Allais paradox using health outcomes. *Journal of Economic Psychology, 24*, 35-48.
Papageorgiou, A. C., Macfarlane, G. J., Thomas, E., Croft, P. R., Jayson, M. I., & Silman, A. J. (1997). Psychosocial factors in the workplace--do they predict new episodes of low back pain? Evidence from the South Manchester Back Pain Study. *Spine, 22*, 1137-1142.
Plsek, P. E., & Greenhalgh, T. (2001). Complexity science: The challenge of complexity in health care. *British Medical Journal, 323*, 625-628.
Reiser, S. (1995). Anatomic thinking: The clinical and social consequences of health care's basic logic. *Family & Community Health, 18*, 26-36.
Richters, J. E. (1997). The Hubble hypothesis and the developmentalist's dilemma. *Development and Psychopathology, 9*, 193-229.
Rieger, M. O., & Wang, M. (2006). Cumulative prospect theory and the St. Petersburg paradox. *Economic Theory, 28*, 665-679.
Riekeman, G. F. & Bolles, S.. (2010, April). *Working Summit Proceedings*. Paper presented at the The Life Source Octagon: A Center for Infinite Thinking, Atlanta, GA.
Ritenbaugh, C., Verhoef, M., Fleishman, S., Boon, H., & Leis, A. (2003). Whole systems research: A discipline for studying complementary and alternative medicine. *Alternative Therapies in Health and Medicine, 9*, 32-36.
Rossi, P. H., Lipsey, W. M., & Freeman, H. E. (2004). *Evaluation: A systematic approach, 7th Edition*. Thousand Oaks, CA: Sage Publications.
Samuelson, P. (1960). The St. Petersburg Paradox as a divergent double limit. *International Economic Review, 1*, 31-37.
Senn, S. (2004). Individual response to treatment: Is it a valid assumption? *British Medical Journal, 329*, 966-968.
Senn, S., & Harrell, F. E., Jr. (1998). On subgroups and groping for significance. *Journal of Clinical Epidemiology, 51*, 1367-1368.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.
Shadish, W. R., & Heinsman, D. T. (1997). Experiments versus quasi-experiments: Do they yield the same answer? *NIDA Research Monograph, 170*, 147-164.
Shen, F. H., Samartzis, D., & Andersson, G. B. (2006). Nonsurgical management of acute and chronic low back pain. *Journal of the American Acadamy of Orthopedic Surgery, 14*, 477-487.
Sherif, M. (1935). A study of some social factors in perception: Chapter 2. *Archives of Psychology, 27*(187), 17-22.
Sherif, M. (1936). *The psychology of social norms*. New York: Harper.
Swift, J. K., & Callahan, J. L. (2009). The impact of client treatment preferences on outcome: A meta-analysis. *Journal of Clinical Psychology, 65*, 368-381.
Thagard, P. (2007). The Concept of Disease: Structure and Change Available from http://cogsci.uwaterloo.ca/Articles/Pages/Concept.html - anchor02
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Valentine, J. M., Neff, J., Park, A. N., Sharp, V., Maynard, C., Christakis, D., et al. (2004). Pediatric hospitalization patterns for selected chronic health conditions using hospital abstract reporting system data: Methods and findings. *Health Services and Outcomes Research Methodology 1*, 335-350.

Waddell, G. (1996). Low back pain: A twentieth century health care enigma. *Spine, 21*, 2820-2825.

Weinstein, J. N., Lurie, J. D., Olson, P. R., Bronner, K. K., & Fisher, E. S. (2006). United States' trends and regional variations in lumbar spine surgery: 1992-2003. *Spine, 31*, 2707-2714.

Weinstein, J. N., Lurie, J. D., Tosteson, T. D., Tosteson, A. N., Blood, E. A., Abdu, W. A., et al. (2008). Surgical versus nonoperative treatment for lumbar disc herniation: Four-year results for the Spine Patient Outcomes Research Trial (SPORT). *Spine, 33*, 2789-2800.

Weinstein, J. N., Tosteson, T. D., Lurie, J. D., Tosteson, A. N., Hanscom, B., Skinner, J. S., et al. (2006). Surgical vs nonoperative treatment for lumbar disk herniation: The Spine Patient Outcomes Research Trial (SPORT): A randomized trial. *Journal of the American Medical Association, 296*, 2441-2450.

Wennberg, J. E. (1996). On the appropriateness of small-area analysis for cost containment. *Health Affairs, 15*, 164-167.