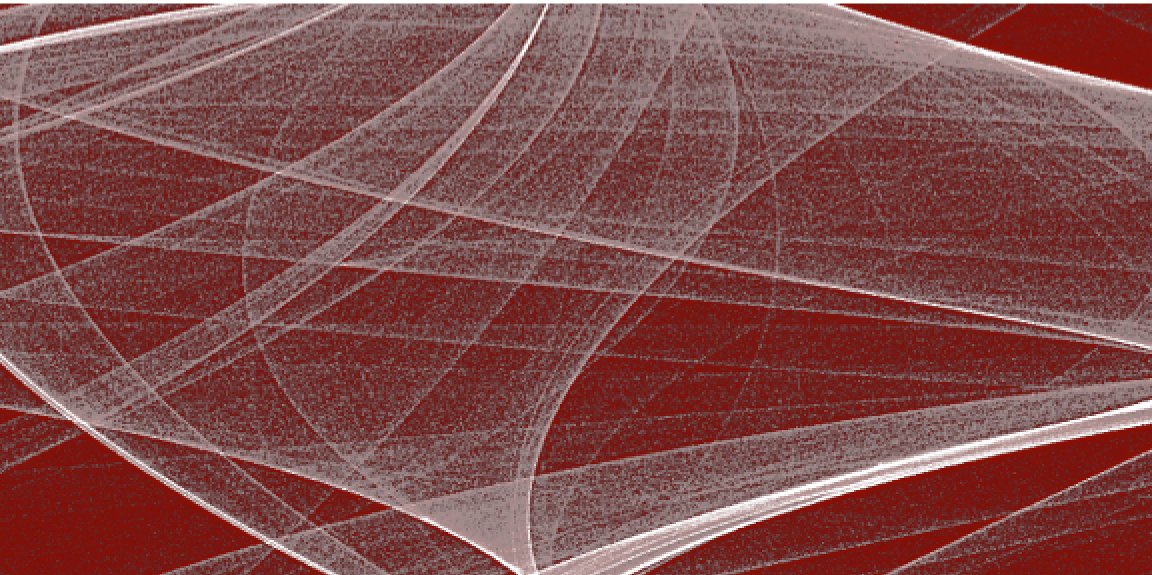
ROUTLEDGE

Understanding Research in Clinical and Counseling Psychology

SECOND EDITION



EDITED BY

Jay C. Thomas and Michel Hersen

Understanding Research in Clinical and Counseling Psychology

SECOND EDITION

Understanding Research in Clinical and Counseling Psychology

SECOND EDITION

EDITED BY

Jay C. Thomas and Michel Hersen

Routledge
Taylor & Francis Group
New York London

Routledge
Taylor & Francis Group
270 Madison Avenue
New York, NY 10016

Routledge
Taylor & Francis Group
27 Church Road
Hove, East Sussex BN3 2FA

© 2011 by Taylor and Francis Group, LLC
Routledge is an imprint of Taylor & Francis Group, an Informa business

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-0-415-99221-3 (Paperback)

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Understanding research in clinical and counseling psychology / edited by Jay C.
Thomas, Michel Hersen. -- 2nd ed.
p. cm.
First published: Mahwah, N.J. : Lawrence Erlbaum Associates, 2003.
Includes bibliographical references and index.
ISBN 978-0-415-99221-3 (pbk. : alk. paper)
1. Clinical psychology--Research. 2. Counseling psychology--Research. 3.
Psychotherapy--Research. I. Thomas, Jay C., 1951- II. Hersen, Michel.

RC467.U53 2011
616.890072--dc22 2010035024

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the Routledge Web site at
http://www.routledgementalhealth.com

Contents

Part I Research Foundations

Part II Research Strategies

Part III Research Practice

Part IV Special Problems

Preface

Understanding Research in Clinical and Counseling Psychology, Second Edition, is a result of our experiences in teaching and working with students in professional psychology and counseling over many years. Although virtually all graduate programs require a course on research, the basis for that requirement is often shrouded in mystery to many students. They enter their graduate training with the admirable ambition of learning skills important for assisting clients to make changes. Although practice may be somehow loosely based on research, for those students, the connection is not clear, and the value of psychological research is not readily apparent. In preparing this book, we set out to create a text that presents research as an indispensable tool for practice.

There are illustrations of how research can improve practice throughout the text. Such improvement can be seen in better assessment, treatment choice, and, most importantly, client outcomes. Research in clinical areas is fascinating, invigorating, and fulfilling to do, but it gains value when it brings about better practice. Practitioners need to know how to utilize research findings and to communicate with researchers what new knowledge is most needed. Our intent is this text builds these basic skills.

This is an edited text. We invited authors who we know to be experts in both psychological research and practice to contribute chapters in their particular areas of expertise. This has the advantage of each topic being presented by authors who are experienced in applying the concepts presented in the chapter and are enthusiastic about how that information can be used by both a practitioner and a researcher. Readers can be sure that the techniques described in this book are used every day to advance knowledge and practice in psychology. The information may at times be complex, but it is never only of interest in the ivory tower. The book reflects research in the real world.

This book is divided into four parts. Part I, Research Foundations, contains four chapters that form the basis for understanding the material in the rest of the book. Part II, Research Strategies, consists of five chapters covering the most important research strategies in clinical and counseling psychology. Each of these chapters includes an illustration and analysis of a study using that strategy, explaining the important decision points encountered by the researcher and, also, how the results of the study can be used to inform practice. Part III, Research Practice, consists of three chapters covering issues related to actually doing planning, and interpreting research and research literature. Finally, Part IV, Special Problems, includes six chapters. Chapter 13, addresses one of the most important controversies in mental health research today: The distinction between "gold standard" efficacy studies and more realistic effectiveness studies. This is followed by a chapter on the challenges in conducting research in various cultures; a topic that is gaining more and more importance as the world changes. These chapters nicely set the stage for Chapter 15, which addresses how a psychologist in practice can operate an empirically-oriented practice and can actually do research in that arena. The remaining three chapters in this section illustrate how to perform research with families, children, and the elderly, respectively. Overall, the coverage of the book gives students all the relevant information needed while still staying at a size appropriate for a semester long course.

Jay C. Thomas
Michel Hersen
Hillsboro, Oregon

Editors

Jay C. Thomas, PhD, ABPP, is Distinguished University Professor and Assistant Dean in the School of Professional Psychology at Pacific University, Hillsboro, Oregon. He received his PhD from the University of Akron in Industrial and Organizational Psychology and is certified as a specialist from the American Board of Professional Psychology (ABPP) in that specialty and in Organizational and Business Consulting Psychology. After many years as a consultant and in private practice he joined Pacific's School of Professional Psychology. His research interests include program evaluation, methodology for field studies, outcome research, and integrating findings and concepts across I/O and clinical/counseling psychology. His teaching has included statistics, research methods, program evaluation, and courses in organizational behavior and consultation, organizational assessment, and career development. He is co-editor of fifteen books, the author of one, and is on the editorial boards of *Aggression and Violent Behavior, Journal of Anxiety Disorders*, and *PsycCRITIQUES: Contemporary Psychology.* He has also published several papers in I/O psychology, program evaluation, and mental health.

Michel Hersen, PhD, ABPP, is Dean of the School of Professor Psychology, Pacific University. He earned his doctorate at the State University of New York at Buffalo, did post-doctoral work at the West Haven Veterans Administration Hospital (Yale University School of Medicine Program). He is a certified specialist in clinical psychology by the American Board of Professional Psychology, a Fellow of the American Psychological Association, past president of the Association for Advancement of Behavior Therapy. In a career spanning four and a half decades he published 226 scientific papers, co-authored and co-edited over 140 books, and is the editor of seven psychological journals, including *Behavior Modification, Aggression and Violent Behavior: A Review Journal, Clinical Psychology Review, Journal of Family Violence*, and *Journal of Developmental and Physical Disabilities.* He is also editor-in-chief of the *Journal of Anxiety Disorders* and *Clinical Case Studies.* Dr. Hersen's research interests are broad, although centered on the development and testing of behavioral approaches to psychological treatment across the life-span.

Contributors

Jennifer Bolden, MS
University of Central Florida
Orlando, Florida

Janice Ka Yan Cheng, BA
University of California
Davis, California

Kyong-Mee Chung, PhD
Yonsei University
Seoul, South Korea

Paul W. Clement, PhD, ABPP
Private Practice
South Pasadena, California

Stanley H. Cohen, PhD
West Virginia University
Morgantown, West Virginia

Lisa R. Christiansen
Pacific University
Hillsboro, Oregon

V. Mark Durand, PhD
University of South Florida
St. Petersburg
St. Petersburg, Florida

Rose F. Eagle, PhD
The Children's Program
Portland, Oregon

Leilani Feliciano, PhD
University of Colorado
at Colorado Springs
Colorado Springs, Colorado

Kurt Freeman, PhD
Oregon Health & Science University
Portland, Oregon

Lisa Selthon French, PsyD
United States Air Force
San Antonio, Texas

Isaac P. Gilman, MLIS
Pacific University
Hillsboro, Oregon

Gerald Goldstein, PhD
VA Pittsburgh Healthcare
System
Pittsburgh, Pennsylvania

Alan Gross, PhD
University of Mississippi
Oxford, Mississippi

Stephen N. Haynes, PhD
University of Hawaii
at Manoa
Honolulu, Hawaii

Allison A. Jay, MA
University of Colorado
at Colorado Springs
Colorado Springs, Colorado

Amanda M. McEnery, MA
Indiana University
Bloomington, Indiana

Heidi Meeke, MA
Pacific University
Hillsboro, Oregon

Paul G. Michael, PhD
Pacific University
Hillsboro, Oregon

Catherine Miller, PhD
Pacific University
Hillsboro, Oregon

Karl A. Minke, PhD
University of Hawaii at Manoa
Honolulu, Hawaii

Tracy L. Morris, PhD
West Virginia University
Morgantown, West Virginia

Katherine H. Moyer, MA
University of Mississippi
University, Mississippi

Mark D. Rapport, PhD
University of Central Florida
Orlando, Florida

Johan Rosqvist, PsyD
Pacific University
Hillsboro, Oregon

Joseph R. Scotti, PhD
West Virginia University
Morgantown, West Virginia

Daniel L. Segal, PhD, ABPP
University of Colorado
at Colorado Springs
Colorado Springs, Colorado

Thomas L. Sexton, PhD, ABPP
Indiana University
Bloomington, Indiana

Elizabeth E. Stacom, MS
West Virginia University
Morgantown, West Virginia

Mary E. Steers, MA
University of Colorado
at Colorado Springs
Colorado Springs, Colorado

Leslie Sue, PhD
University of La Verne
La Verne, California

Stanley Sue, PhD
Palo Alto University
Palo Alto, California

Jay C. Thomas, PhD, ABPP
Pacific University
Hillsboro, Oregon

Paula Truax, PhD
Kaiser Permanente
Portland, Oregon

Mo Wang, PhD
University of Maryland
College Park, Maryland

Michael C. Winder
University of California, Merced
Merced, California

Alyson Williams, PhD
Pacific University
Hillsboro, Oregon

Laura R. Wilson, MEd
Indiana University
Bloomington, Indiana

Brian Yochim, PhD, ABPP
Sierra Pacific Mental Illness
Research, Education,
and Clinical Centers
(MIRECC)
Palo Alto, California

I
Research Foundations

1
Introduction

Science in the Service of Practice

JAY C. THOMAS and JOHAN ROSQVIST

Contents

Today, psychologists are called on to help solve an ever widening scope of personal and social problems. It has been recognized that a large proportion of the population can benefit from psychotherapeutic services. Current estimates of the prevalence of mental disorders indicate that such disorders are common and serious. The National Institute of Mental Health (2009) estimates that up to one in four American adults suffer from a diagnosable mental disorder. The provision of psychotherapy services is a multibillion-dollar industry (Sexton, Whiston, Bleurer, & Walz, 1997), with certain very common phenomena (i.e., anxiety disorders) representing economic burdens to a tune of $42.3 billion annually of the total U.S. mental health bill of $148 billion (Greenberg et al., 1999). In addition, clinical and counseling psychologists are asked to intervene in prevention efforts and in situations involving individuals and families, prisons, and schools, along with playing their role in industrial and organizational work settings.

When so many people trust the advice and assistance of psychologists and counselors, it is important that professionals rely on a foundation of knowledge and evidence that is known to be tried and tested. Many students in clinical and counseling psychology wonder about the relevance of research courses and of research in general pertaining to their chosen profession, which mirrors a field in which science and practice of psychotherapy almost invariably inhabit different worlds (Lebow, 2006). These students often primarily value the role of the psychologist as helper and expect to spend their careers "helping" clients in dealing with important issues. This is indeed a very worthy ambition, but we argue that "effective" *helping* can occur only when the best techniques are utilized, and that it is only through scientific research that we can determine what is "best." We illustrate this point through a brief history of treatment for obsessive–compulsive disorder (OCD) in which a client, Sue, received the assistance she needed through the systematic, targeted application of an empirically based treatment.

The Case of Sue

Sue, a 28-year-old married woman, was engaged in a broad range of avoidant and compulsive behaviors (Rosqvist, Thomas, Egan, Willis, & Haney, 2002). For example, she executed extensive checking rituals—hundreds of times per day—that were aimed at relieving obsessive fears that she, by her thoughts or actions, would be responsible for the death of other people (e.g., her one-year-old child, her husband, other people whom she cared for, and sometimes even strangers). She was intensely afraid of dying herself. She also avoided many social situations because of her thoughts, images, and impulses.

As a result of these OCD symptoms and the resultant avoidant behavior, Sue was left practically unable to properly care for herself and her child. In addition, she was grossly impaired in her ability to perform daily household chores, such as grocery shopping, cleaning, and cooking. Her husband performed many of these activities for her, as she felt unable to touch many of the requisite objects, like pots and pans, food products, cleaning equipment, and so on.

Additionally, Sue was unable to derive enjoyment from listening to music or watching television because she associated certain words, people, and noises with death, dying, and particular fears. She also attributed the loss of several jobs to these obsessions, compulsions, and her avoidant behavior. Sue reported feeling very depressed due to the constrictive nature of her life that was consumed with guarding against excessive and irrational fears of death.

Sue eventually became a prisoner of her own thoughts and was unable do anything without horrendous feelings of fear and guilt. For all intents and purposes, she was severely disabled by her OCD symptoms, and her obsessions, compulsions, and avoidant behavior directly impacted her child and husband.

In fact, her fears were so strong that she eventually became uncertain that her obsessions and compulsions were irrational or excessive and unreasonable (i.e., she demonstrated "poor insight"). She strongly doubted the assertion that her fears will not come true, although she had little, if any, rational proof for her beliefs. She was unsuccessful in dismissing any of her obsessive images, impulses, or thoughts and beliefs. She had very little relief from the varied intrusions, and she reported spending almost every waking hour on some sort of obsessive-compulsive behavior. She felt disabled by her fears and doubts, and felt that she had very little control over them.

Obviously, Sue was living a life of very low quality. Over the course of some years, she was treated by several mental health practitioners and participated in many interventions, including medication of various kinds, psychodynamic, interpersonal, supportive, humanistic, and cognitive behavioral therapies, both individually and in groups, as both an inpatient and an outpatient. However, Sue made little progress and was considered for high-risk neurological surgery. As a last-ditch effort, a special home-based therapy emphasizing exposure and response prevention (ERP) along with cognitive restructuring was devised. This treatment approach was chosen because the components had the strongest research basis and empirical support. Within a few months, her obsessive and compulsive symptoms remitted, and she eventually became sufficiently free of them to return to work and a normal family life. Thus, by the application of research-based treatment, Sue, who was previously considered "treatment refractory," was effectively helped to regain her quality of life.

The Role of Research in Treatments for Obsessive-Compulsive Disorder (OCD)

OCD has a long history. For example, Shakespeare described the guilt-ridden character of Lady Macbeth as prone to obsessive hand washing. Other, very early, descriptions of people with obsessional beliefs and compulsive behaviors also exist, such as people having intrusive thoughts about blasphemy or sexuality. Such people were frequently thought (both by the sufferer and the onlooker) to be possessed, and they were typically "treated" with exorcisms or other forms of what would now be deemed torture.

Obsessions and compulsions were first described in the psychiatric literature in 1838, and throughout the early 1900s, it received attention from pioneers such as Pierre Janet (1859–1949) and Sigmund Freud (1856–1939); however, OCD remained virtually an intractable condition, and patients suffering from it were frequently labeled as psychotic and little true progress was thought possible. That was until the mid 1960s, when Victor Meyer in 1966 first described the successful treatment of OCD by ERP (Meyer, 1966).

Since Meyer's pivotal work, the behavioral and cognitive treatment of OCD has been vastly developed and refined. Now, it is generally accepted that 60%–83% of patients can make significant improvement with specifically designed techniques (Foa, Franklin, & Kozak, 1998; Salomoni et al., 2009). Also, patients who, initially, prove refractory to the current standard behavioral treatment can achieve significant improvement with some additional modifications. In any case, OCD does not appear to be an incurable condition any longer.

This change has been made possible only by the systematic and deliberate assessment and treatment selection for such patients. That is, interventions for OCD, even in its most extreme forms, have been scientifically derived, tested, refined, retested, and supported. Without such a deliberate approach to developing an effective intervention for OCD, it would possibly still remain intractable (as it mostly was just 35 years ago). In truth, recalcitrance is largely a myth promulgated by people who drift away from science-informed or evidence-driven treatment (Waller, 2009).

The empirical basis of science forms the basis for effective practice, such as what has made OCD amenable to treatment. Such empirical basis is embodied in the scientific method, which involves the systematic and deliberate gathering and evaluation of empirical data, and generating and testing hypotheses based on general psychological knowledge and theory, in order to answer questions that are answerable and "critical."

The answers derived should be proposed in such a manner that they are available to fellow scientists to methodically repeat them. In other words, science, and professional effectiveness, can be thought of as the observation, identification, description, empirical investigation, and theoretical explanation of natural phenomena.

Ideally, conclusions are based on observation and critical analyses, and not on personal opinions (i.e., biases) or authority. This method of reaching conclusions is committed to empirical accountability, and in this fashion, it forms the basis for many professional regulatory bodies. It remains open to new findings that can be empirically evaluated to determine their merit, just as the professional is expected to incorporate new findings into how he or she determines a prudent course of action.

Consider, for example, how the treatment of obsessions has developed over time. Thought stopping technique is a behavioral technique that has been used for many years to treat unwanted, intrusive thoughts. In essence, the technique calls for the patient to shout "stop" or make other drastic responses to intrusions (e.g., clapping hands loudly, or snapping a heavy rubber band worn on his or her wrist) in order to extinguish the thoughts through a punishment paradigm. It has since been determined that thought-suppression strategies for obsessive intrusions may have a paradoxical effect (i.e., reinforcing the importance and veracity of the obsession by specifically focusing attention and energy on it) rather than the intended outcome (Rosqvist, 2005). It has been established, through empirical evaluation and support, that alternative, cognitive approaches (e.g., challenging the content of cognitive distortions)—like correcting overestimates of probability and responsibility—are more effective in reducing not only the frequency of intrusions but also the degree to which they distress the patient.

An alternative to thought-stopping strategy, the exposure-by-digital-loop method, has been systematically evaluated and its effectiveness has been scientifically supported. In this technique, the patient is exposed to endless streams of "bad" words, phrases, or music. As patient's obsessions frequently center on the death of loved ones, they may develop substantial lists of words that are anxiety producing (e.g., Satan, crib death, sudden infant death syndrome [SIDS], devil, casket, coffin, cancer). These intrusive thoughts, images, and impulses are conceptualized as "aversive stimuli," as described by Rachman (Emmelkamp, 1982). Such distortions and intrusions are now treated systematically by exposure by digital loop (and pictures) so that the patient can habituate to the disturbing images, messages, and words. This procedure effectively reduces their emotional reactivity to such intrusions and lowers overall daily distress levels. Reducing this kind of reactivity appears to allow patients to more effectively engage in ERP (van Oppen & Emmelkamp, 2000; Wilson & Chambless, 1999; van Oppen & Arntz, 1994).

The point of this OCD example is to show that over time more and more effective methods of treatment are developed by putting each new technique to empirical testing and refining it based on the results. In addition, the research effort has uncovered unexpected findings, such as the paradoxical effect of thought suppression. The traditional thought-stopping technique is in essence a method of thought suppression, whereby the individual by aversive conditioning attempts to suppress unwanted thoughts, images, or impulses. However, systematic analyses have revealed that efforts at suppressing thoughts (or the like), in most people, lead to an increased incidence of the undesired thoughts. It is much like the phenomenon of trying to not think about white bears when instructed to not think about them; it is virtually impossible! What has been supported as effective in reducing unwanted thoughts, whether about white bears, the man behind the curtain, or about germs and death, is exposure by loop. This method does not attempt to remove the offending thought, but "burns it out" (i.e., reaction to specific content) through overexposure.

In light of this experience, it is prudent for the professional to incorporate these techniques into treating intrusive thoughts. Although a therapist may be very familiar with thought-stopping strategy, it is reasonable to expect that the scientifically supported techniques will be given a higher value in the complete treatment package. This follows the expectations of many managed care companies, and this also adheres to the ethical necessity to provide the very best and most appropriate treatment possible for any given clinical presentation. To do anything less would be a great disservice to the patient, as well as put him or her into possible jeopardy by providing substandard care.

In these days of professional accountability and liability for a product, it has become necessary to be able to clearly demonstrate that what we do is prudent given the circumstances of any particular case. Most licensing boards and regulatory bodies will no longer accept arbitrary, individual decisions on process; rather, they dictate and expect that a supported rationale is utilized in the assessment and treatment process.

With this in mind, it has become increasingly necessary, if not crucial, that the professional engage in a systematic method for assessment and treatment selection in order to create the most effective interventions possible (given current technology and methodology). Today, the empirical basis of science forms the basis of effective practice.

Scientific Method and Thought

Early in the twentieth century, the great statistician Karl Pearson became embroiled in a heated debate over the economic effects of alcoholism on families. Typical of the scientific battles of the day, the issue was played out in the media with innuendoes, mischaracterizations and, most importantly, spirited defense of pre-established positions. Pearson, frustrated by the lack of attention on the central issue, raised a challenge that we believe serves as the foundation for any applied science. “Pearson’s challenge” was worded in the obscure language of his day, and has been updated by Stigler (1999) as follows:

> If a serious question has been raised, whether it be in science or society, then it is not enough to merely assert an answer. Evidence must be provided, and that evidence should be accompanied by an assessment of its own reliability (p. 1).

Pearson went on to state that adversaries should place their “statistics on the table” (Stigler, 1999, p. 1) for all to see. Allusions to unpublished data or ill-defined calculations were not to be allowed. The issue should be answered by the data at hand, and everyone was free to propose their own interpretations and analyses. These interpretations were to be winnowed out by the informed application of standards of scientific thought and method. This required clear and open communication of methods, data, and results.

The classic scientific method involves the objective, systematic, and deliberate gathering and evaluation of empirical data, and generating and testing hypotheses based on general psychological knowledge and theory, in order to answer questions that are answerable and critical. The answers derived should be proposed in such a manner that they are available to fellow scientists to methodologically repeat them. Conclusions are based on observation and critical analyses, and not on personal opinions (i.e., biases) or authority. This method of reaching conclusions is committed to empirical accountability. It is open to new findings that can be empirically evaluated to determine their merit. Findings are used to modify theories in order to account for discrepancies between theory and data. Results are communicated in detail to fellow scientists.

We accept the general outline of the scientific method just described. It has had its critics who object to one or another of the components. We will explore each component in somewhat more detail and address some of the more common objections.

Objective, Systematic, and Deliberate Gathering of Data

All research involves the collection of data. Such data may be collected from self-report, surveys, tests, or other psychological instruments; physiological measurement; interview; or a host of other sources. The most common approach is to design a data collection procedure and actually collect data purposely for a particular study. It is also possible to perform archival studies, in which data that might bear on an issue are pulled from files or other archival sources, although the information was not originally collected for that purpose. In either case, the idea is to obtain information, which is free from the investigator's expectations, values, and preferences, as well as from other sorts of bias. Originally, it was expected that data that was completely free from bias and atheoretical could be obtained. Although this has not been proved possible, objectivity in data gathering as well as analysis and interpretation remains the goal for the scientist. No other aspiration has been proved as effective (Cook, 1991; Kimble, 1989).

Generating and Testing Hypotheses

Hypotheses are part of everyday life in psychological practice. A treatment plan, for example, contains implicit or explicit hypotheses that a particular intervention will result in an improvement in a client's condition. In the case of Sue, the hypothesis was that home-based ERP would reduce her OCD symptoms to the point where she would no longer be a potential candidate for neurosurgery. Many research hypotheses are more complex than that one, but they serve an important purpose in meeting Pearson's challenge. They specify what data are relevant and predict in advance what the data will show. Hypotheses are derived from theories, and it is a poor theory that fails to allow us to make relevant predictions. Thus, by comparing our predictions against the obtained data, we put theories to test.

Theories are used to summarize what is known and to predict new relationships between variables and, thus, they form the basis of both research and practice. John Campbell (1990) provided an overall definition of theory as " … a collection of assertions, both verbal and symbolic, that identifies what variables are important for what reasons, specifies how they are interrelated and why, and identifies the conditions under which they should be related or not related" (p. 65). Campbell goes on to specify the many roles that a theory may play:

1. Theories tell us that certain facts among the accumulated knowledge are important, and others are not.
2. Theories can give old data new interpretations and new meaning.
3. Theories identify important new issues and prescribe the most critical research questions that need to be answered to maximize understanding of the issue.
4. Theories provide a means by which new research data can be interpreted and coded for future use.
5. Theories provide a means for identifying and defining applied problems.
6. Theories provide a means for prescribing or evaluating solutions to applied problems.
7. Theories provide a means for responding to new problems that have no previously identified solution strategy (Campbell, 1990, p. 65).

From abstract theories we generate generalizations, and from generalizations specific hypotheses (Kluger & Tikochinsky, 2001). A useful theory allows for generalizations beyond what was previously known and often into surprising new domains. For example, Eysenck's (1997; cited in Kluger & Tikochinsky, 2001) arousal theory of extroversion predicts that extroverts will not only prefer social activities but also other arousing activities like engaging in crimes such as burglary.

Popper (1959), one of the most influential philosophers of science, maintained that it is not possible to confirm a theory, but instead all we can do is disconfirm it. If our theory is "all ravens are black" (this is a classic example dating back to the ancient Greeks), all we can say by way of confirmation is that we have not observed a nonblack one. However, observing a single nonblack raven is sufficient to disprove the theory. The problem is compounded by the fact that one day one of the authors of this chapter (Jay C. Thomas) observed a raven, or what he thought was a raven, and in the early morning sunlight its feathers had a dark blue iridescent sheen. Thomas concludes that the theory "all ravens are black" is disproved. But, two issues remain. Is a "blue iridescent sheen" over a basically black bird what we mean by a nonblack raven? Second, how do we know it was a raven? Although Thomas reports seeing such a raven, Johan Rosqvist retorts that Thomas is by no means a competent ornithologist and his description cannot be trusted and, consequently, the theory has not been disproved.[1] Before we can put a theory to a convincing test, we must be very careful to specify what we are looking for. This level of attention to detail has been rare in psychology. It is sometimes noted that few theories have ever been completely rejected on the basis of research evidence (Mahrer, 1988). There are two major reasons for reaching this conclusion.

One is the naive confusion of null hypothesis significance testing (NHST) from inferential statistics with theory testing, or, as Meehl (1997) prefers to call it, “theory appraisal.” The NHST is a tool for the researcher to use, just as a carpenter may use a hammer for joining boards. But it is not the only tool or even necessarily the optimal one. This testing has many problems, as described in Chapter 9, and the method itself has little to do with theory testing (Meehl, 1997).

The second reason why psychology has so often failed to reject theories is because of the occurrence of “auxiliary theories” (Lakatos, cited in Serlin & Lapsley, 1993; Meehl, 1997). Auxiliary theories are not part of the content of a theory, but they are present when we try to put the theory into action, that is, when we try to test it. The problem with auxiliary theories is that the validity of one or more auxiliary theories may impact the results of a study so that it is not possible to determine whether the results bear on the original theory. In the case of Sue, we had a hypothesis that home-based ERP would change her OCD symptoms. This hypothesis was derived from ERP theory in response to the failure of ERP to have any effect on the patient condition in its usual clinic-based administration. One auxiliary theory related to Sue’s treatment was that ERP therapy was competently conducted. Had the therapy failed, we would be more inclined to suspect a problem in implementation than a problem in the theory itself. Auxiliary theories reside in almost every aspect of research, from instrumentation to design and analysis. Later, when we examine the hallmarks of gold standard clinical research in Chapter 11, it is seen that the standard has been designed to minimize the ability of auxiliary theories to influence our conclusions.

Replication of Research Findings

Replication is critical for science. A given finding may be the result of many factors besides the effects specified by a theory or the researcher. Random chance is a common culprit; others include unusual features of a study’s design, biased sampling or observation, inconclusive statistical analyses, and even the researcher’s hopes and dreams. The most famous instance of this effect in recent years is that of “cold fusion.” Cold fusion was the supposed fusion of two atomic nuclei at much lower temperatures than previously thought possible. If such a thing were possible, the world would have been vastly changed by the availability of abundant, inexpensive, and non-polluting power. Such a development would have had unimaginable benefits. There was one problem. The effect could not be consistently obtained in other laboratories (Park, 2000). Not only did other laboratories find it impossible to duplicate the energy release predicted by cold fusion but they could also not observe the expected by-products of fusion such as lethal doses of nuclear radiation. Today, the cold fusion concept is stone-cold dead.

Science relies on two types of replications. "Exact replication" involves repeating the original study in every detail to see if the same result is obtained. This is what the replicators of cold fusion set out to do; but they were hampered by the failure of the original "discoverers" to provide sufficient detail about the experiment. Cold fusion as a research topic lasted a bit longer because of this, but met its demise despite its originators' obstructionism. Psychology has not done well by exact replication. Journals prefer to publish original findings and are rarely interested in exact replications. This has led to an emphasis on "conceptual replications," that is, testing the same or a similar hypothesis, but using different measures or conditions. The idea seems to be that if the effect is large enough, it will again be observed. The problem is that when an effect is not replicated, we do not know why. It could be the original finding was spurious, the changes in the research design were sufficient to mask or eliminate the finding, or the replication may have lacked sufficient power to detect the effect.

Limitations of conceptual replications are illustrated by a recent controversy on the value of a recently introduced psychotherapy technique, eye movement desensitization and reprocessing (EMDR). The original developer of EMDR, Francine Shapiro, and the proponents of the method had reported substantial success with this technique. However, other researchers failed to obtain positive results. Shapiro (1999) argued that the failed replications have been characterized by inadequate treatment fidelity. In other words, the studies did not properly implement the technique, and so the failure to replicate results is not surprising. Rosen (1999), meanwhile, contends that the issue of treatment fidelity is a red herring that distracts the reader from a negative evaluation of a theory and permits its perpetuation. This is an example of an auxiliary theory in action. On the one hand, the EMDR theory is protected by the supposedly inept implementation of EMDR practice, whereas on the other, if there is anything to the theory it should work despite imperfect fidelity. We take no position on the issue except to note three things: (1) This controversy would not exist if exact replication had been attempted. (2) Although claims of inadequate treatment fidelity may well be a legitimate issue, this general tactic is often abused and its employment has been a red flag throughout history (cf. Park, 2000; Shermer, 2001). (3) Conscientious researchers examine their own findings from many angles to ensure that they have eliminated as many competing explanations as possible. This may mean running studies two, three, or more times with slight modifications for the researchers to determine by themselves how robust their findings are.

We cannot replicate many natural phenomena; natural catastrophes and the horrors of war are two examples. We can still fulfill the replication requirement in two ways. First, we can attempt to combine the observations of multiple observers. Bahrick, Parker, Fivush, and Levitt (1998) examined the impact of varying levels of stress on the memories of young children about Hurricane Andrew. Children between the ages of three and four were interviewed a few months after the hurricane about what had happened during the storm. Interviews were recorded and scored for several facets of memory. By having two raters score each transcript and by comparing their scoring, Bahrick et al. (1998) demonstrated that similar scores would be derived by different raters. This represents a replication within the study. Bahrick et al. (1998) also provided detailed information about how the data were collected and the nature of the analyses that had been carried out by them. This makes it possible for other researchers to attempt to replicate the results after some other disaster. We would expect the impacts of hurricanes, tornadoes, floods, and the like to be comparable, and that other researchers can replicate the results following another disaster. Thus, whereas exact replication is impossible in these cases, conceptual replication is possible and it should be expected to establish the validity of any important finding from such circumstances.

Modification of Theories Using Findings

Good theories account for past results. They also predict new results beyond what other theories are capable of predicting. Unfortunately, sometimes the data do not support the theory. Although this may be due to some of the reasons we have already presented, it may be that the theory is actually wrong in some respects. We expect our theories to be wrong in at least some respects. That is why we test them. Still, many researchers, particularly, those just beginning their careers, will often conclude that "they" have failed when the data do not come out as expected. If the idea were sound in the first place and the study has been conducted as well as possible, then failure of a prediction is an opportunity to learn more and create an even better understanding of behavior. Petroski (1985), a noted structural engineer, made the case that without failure, engineering would not advance. That the Roman aqueducts have stood for hundreds of years is instructive, but studying the collapse of a newly built bridge can be even more so. Applied psychology is like engineering in this respect; we must learn from failure. It is a rare theory that does not change over time to accommodate new findings. The modified theory should make predictions different from those of the old one and, thus, needs to be tested again. Critics of theory testing may be correct in stating that often theories do not die out from lack of empirical support, but these critics forget that theories evolve. Perhaps the most memorable statement to this effect is that of Westen (1998), who wrote on the scientific legacy of Sigmund Freud: "Freud's critics largely lambast the theory as it stood in the early 1920s while the theory had changed substantially by the time Freud died in 1939 even though since then 'he has been slow to undertake further revisions'" (p. 333).

Clear and Open Communication of Methods, Data, and Results

If it is not answered, Pearson's challenge means nothing. Research must include the dissemination of results so that others can study, evaluate, and contest or use them. In the cold fusion debacle, what irreparably damaged the researchers' reputations in the scientific community was not that they made an error—that could, and should, happen in cutting-edge research—but that they refused to divulge details of their procedure, making it difficult to replicate and evaluate the phenomenon (Park, 2000). There are norms in science for effectively communicating information. The *Publication Manual of the American Psychological Association* (APA: APA, 2010) provides guidelines for what information should be included in research reports. In addition to following these guidelines, researchers are expected to make copies of their data available to others on request. Of course, care must be taken to ensure that all participant-identifying information has been removed so that there is no possible breach of confidentiality (cf. Chapter 10).

Theory of Causality

Clinical and counseling psychology can get by with a straightforward theory of causality. Interventions, such as psychotherapy, are implemented because it is assumed that the intervention causes changes in the clients. Similarly, life events are often expected to cause changes in people, which may later lead them to become clients (Kessler, 1997). But, it is a big leap from believing there is a causal relationship to developing a convincing demonstration for the actual existence of that relationship in a causal fashion.

The nature of causality and the proof of causality have been a favorite topic of philosophers for centuries. The most widely employed analysis was the one formulated by the nineteenth-century philosopher John Stuart Mill. His formulation (cited in Shadish, Cook, & Campbell, 2002) consisted of three tests: (1) The cause must precede the effect in time, (2) the cause and the effect must covary, and (3) there must be no other plausible explanations for the effect than the presumed cause.

Mill's Requirement 1: Cause Must Precede the Effect

This is the least controversial of Mill's tests. Due to the lack of a time machine, no one has ever figured out how to change an event after it has happened. It is very unlikely that a researcher would make the error of attributing the status of cause to something that occurred after the observed effect. However, comparable errors are sometimes made in cross-sectional studies in which two variables are measured at the same time. Although we may have a theory that self-esteem has a causal influence on school performance, we may measure both variables at the same time and no causal conclusions can be drawn. Sometimes, a study will be retrospective in nature: People are asked to remember their condition prior to a given event, for example, how much alcohol they consumed a day prior to the onset of some disease or the occurrence of an accident. Unfortunately, circumstances that arise after the event has occurred may influence memory (Aiken & West, 1990), so the timing of the variables is now reversed: The effect (disease or accident) now precedes the presumed cause (amount of alcohol consumed), and no causal conclusions can be drawn.

Mill's Requirement 2: Cause and Effect Must Covary

In a simple world, this test would specify that when the cause is present the effect must be present, and when the cause is absent the effect must be absent. Unfortunately, we do not live in such a simple world. Take a dog to a park and throw a stick. That action is sufficient to cause the dog to run. But dogs run for other reasons too, such as a squirrel digging in the dirt nearby. Throwing the stick is not a necessary cause for the dog to run. "Sufficient causes" are those that by themselves "may" cause the effect but do not have to consistently result in the effect. For example, a well-trained guide dog on duty when the stick is thrown will probably not run. "Necessary causes" must be present for the effect to occur, but they do not have to be sufficient. Driving too fast may be a necessary cause for a speeding ticket, but most drivers have exceeded the speed limit on occasion without getting cited. As if this is not confusing enough, consider the case of schizophrenia. Schizophrenia is thought to have a genetic basis; yet, a family background cannot be found in all people with schizophrenia, indicating that there are other causal factors (Farone, Tsuang, & Tsuang, 1999). Many people appear to have at least some of the genes related to schizophrenia, but show no symptoms. Thus, a family background of schizophrenia can be considered a "risk factor" for schizophrenia. If the genetic background is present, schizophrenia is more likely than if the family background is not present. Risk factors may or may not have a causal relationship with an event; they may simply be correlated with it.

"Correlation does not prove causation" is a statement whose significance every aspiring psychologist should learn to appreciate. The statement says that Mill's second criterion is a necessary but not sufficient reason to attribute causality. A study may find a negative correlation between depression and self-esteem such that people with lower self-esteem are found to report higher levels of depression. The temptation is to conclude that people are depressed because they have low self-esteem (and that by raising self-esteem, depression would be reduced). This temptation must be resisted because nothing in the data lends support to a causal inference. Seligman, Reivich, Jaycox, and Gillham (1995) cogently argued that there may be a third factor that causes both low self-esteem and depression. Seligman and his colleagues have gone so far as to argue that ill-advised attempts to raise self-esteem in the general population may have set up many people for a propensity toward depression. So, we must be very careful and not assume that a correlational relationship implies a causal relationship.

Sometimes a third variable influences the causal relationship between two others. It has often been noted that even the best psychological interventions fail to help some people. Prochaska and DiClemente (cf. Prochaska, 1999) postulated that clients may have differential readiness to change. Some may have never considered making changes in their lives or do not wish to do so. It is unlikely that such clients will benefit from interventions designed to create change, whereas clients motivated to change may well benefit from such therapies. What is variously called *stage of change* or *readiness to change*, if supported by further research, could be a moderator of the causal impact of psychotherapy on a client's outcome.

Mill's second test gets even more complicated when we consider the possibility of "reciprocal causation." Sometimes, two or more factors cause each other. A basic tenet of economics lies in the relationship between supply and demand. If a desirable good is in short supply, its demand increases. As demand increases, producers initiate a ramp-up in production until it eventually satiates demand, which then falls. Thus, supply and demand are reciprocally related. Psychology does not have such well-defined examples, but there are probably many cases of reciprocal causation. Lewinsohn's (1974) behavioral theory of depression, for example, postulates that lack of reinforcement leads to a depressed mood; this mood leads to less activity, which, in turn, leads to less reinforcement. A study that examines these factors at only two points in time will miss this reciprocal relationship.

The statement "correlation does not prove causation" does contribute its share of mischief to the field due to a misunderstanding of the meaning of "correlation." Correlation in this sense refers to the co-occurrence of two or more variables. It does not refer to the set of statistics known as coefficients of correlation. No statistic or statistical procedure indicates or rules out causation. Our ability to infer causation depends on the study design, not the statistical analysis of data. Although some analytic methods have been developed to facilitate the investigation of causation, the conclusions regarding possible causal relationships depend on how, where, when, and under what conditions the data were gathered.

Mill's Requirement 3: There Must Be No Other Plausible Explanations for the Effect Other Than the Presumed Cause

Mill's third requirement is the one that causes the most problems for researchers and, except for effectiveness research, most study designs have been developed with this requirement in mind. Sherlock Holmes once told Dr. Watson that " … when you have eliminated the impossible, whatever remains, *however improbable* (emphasis added), must be the truth" (Doyle, 1890/1986, p. 139). But, if Holmes cannot eliminate the alternatives as being impossible, then he cannot deduce the answer. There are innumerable alternative causes of an observed effect in psychological research. Consider a study comparing two different treatments for OCD. Sampling may be faulty; assigning people to different treatments in a biased manner eliminates our ability to say that one treatment caused greater change than another. Failure to control conditions may influence the results; for example, if people in one treatment have a friendly, warm, empathic therapist while those in another treatment have a cold, distant therapist, we cannot determine if any observed effect was due to differences in the treatment or differences in the therapists.

The key in Mill's third criterion is to rule out "plausible" alternative explanations. It takes a great deal of expense and trouble to control outside factors that might contaminate results. Therefore, we expend most of our budget and effort in controlling those that offer the most compelling alternative explanations. Space aliens could abduct the members of one of our study's treatment groups and subject them to some strange "cure"; but this possibility is considered so improbable that no one ever controls for the effects of an alien abduction. Outside the bizarre, deciding which alternatives are plausible requires an understanding of the rationale underlying research design and the phenomenon under study. As a consumer of research, you need to pay close attention to the Method section of research articles because that is where you will find how the researchers chose to control what they believed were the most plausible alternative explanations, the Results section because more control is exerted there, and the Discussion section because that is where researchers often confess to any remaining limitations of the study.

Science in the Service of Practice

Influential clinicians recognized a few years ago that it was desirable to carefully examine and enumerate those treatments that could be described as having shown to have an efficacious effect on client outcomes (Seligman, 1998a). A consensus developed that professional psychologists should be competent in scientifically engaged practice (Kaslow, 2004). This led to an ambitious effort by the Society for Clinical Psychology (Division 12 of the APA) to do exactly that. The findings, first published in 1995 (Division 12 Task Force, 1995), were controversial in that many popular methods in long use did not make the list. How can this be? Usually, it was not so much a consequence of documented treatment failures as a paucity of outcome research on these treatments (Seligman, 1998b). It could not be determined that those treatments are effective because adequate studies had not been conducted. The Division 12 effort continues, updates are periodically posted on the Society's Web page (http://www.apa.org/divisions/div12/homepage.shtml). In addition, an APA Presidential Task Force (2006) made important recommendations on how research evidence can be effectively implemented in practice. It is important for clinical and counseling psychologists to develop the knowledge and skills to interpret the results of research, if not to contribute to it, because research results have shaped practice and will do so to an even greater extent in the coming years.

Because of stories like Sue's, clinical and counseling psychologists have an interest and responsibility in demonstrating that their interventions are effective and in using the scientific method in advancing practice. A third-party payer (i.e., an insurance company or a government body) also has a legitimate interest in verifying that the services it pays for are effective, and clients and their families are also concerned that treatments result in real change (Newman & Tejada, 1996). Still, some practitioners ask, "What difference does it make if our clients feel better after therapy? Do we really need to fuss around with all this research stuff if it is secondary to feeling better?" These questions were actually raised by a graduate student in the senior author's research methods class. Despite the author's own apoplexy in response to the question, these are legitimate and proper issues that must be raised. They deserve an answer. If "feeling better" is the objective of the work with a client, then how are other outcomes relevant, as assessed by standardized measures? If the outcomes employed in outcome studies are not relevant, then the studies themselves are a poor foundation for practice. If progress in treatment, ethics, concerns of leading thinkers, demands of third-party payers, and social imperative are not enough bases for relying on research, there is still one more excellent reason that justifies an emphasis on research-based practice. Throughout most of history, people with psychological disorders were stigmatized and denied the same rights and dignity as others (Stefan, 2001). This treatment was considered justified because such people were considered to be weak, have flawed characters, and be unreliable and, worse, unchangeable. Social and legal opinion has changed over the past 30 years or so; but those changes can only be sustained by continual rigorous demonstrations that personal change is possible, that people with disorders are not fated to a low quality of life. That is the lesson to be learned from Sue's OCD. A few years ago, she would undoubtedly be institutionalized, probably for the rest of her life. Today, as a result of effective, empirically based treatment, she has come back to work and leads a normal home life. She is indistinguishable from any other member of "normal" society. She "feels better" too.

We subtitled this chapter "Science in the Service of Practice" because, although it is possible to pursue science for its own sake, we expect that most readers of this book will be mostly interested in learning about clinical or counseling practice. Science can make for a stronger, more effective practice. So far, we have concentrated on the scientific investigation of treatment effects. Research impacts practice in many other ways: in determining causes of disorders, validation of measures, cultural effects, human development, and even practitioners' acceptance of treatment innovations (e.g., Addis & Krasnow, 2000), to name a few. The history of science shows that there have been few important scientific findings that have not had some effect on practical affairs; but when science is purposely employed to advance practice, it can be an exceptionally powerful method. Applied science differs a bit from so-called *pure* science in that some issues appear that are not the concern of the pure scientist. For example, the distinction between "efficacy" and "effectiveness" studies (see Chapter 11) does not surface in the laboratory. In efficacy studies, we are concerned with showing a causal relationship between a treatment and an outcome. Effectiveness studies are not designed to show causality, but are concerned with the conditions under which an established causal relationship can be generalized.

The Local Clinical Scientist

One model of practice that encourages the incorporation of scientific method into the provision of services is the "local clinical scientist" (Stricker & Trierweiler, 1995). This model applies to psychological science in two ways: (1) approaching the local situation in a scientific way (i.e., gathering and evaluating data, and generating and testing hypotheses based on general psychological knowledge and theory), and (2) systematically questioning how local variables impact the validity of generalizing such knowledge to the local situation. Local is contrasted with universal or general in four ways: (1) local as a particular application of general science; (2) local culture consists of persons, objects, and events in context, including the way people speak about and understand events in their lives (i.e., in the local perspective, science itself is a local culture, which practitioners bring into the open systems of their clients' local cultures); (3) local as unique (i.e., some aspects of what the practitioner observes will fall outside the domain of available science, like a local phenomenon that has not yet been adequately studied because it is not (yet) accessible to the methods of scientific inquiry; and (4) space–time local (i.e., not just the physical and temporal properties of the object under study, but also the specific space–time context of the act of judgment).

The effective local clinical scientist knows the research methods and results in the areas in which he or she works (Spring, 2007) and utilizes the scientific method in their practice. Table 1.1 illustrates how the phases of clinical practice and scientific investigation have common elements and how the scientific approach can be incorporated into practice.

Skepticism, Cynicism, and the Conservative Nature of Science

One of the authors (Jay Thomas) teaches a course in statistics. After going over one assignment with the class (reading Huff's, 1954, *How to Lie with Statistics*), one student commented that he was now more cynical than ever when it comes to reading research reports. To become cynical is to doubt the sincerity of one's fellows, to assume that all actions are performed solely on the basis of self-interest, and to trust anyone's reports is naive. Developing cynicism in students is hardly a desirable outcome of studying research and statistical methods, particularly because it is hard to believe that a cynical clinician will be very successful in practice. We do hope that students become skeptical, doubting assertions until evidence is submitted to substantiate the claims. To be skeptical is to be "not easily persuaded or convinced; doubting; questioning" (Guralnik et al., 1978, p. 1334). Effective clinicians do not believe everything they hear

Table 1.1 Incorporating Research Knowledge into Practice

Client Phase	Practice Issue	Scientific Method	Scientific Issue
1. Intake	• What brought the client here? • What is troubling the client? • What are the client's strengths, weaknesses, and resources? • What is salient about the client's background and history? • What is relevant about client's background and history for presenting problem? • What are the client's expectations about your services? • What is the client's stage of change? • Who is the client?	1. Observe	• Attend to subject expectancies, experimenter expectancies, and demand characteristics. • Utilize multiple sources of information to maximize reliability and validity. • Ask questions in a way that elicits useful information. • Obtain information in as value-free a manner as possible. • Obtain assessment information that may help clarify the client's situation.
2. Develop diagnosis	• What makes this client similar to other clients? • What makes this client unique? • What parts of the client's presentation are credible? • What parts need further checking? • What has the client not told you? • Evaluate the client on case conceptualization factors: 1. Learning and modeling 2. Life events 3. Genetics and temperament 4. Physiological factors affecting psychological factors 5. Drugs affecting physiological factors 6. Sociocultural factors	2. Develop hypotheses	• Do client's symptoms or complaints match diagnostic criteria? • What about symptoms that overlap with other diagnoses? • What are the base rates? • What is the comorbidity rate? • What additional information do you need? • What is the evidential basis for your conclusions on the conceptualization factors?

3. Develop treatment plan	• What priorities make sense for this client? • What is apt to work for this client given the resources? • What will the client agree to? • What are you and the client comfortable trying? • How can you monitor progress?	2. Develop hypotheses	• What is known to work with clients similar to this one? • What is known to "not" work with similar clients? • If no "standard of care," what methods can be said to have the best chance of being effective? • Develop plan for data collection as part of ongoing treatment. • Ensure clear operational definitions of goal attainment, behaviors, and results. • Behavioral specificity is preferred over vague statements.
4. Implement treatment plan	• How is client reacting to treatment? • Is client complying with treatment assignments? • Is therapist adhering to the treatment plan? • Are therapist and client maintaining a satisfactory alliance? • Is client making progress toward goals?	1. Observe. 3. Test hypotheses. 4. Observe results. 5. Revise hypotheses. 6. Test new hypotheses.	• Is client attending sessions? • Is client showing change? • Is change consistent with what was expected? • Has new information surfaced that would change the hypotheses? • Are there trends that might indicate that a change in treatment plan is needed?
5. Verify results	• Did client meet goals? • Do other clients meet goals?	7. Observe results. 8. Revise hypotheses. 9. Test new hypotheses. 10. Disseminate results.	• How can you perform an unbiased assessment of your own work? • Can you demonstrate a causal relationship between treatment and change? • How can you modify your practice based on results? • Would these results be of interest to others?

or read. They ask for, and evaluate, the evidence based on their understanding of the principles and methods of science. This is especially necessary in the age of the Internet and the World Wide Web. Today, information can be disseminated at a fantastic pace. It is not all good information and cannot be relied on by a professional until it is vetted and proved to be reliable.

To be a skeptic is not the same as being a pugilist. Although some scientists on opposite sides of a theoretical controversy go at one another with the ferocity of heavyweight boxers fighting for the world championship, such ferocity is not necessary. Skepticism demands that we examine the evidence; when we find it weak or otherwise unpersuasive we can declare our distrust of the evidence, usually without distrusting or disrespecting those who reported it. In fact, Shadish, Cook, and Campbell (2002) go so far as to state that "the ratio of trust to skepticism in any given study is more like 99% trust to 1% skepticism than the opposite" (p. 29). They continue to assert that "thoroughgoing skepticism" is impossible in science. We assert that the issue revolves around who should be trusted, what should be trusted, and in what circumstance.

Huff (1954) used actual examples from the media to demonstrate many tricks that will lead a reader to draw a conclusion that is not supported by the data. This is the book that the student believed made him a cynic, but it should have turned him into a skeptic. At the end of the book, Huff provides five questions that the alert and skeptical reader can use to determine whether a statistic, a study full of statistics, or an author can be trusted. Huff's questions are discussed here.

"Who Says So?" (Huff, 1954, p. 123) The nonspecialist in a field has no idea regarding who has a track record of doing excellent work; so they often look for an institutional or professional affiliation for guidance. Being associated with a famous institution affords an author an "OK name," whether or not it is deserved. Several years ago, a physician wrote a book on sex, which became a bestseller. The good doctor claimed to be a psychiatrist and to have received his medical education from Harvard University. Neither claim proved to be true. In general, watch out for the researcher or institution with a vested interest in proving a point. Much of the evidence in favor of psychopharmacological remedies originates from the companies who produce the medications. This concerns us.

"How Does He (She) Know?" (Huff, 1954, p. 125) Ask where the data came from, how large the sample size was, and how it was obtained. Very large and very small samples can be misleading, and a biased sample should always be considered misleading until proved otherwise.

"What's Missing?" (Huff, 1954, p. 127) Pearson's challenge demands that evidence be provided with an assessment of its own reliability. For statistics, that means confidence intervals, standard errors, or effect sizes. It also means

defining one's terms. If an "average" is reported, ask which kind. Means, medians, and modes are impacted by different factors, and a cheat will report the one that best states his or her case. In examining research reports in general, ask how well the design of the study matches with the principles covered in this book.

"Did Somebody Change the Subject?" (Huff, 1954, p. 131) Suppose a researcher surveys clients about their satisfaction with therapy and rapport with their therapist, finds a relationship between the two variables, and reports greater rapport leads to better treatment outcomes. Notice the change from "satisfaction" to "outcome." The two words are by no means synonymous. This is a case of switching the subject. The clinical literature is replete with examples. Other forms of changing the subject include using far different definitions of terms from what the audience expects and either not providing that information or burying it so the reader tends to skip over it. Gernsbacher, Dawson, and Goldsmith (2005) documented one such switch in the case of the so-called *autism epidemic*. Many people believe that autism is increasing at a tremendous rate. It has increased, but only because the definition of autism has been broadened so that many people who would not have received the diagnosis in years past now qualify.

"Does It Make Sense?" Huff (1954, p. 137) reminds us that sometimes a "finding" makes no sense and the explanation is there is no intrinsic reason for it to do so. As an example, he cited a physician's statistics on the number of prostate cancer cases expected in this country each year. It came out to 1.1 prostates per man, a spurious figure. A few years ago, a method was devised that supposedly allowed autistic children to communicate with parents, teachers, and therapists (McBurney, 1996). "Facilitated communication" involved having a specially trained teacher hold the autistic child's hand, and the child held a marking device over a board on which the letters of the alphabet were printed. Wonderful results were reported. Children who found it impossible to communicate even simple requests were creating complex messages even beyond what would be expected of other children their age. Was it too good to be true? It was. Was it sensible? It was not. Skepticism may have seemed cruel in denying the communicative abilities of these children, but even crueler was the discovery that the communication unconsciously sprang from the facilitator, not the child.

The most difficult aspect of being a skeptic is being a fair skeptic. If a study supports what we already believe, we are much less likely to subject it to the same scrutiny as a study whose results are contrary to our preferences. Corrigan (2001) illustrated this in *The Behavior Therapist*, the newsletter of the Association for the Advancement of Behavior Therapy (AABT). There are some psychotherapies for which behavior therapists have a natural affinity and other therapies that they view with some suspicion, a case in point being EMDR. Corrigan (2001) found after a fairly simple and brief literature search that there appears to be as much empirical support for EMDR as there is for the preferred therapies. Corrigan did not attempt to compare results nor did he examine the quality of the studies. His goal was simply to point out that without going to that effort there is no more *a priori* reason to reject EMDR than there is to accept the others. We can only add that the best strategy is to redouble one's efforts in double-checking results when the results fit one's previously established preferences. In fact, Herbert and his collegues (2000) did exactly this by more carefully examining the scientific and nonscientific ingredients of EMDR, in which they succinctly concluded that the active ingredients were exposure and cognitive restructuring (i.e., already known as efficacious, effective, and efficient), whereas the inactive ingredient was eye movement, which ironically is what Shapiro heralded—until disassembly research was allowed—as the new and central change agent within her protocol. Additionally, because she also made such extreme claims about the rapidity, permanence, and generality of its effects, such claims, according to the philosopher Hume (1748/1977), would require "extraordinary" evidence, something which Shapiro has not provided. Instead, aggressive marketing and promotion of EMDR, without an appropriate level of methodological rigor in its validation, has been its mainstay.

Science is conservative due to its need for skepticism and evidence. There are always new ideas and techniques that fall outside the domain of science. Some fall into what Shermer (2001) calls the "borderlands of science," not quite scientific, although potentially can be so. Often, however, the latest fads fail to have much of a lasting impact on science and practice just as 10-year-old clothing fashions have little influence on the current mode of dress. It takes time to weed out what is of lasting value when it comes to the cutting edge. This means there are potentially helpful interventions that the local clinical scientist does not employ, and this does represent a cost of ethical practice. There is, however, an even greater cost to clients, payers, the profession, and society at large if skepticism and the rigorous inspection of evidence are abandoned and every fad is adopted on the flimsiest of support (Dunnette, 1966). There are tremendous demands from clients and the market to give in to instant gratification, but that is not what a professional does. Be skeptical, ask questions, and generate answers.

References

Addis, M. E., & Krasnow, A. D. (2000). A national survey of practicing psychologists' attitudes toward psychotherapy treatment manuals. *Journal of Consulting and Clinical Psychology, 68*, 331–339.

Aiken, L. S., & West, S. G. (1990). Invalidity of true experiments: Self-report pretest biases. *Evaluation Review, 14*, 374–390.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

American Psychological Association Presidential Task Force on Evidence-based Practice (2006). Evidence-based practice in psychology. *American Psychologist, 61*(4), 271–285.

Bahrick, L. E., Parker, J. F., Fivush, R., & Levitt, M. (1998). The effects of stress on young children's memory for a natural disaster. *Journal of Experimental Psychology: Applied, 4*, 308–331.

Campbell, J. T. (1990). The role of theory in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 39–74). Palo Alto, CA: Consulting Psychologists Press.

Cook, T. D. (1991). Postpositivist criticisms, reform associations, and uncertainties about social research. In D. S. Anderson & B. J. Biddle (Eds.), *Knowledge for policy: Improving education through research* (pp. 43–59). London: The Falmer Press.

Corrigan, P. (2001). Getting ahead of the data: A threat to some behavior therapies. *The Behavior Therapist, 24*(9), 189–193.

Division 12 Task Force. (1995). Training in and dissemination of empirically validated psychological treatments: Report and recommendations. *The Clinical Psychologist, 48*, 3–23.

Doyle, A. C. (1890/1986). The sign of four. In A. C. Doyle (Ed.), *Sherlock Holmes: The complete novels and stories* (Vol. 1). New York: Bantom Books.

Dunnette, M. D. (1966). Fads, fashions, and folderol in psychology. *American Psychologist, 21*, 343–352.

Emmelkamp, P. M. G. (1982). *Phobic and obsessive-compulsive disorders: Theory, research, and practice*. New York: Plenum Press.

Farone, S. V., Tsuang, M. T., & Tsuang, D. W. (1999). *Genetics of mental disorders*. New York: Guilford Press.

Foa, E. B., Franklin, M. E., & Kazak, M. J. (1998). Psychosocial treatments for obsessive-compulsive disorder: Literature review. In R. P. Winson, M. M. Antony, S. Rachman, & M. A. Richter (Eds.), *Obsessive-compulsive disorder: Theory, research, and treatment* (pp. 258–276). New York: Guilford Press.

Gernsbacher, M. A., Dawson, M., & Goldsmith, H. H. (2005). Three reasons not to believe in an autism epidemic. *Current Directions in Psychological Science, 14*, 55–58.

Greenberg, P. E., Sisitsky, T., Kessler, R. C., Finkelstein, S. N., Berndt, E. R., Davidson, J. R. T., et al. (1999). The economic burden of anxiety disorders in the 1990s. *Journal of Clinical Psychiatry, 60*, 427–435.

Guralnik, D. B. et al. (1978). *Webster's new world dictionary of the American language* (2nd college ed.). William Collins + World Publishing Company.

Herbert, J. D., Lilienfeld, S. O., Lohr, J. M., Montgomery, R. W., O'Donohue, W. T., Rosen, G. M., et al. (2000). Science and pseudoscience in the development of eye movement desensitization and reprocessing: Implications for clinical psychology. *Clinical Psychology Review, 20*, 945–971.

Huff, D. (1954). *How to lie with statistics*. New York: Norton.

Hume, D. (1748/1977). *An inquiry concerning human understanding*. Indianapolis, IN: Hackett. (Original work published in 1748)

Kaslow, N. (2004). Competencies in professional psychology. *American Psychologist, 59*(8), 774–781.

Kessler, R. (1997). The effects of stressful life events on depression. *Annual Review of Psychology, 48*, 191–214.

Kimble, G. A. (1989). Psychology from the standpoint of a generalist. *American Psychologist, 44*, 491–499.

Kluger, A. N., & Tikochinsky, J. (2001). The error of accepting the "theoretical" null hypothesis: The rise, fall, and resurrection of common sense hypotheses in psychology. *Psychological Bulletin, 127*, 408–423.

Lebow, J. (2006). *Research for the psychotherapist: From science to practice.* New York: Routledge.

Lewinsohn, P. M. (1974). A behavioral approach to depression. In R. M. Friedman & M. M. Katz (Eds.), *The psychology of depression: Contemporary theory and research* (pp. 157–185). New York: Wiley.

Mahrer, A. R. (1988). Discovery oriented psychotherapy research: Rationale, aims, and methods. *American Psychologist, 43*, 694–702.

McBurney, D. H. (1996). *How to think like a psychologist: Critical thinking in psychology.* Upper Saddle River, NJ: Prentice-Hall.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests with confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests*? (pp. 393–425). Mahwah, NJ: Lawrence Erlbaum Associates.

Meyer, V. (1966). Modification of expectations in cases with obsessional rituals. *Behavior Research and Therapy, 4*, 273–280.

National Institute for Mental Health. (2009). *The numbers count: Mental disorders in America*. Retrieved December 16, 2009, from http://www.nimh.nih.gov/health/publications/the-numbers-count-mental-disorders-in-america/index.shtml

Nehls, H., Aversa, T., & Opperman, H. (2004). *Birds of the Willamette Valley region*. Olympia, WA: R. W. Morse Company.

Newman, F. L., & Tejada, M. J. (1996). The need for research that is designed to support decisions in the delivery of mental health services. *American Psychologist, 51*, 1040–1049.

Park, R. (2000). *Voodoo science: The road from foolishness to fraud*. New York: Oxford University Press.

Petroski, H. (1985). *To engineer is human: The role of failure in successful design*. New York: St. Martin's Press.

Popper, K. (1959). *The logic of scientific discovery*. New York: Basic Books.

Prochaska, J. O. (1999). How do people change and how can we change to help many more people change? In M. A. Hubble, B. L. Duncan, & S. D. Miller (Eds.), *The heart and soul of change: What works in therapy* (pp. 227–255). Washington, DC: American Psychological Association.

Rosen, G. M. (1999). Treatment fidelity and research on Eye Movement Desensitization and Reprocessing (EMDR). *Journal of Anxiety Disorders, 13*, 173–184.

Rosqvist, J. (2005). *Exposure treatments for anxiety disorders: A practitioner's guide to concepts, methods, and evidence-based practice.* New York: Routledge.

Rosqvist, J., Thomas, J. C., Egan, D., Willis, B. S., & Haney, B. J. (2002). Home-based cognitive-behavioral treatment successfully treats severe, chronic, and refractory obsessive-compulsive disorder: A single case analysis. *Clinical Case Studies, 1*, 95–121.

Salomoni, G., Grassi, M., Mosini, P., Riva, P., Cavedini, P., & Bellodi, L. (2009). Artificial neural network model for the prediction of obsessive-compulsive disorder treatment response. *Journal of Clinical Psychopharmacology, 29*, 343–349.

Seligman, M. E. P. (1998a). Foreword. In P. E. Nathan & J. M. Gorman (Eds.), *A guide to treatments that work* (pp. v–xiv). New York: Oxford University Press.

Seligman, M. E. P. (1998b). Afterword. In P. E. Nathan & J. M. Gorman (Eds.), *A guide to treatments that work* (pp. 568–572). New York: Oxford University Press.

Seligman, M. E. P., Reivich, K., Jaycox, L., & Gillham, J. (1995). *The optimistic child*. Boston: Houghton Mifflin.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Mahwah, NJ: Lawrence Erlbaum Associates.

Sexton, T. L., Whiston, S. C., Bleuer, J. C., & Walz, G. R. (1997). *Integrating outcome research into counseling practice and training*. Alexandria, VA: American Counseling Association.

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Shapiro, F. (1999). Eye Movement Desensitization and Reprocessing (EMDR) and the anxiety disorders: Clinical and research implications of an integrated psychotherapy treatment. *Journal of Anxiety Disorders, 13*, 35–67.

Shermer, M. (2001). *The borderlands of science*. New York: Oxford University Press.

Spring, B. (2007). Evidence-based practice in clinical psychology: What it is, why it matters, what you need to know. *Journal of Clinical Psychology, 63*(7), 611–631.

Stefan, S. (2001). *Unequal rights: Discrimination against people with mental disabilities and the Americans with Disabilities Act*. Washington, DC: American Psychological Association.

Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.

Stricker, G., & Trierweiler, S. J. (1995). The local clinical scientist: A bridge between science and practice. *American Psychologist, 50*, 995–1002.

van Oppen, P., & Arntz, A. (1994). Cognitive therapy for obsessive-compulsive disorder. *Behaviour Therapy and Research, 32,* 273–280.

van Oppen, P., & Emmelkamp, P. M. G. (2000). Issues in cognitive treatment of obsessive-compulsive disorder. In W. K. Goodman, M. V. Rudorfer, & J. D. Maser (Eds.), *Obsessive-compulsive disorder: Contemporary issues in treatment* (pp. 117–132). Mahwah, NJ: Lawrence Earlbaum.

Waller, G., (2009). Evidence-based treatment and therapist drift. *Behaviour Research and Therapy, 47*, 119–127.

Westen, D. (1998). The scientific legacy of Sigmund Freud: Toward a psychodynamically informed psychological science. *Psychological Bulletin, 124,* 333–371.

Wilson, K. A., & Chambless, D. L. (1999). Inflated perceptions of responsibility and obsessive-compulsive symptoms. *Behaviour Therapy and Research, 37*, 325–335.

Note

1. Indeed, according to Nehls, Aversa, and Opperman (2004), it was probably a crow. However, those authors note that in this part of the country all ravens and all crows are black.

2
Measurement Theory in Research

JAY C. THOMAS and LISA R. CHRISTIANSEN

Contents

Imagine you are a World Records representative investigating two competing claims of "the world's longest mustache." One contender states his mustache extends to the bottom of his ribcage, whereas the other states his mustache extends to his hipbones. Might you be inclined to conclude the second contender's mustache is longer? What if the first gentleman was quite a bit taller than the second? What if the second gentleman had a much shorter torso than the first? Without a measuring device, how confident can you be in your conclusion? In any research question, accurate measurement is an absolute necessity. In the field of psychology, physical measurement of a trait such as height, weight, or mustache length is rarely used; instead, psychology deals with abstract concepts that are difficult to directly observe and sometimes even difficult to define. Developing and utilizing the best methods for measuring such traits is therefore an essential component of psychological research.

Reliability

In a baseball game, the home plate umpire calls balls and strikes. The rules of baseball define the strike zone, the rectangle over the plate within which a pitch is considered a strike and outside which the pitch is a ball. Television broadcasters can superimpose an outline of the strike zone over the plate, so viewers can see whether a pitch is a ball or strike, but the umpire does not have the outline. Consequently, umpires may develop tendencies such that one may tend to treat lower pitches or pitches that are slightly outside (away from the batter), or similar errors, as strikes. Although all umpires would ideally be perfect in their calls of balls and strikes, players have to settle for an umpire being consistent so that once a slightly low or outside pitch is called a strike, any pitch that is low or outside is a strike. This example illustrates the concepts of validity and reliability. When a pitch is accurately termed a ball or strike exactly in accord with the definition of the strike zone, the measuring instrument, the umpire, is making valid calls. Consistently calling pitches in a given location as balls or strikes, regardless of where the rules state the strike zone ought to be, means the umpire is reliable. A reliable umpire may not be validly calling strikes and balls, but a valid umpire is reliable.

Measuring psychological attributes is much like calling balls and strikes. Even if we can exactly define what we are measuring, there typically is not something like the broadcaster's rectangle that can help us verify the measurement. The umpire has an advantage on those of us measuring psychological factors; the ball did pass through space at some point. Instead, our instruments are often tests, surveys, ratings, checklists, and so on from which we infer that we are measuring what we think we are measuring. For example, we cannot often directly observe an individual's emotional state. Even if we notice someone crying, we cannot necessarily assume that they are sad. People have been known to cry especially when happy or even when they are angry. Instead, we rely on their truthfulness of self-report when we ask them, "How are you feeling?" In essence, much of our psychological measurement techniques are like asking someone "How are you feeling?" in a lot of different ways. From the consistency and content of their responses, we hope to infer that we are indeed measuring what we think we are measuring, that is, their emotional state.

So, reliability deals with consistency of measurement. We can measure reliability in several ways; we present the four most common methods: (1) test–retest, (2) alternate forms, (3) internal consistency, and (4) interrater reliability. Things can rapidly get complicated, so to start, we will present these different types of reliability in as nontechnical language as possible, and then present the theory of reliability more rigorously once the basic ideas are established.

To begin, we switch from baseball to another analogy, measuring a person's height, to allow for the simplest possible introduction. A person's height is a straightforward physical measurement. In many homes, the height of family members is inscribed on a door jamb showing how the child or children have grown over the years. Height is based on the distance from the floor to the top of one's head. It would seem to be an exact number, but its reliability would only be limited by the precision of measurement. That precision is determined by the gradations on the ruler as well as how near it is held to the horizontal when placed on the person's head and against the wall. The angle of the ruler introduces errors that can be slight or can range up to a couple of inches. Even the perspective of the person recording the height can introduce errors. One of us (Jay C. Thomas) was a subject in a longitudinal study examining changes in health in "older" adults. Each year, he would arrive at the laboratory, fill out questionnaires, and have his weight and height recorded. Height was measured by a tape measure mounted on the wall. For a known mounting height, presumably the subject could stand under the tape measure, have the tape pulled down to the top of the head, and the research assistant (RA) could read the height. One year, the reading was three-fourth of an inch shorter than he had ever been since reaching adulthood. This was disturbing enough, but after he pointed out the discrepancy, the RA commented "Well, people your age do start getting shorter." Actually, the RA was rather short, and the angle between her eyes and the tape measure 3 feet above her head resulted in a biased measurement. Once it was discovered, several people had to be remeasured. An error that consistently results in measures that are too high or too low does not affect reliability, but it can affect validity just as in the case of baseball umpires.

You probably have a good idea of how tall you are and have no lack of confidence in reporting it on forms such as those used for applying for driver's licenses. Whatever your actual height is, it can be considered a *true score.* The heights we measure against the wall are *observed scores* and are a composite of true scores *plus or minus error.* Assuming for simplicity that we can measure without bias as in the example of Jay Thomas's height, the errors are assumed to have a mean of zero and to be normally distributed above and below that mean. Assumption of normality of errors boils down to expecting that, over a great many measurements, there will be more small errors than large errors and that positive and negative errors balance out. We also have to assume that the errors are independent of the true score. This is not the case with some types of measures, such as mechanical scales that are rated to be accurate within a percentage of the actual weight. A bathroom scale may be rated for accuracy within ±2%, meaning the observed weight of a person weighing 100 pounds could be between 98 and 102 pounds, whereas for a person weighing 200 pounds, the interval is from 196 to 204 pounds. Given all these assumptions, we have the equation

$$X = t + e$$

where X is the observed score, t is the true score, and e is the error. Because the e's are just as likely to be positive as negative, we do not have to include "plus or minus" in the equation, but you should remember that the error can result in a score that is higher or lower than the true score.

There is a good deal of variability across people in height. You might recall that we can measure variability through a variance. The sample variance[1] is defined as

$$s_x^2 = \left(\frac{\sum_i^n (x_i - \bar{x})^2}{n-1} \right)$$

Recall also that the cousin of the variance, the standard deviation, s_x, is simply the square root of the variance. Because true scores and errors are uncorrelated, the variance of the observed scores is the sum of the variance of true scores (s_t^2) and the variance due to error, s_e^2 (we say that variances are additive, which is why they are preferred in many computations over the standard deviation; the strength of the latter is that it is in the same measurement units we began with and so is more useful for descriptive purposes).

At this point, we have s_x^2 that represents 100% of the observed variance and s_t^2 and s_e^2, each representing proportions of that total variance. You might also recall from statistics that the squared correlation (r^2, the coefficient of determination) represents the amount of variance the two variables share; thus, it is often said for two variables, X and Y, r^2 is the proportion of variance in Y explained by X. If we could calculate s_t^2, we could determine directly what proportion of the total observed variance is accounted for by true scores. We cannot make that calculation, but a little algebra leads to the conclusion that the squared correlation between observed scores and true scores represents the proportion of observed variance due to true score variance. This squared correlation is defined as *reliability*, denoted r_{xx}, or more directly, reliability is the squared correlation of observed scores with true scores.[2]

Test–Retest Reliability

If we repeat the measurements of the heights of a large number of people twice, we should obtain similar results on each occasion, a type of reliability known as test–retest reliability. Yet, you have probably heard that a person's height varies somewhat throughout the day being slightly taller in the morning, a little shorter in the evening. Thus, consistency depends on when we take the measurement, and our test–retest reliability of height may be limited if people are tested at different times during the day. There are also long-term effects. Young people grow, elderly people may shrink. Developmental trends can, depending on who is studied, affect the test–retest reliability even though the trends are predictable.

Test–retest reliability is typically evaluated using a correlation. We simply take measurements from the first occasion and correlate them with the second occasion. In theory, correlations can range from –1.0 to +1.0; a result close to +1.0 or –1.0 indicates a very high relationship, and a result close to 0 indicates very little relationship.[3] A coefficient of correlation is a small value that packs a lot of information. One type of information it includes is how well the relative ranking of measurements held up from time 1 to time 2. Suppose we measure the height of children at 3 years of age and then repeat the measurements on the same children 2 years later. Children can grow a lot between three and five. Further suppose that the children tended to grow about the same amount so that the rank ordering of heights remained about the same. We would obtain a large correlation, leading us to conclude that the measurement system had high reliability even though there was a large change in the measurements. So, confounding of developmental trends into the study can substantially reduce the credibility of the results. The situation is further complicated by some psychological variables, such as depression, occurring in phases, rising for a period of weeks, then dropping down to more typical levels for awhile, and so on.

The lesson here is that test–retest reliabilities should be considered confounded with stability of the characteristic being measured. Some authors suggest using the term *stability coefficient* in place of test–retest reliability (Ghiselli, Campbell, & Zedeck, 1981). The confounding with stability limits the use of test–retest reliability in practice and research. Since test–retest reliabilities are evaluated using correlations, an artificially high value may be obtained when all or most of the people being studied change approximately the same amount and maintain much the same rank ordering.

When test–retest reliabilities are cited in research studies, they should be from studies that covered approximately the same period of time. The degree of approximation depends on what is known about the stability of the trait or attribute being measured. For studies of depression, test–retest duration should be within a few weeks of the duration of the study. For studies of intelligence, a more stable trait, the test–retest duration, could differ from the study duration by some months, or even years for a long-term study.

A second limiting factor on use of test–retest reliability is rooted in tactics; it is often difficult or expensive to get study participants to return for a separate testing. Since test–retest reliability necessarily has to be evaluated on people who are not part of an experimental test of something intended to change them, separate studies on similar people are needed to establish the test–retest reliability of a measure.

A third limiting factor is that the people being tested may learn the contents and remember their previous answers, thus confounding reliability with memory effects. A related problem transpires when the person is somehow changed by the test; answering questions about personality or mood could cause the person to begin introspecting about those matters, and this could result in some internal change (Ghiselli et al., 1981). The test in this case could be validly measuring the trait, but because taking the test contributed to the observed change, the results could not be interpreted as reliable. Longwell and Truax (2005) conducted a study involving administering the

Beck Depression Inventory II (BDI-II; Beck et al., 1996) once each week to a group of undergraduate students over a two-month period. Another group took the BDI-II bimonthly, and a third group took it monthly. All participants also took another depression measure at the beginning and end of the study and a third measure just at the end. There were no differences between the groups on the pretests or posttests for the alternative measures, but the weekly BDI-II group demonstrated a consistent decrease in BDI-II scores over time. Thus, something about completing the inventory repeatedly on a weekly basis was associated with changing scores.

The use of test–retest reliability should be carefully considered whenever people are being studied over a time period in which the characteristic can reasonably be expected to change. Mood fluctuates over days or weeks, arousal over minutes or hours, whereas personality and intelligence may remain stable for a few years. Some researchers attempt to avoid some of these problems by administering the instrument repeatedly over time, say once a month for several months. This is thought to allow for the recognition of at least phase-related trends and, possibly, developmental trends. Obviously in such studies, reliability is no longer expressed as a simple correlation, and the researchers must examine the data carefully to determine whether results are credible over the duration they desire to study. The researchers must also take care to determine that the Longwell and Truax (2005) effect is not occurring. Finally, practical problems in obtaining this type of data dissuade many researchers from even trying.

Alternate Forms

Problems with test–retest reliability were recognized relatively early in the history of psychometrics. One could eliminate the stability problems as well as some of the tactical problems associated with test–retest by shortening the interval between test administrations to no time at all; participants finish the test and start over on it again right away. This has the obvious limitation of participants remembering their earlier answers and simply repeating them or benefiting from practice or losing interest, so the test developers had to prepare alternate versions of the test covering the same content but with different items. If the test were of arithmetic skills, the two forms might look something as shown in Table 2.1.

Table 2.1 Example Items on Two Forms of an Arithmetic Test

Form A	Form B
24 × 53 =	76 × 43 =
87 × 32 =	13 × 65 =
93/21 =	77/43 =
65/57 =	98/53 =

If scores on both forms are highly correlated, then the tests have alternate form reliability. This works pretty well for tests such as this, although what if one form turns out to have somewhat more difficult items? Ideally, test forms should be interchangeable, so it is desirable that the forms be parallel, meaning they have the same mean and standard deviation, correlate to a very high degree, and correlate with other variables to the same extent.[4] This results in what are called *parallel tests* (Ghiselli et al., 1981). Actually developing tests that are so closely comparable is difficult and requires statistical and methodological expertise far beyond our scope here. For more details, see the book edited by Dorans, Pommerich, and Holland (2007). For now, we will proceed making the assumption that such tests can be developed.

Much of the reliability theory was developed in educational testing and in the development of cognitive ability tests. It is fairly easy to come up with, say, two 20-item forms of problems involving two-digit multiplication and division, such as those shown in Table 2.1. Other types of instruments require much more creativity. For some commercial personality tests, the authors begin with hundreds or even thousands of potential items, which are culled down first for content, and then more items are eliminated based on the analysis of data from trial administrations. It is not easy to write so many items and certainly not easy to write several distinct items on the same construct so that they consistently have the same content. For example, a scale for the trait of extroversion might have items asking about socializing with groups of friends. Here is a good exercise; try writing two sets of 10 different items on this topic so that you are assured that each set of 10 has the same content and likely the same pattern of answers. If your experience is like ours, the first few items are easy to generate, but it gets harder and harder to complete the task. It often seems as if the first few items exemplify the construct much better, or at least differently, than the last few items when creativity is stretched. If you were able to write two comparable sets of items, then you may have a future in the testing industry.

Writing alternate items hoped for parallel forms is hard enough for personality inventories. It is even harder for behavioral measures and inventories of symptoms. How many ways can you ask about difficulty sleeping at night, use of alcohol, eating habits, or sexual difficulties such as those found on a depression inventory such as the BDI-II? This limitation in the ability to create comparable items across forms is the first limitation in the use of alternate forms reliability.

The second limitation was alluded to the above one. Developing anything close to parallel tests for the population to be studied requires very sophisticated sampling, methodological, and statistical expertise, as well as very large resources to pay for these studies. Consequently, today we usually see alternate forms and parallel tests only in widely used published psychological measures and educational tests.

Alternate forms' reliability must be assessed whenever more than one form of an instrument is employed in a study. Studies examining developmental progress often have to use different forms of the same measure because the same trait may be expressed differently throughout the lifespan. This problem is known as *scale alignment*. Aligning scores on such measures is even more complicated than developing parallel tests. The book by Dorans et al. (2007) introduces these methods, and a high level of statistical expertise is required to use the information.

Internal Consistency Reliability

Evaluating alternative forms of reliability had a number of practical difficulties that were noticed very early in the development of psychometrics: problems writing enough similar items, getting people to sit through two administrations, practice effects, and others. A simple solution was to break the test in half and treat the two halves as alternate forms. Comparing the first half of the items with the second half did not control for experience, but that could be accomplished by comparing the odd-numbered items with the even-numbered items. An odd–even split is just one of many possible splits, but it is generally the easiest to work with—a major consideration in precomputer days. Thus, split-half analyses have most often been odd–even splits. As usual, there were problems associated with this method. Right away, you can see that the two halves must be equated for content and difficulty, so that technical problem is not escaped, just diminished by using half the items needed for two full forms. However, the biggest problem with split-half reliability comes from the use of half as many items in the analysis. You probably already know that in sampling, more is better (all else being equal). In this case, the sample of items is half as large in each "form" as it would be comparing two complete forms. More observations or items lead to greater reliability, so split-half reliability analyses provide an underestimate of the actual reliability. If we assume that the two halves of the test meet the standard of parallel forms, the actual reliability can be estimated through the use of the Spearman–Brown Prophecy formula, which allows inquiries of how lengthening or shortening a test can influence reliability. In the formula, r_{kk} is the expected reliability of the relengthened test, r_{xx} is the current reliability of the halves (remember, we assume they are parallel and, thus, have equal reliability), and k is the factor by which the test is lengthened or shortened. If the test is doubled (i.e., the two halves combined into a single test), $k = 2$.

$$r_{kk} = \frac{kr_{xx}}{1+(k-1)r_{xx}}$$

The Spearman–Brown formula is based on the assumption that any additional items are similar to the existing items in terms of difficulty and that they represent the same content domain. In addition to correcting split-half estimates, the formula is useful for examining the effects of proposed changes. For example, suppose a current 10-item instrument is not sufficiently reliable for some purpose. The researcher can determine how many more items are needed to achieve the desired reliability. Alternatively, an instrument is too long to include in a research study, and the researcher is interested in what would happen to the reliability if the number of items were cut by a third (k = 1/3). Because of these uses, the Spearman–Brown formula still exists despite the split-half method being rarely used today.

Development of split-half reliability focused on the internal consistency of a test, that is, how well the items relate to one another. Several reliability indexes were developed under different sets of assumptions; eventually Cronbach developed the most widely used measure of internal consistency reliability, Cronbach's α,[5] or coefficient α, as Cronbach termed it. Split-half and some other earlier forms of internal consistency were based on what is known as *strong true score theory.* In this theory, the true score is thought to actually exist, just as a person's height exists as some figure. As statistical theory was applied to psychometrics, there was a shift to a weaker form of true score, one that is defined theoretically and is not expected to have an actual value. In the multiplication test described above, we used four problems in each form. However, there are 9,801 possible sets of two two-digit numbers to multiply (if we include 00, 01, 02, etc., as two-digit numbers; otherwise there are only 8,100 sets). We had to sample from the possible items. We now have a situation in which we are sampling both people and items. The sampling of the items leads to *domain sampling theory.* In domain sampling theory, the true score is the average (mean) score a person would obtain if the person could take all possible versions of the test made up of items from the domain (under the same conditions). Domain sampling theory allows for some additional statistical theory in the development of reliability theory. This domain sampling theory imposes some limits on the use and interpretation of Cronbach's α.

The formula for Cronbach's α looks more complicated than it is. The formula is as follows:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum \sigma_k^2}{\sigma_{\text{total}}^2}\right]$$

where k is the number of items, σ_k^2 is the variance of an item across all test takers, and σ_{total}^2 is the variance of total scores.

It is important to understand this formula to understand the meaning and limitation of Cronbach's α, so bear with us as we enter into a little bit of statistics. For understanding purposes, you do not need to worry about the $k/(k-1)$, and we will ignore aspects of other formulas that do not contribute to the understanding level. You might recall the definitional formula for a sample variance as follows:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

Also recall the formula for a standard, Pearson Product–Moment correlation as follows:

$$r_{xy} = \sqrt{\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

The numerator of the correlation (inside the square root sign) is known as the covariance. If $x = y$, then the covariance would equal the numerator of the variance and, of course, $r = 1$.

Now, imagine we gave the four-item multiplication test in Table 2.1 to a large number of fifth graders. Some children would get all four problems correct, some would be correct on three, others on two or one, and some would not get any correct. Across all of the children, we could compare each pair of items and calculate the variances and covariances. These are usually displayed in a variance–covariance matrix as shown in Table 2.2. If you add up all the elements in the matrix, the sum equals the total variance on the test (using all the total scores of all the children). There are twice as many covariances as there are variances in the matrix. The covariances represent the amount of the total variance that is shared across items. The total of the variances represents variances associated with individual items. Therefore, in the formula for Cronbach's α, the fraction $\left(\left(\sum\sigma_k^2\right)/\sigma_{\text{total}}^2\right)$ represents the proportion of variance due to the individual items. Subtract it from one, and you get the proportion of variance that is shared among the items. When you get down to it, that is what Cronbach's α represents. Because of the way it is defined and derived, Cronbach's α represents a lower bound on internal consistency reliability; the actual reliability could be higher (Thompson, 2003).

Because Cronbach's α is so widely used in research, it is important to address some myths and misunderstandings about it. Earlier we said it

Table 2.2 A Variance-Covariance Matrix for a Four-Item Test

	Item 1	**Item 2**	**Item 3**	**Item 4**
Item 1	s_1^2	cov_{12}	cov_{13}	cov_{14}
Item 2	cov_{12}	s_2^2	cov_{23}	cov_{24}
Item 3	cov_{13}	cov_{23}	s_3^2	cov_{34}
Item 4	cov_{14}	cov_{24}	cov_{34}	s_4^2

represents the variance shared among items; that is, the extent to which they are intercorrelated. However, α can increase as more items are added to a scale (Streiner, 2003a). Researchers and practitioners need to be wary of high reliability for an instrument with a very large number of items because the reliability can be artificially increased by so many items. Streiner (2003a) also points out that reliability can also be high due to redundancy between items. If you ask the same question over and over, you may get high reliability, but not much useful information. Therefore, contrary to common opinion, bigger is not always better.

We often see or hear references to a particular test or other instrument as being "highly reliable." Unfortunately, reliability is as much or more a product of the sample from which the data were collected as it is the test (Streiner, 2003a; see also the volume edited by Thompson, 2003). Referring back to the multiplication test, we might guess that the average fifth grader would get half of the problems correct. If the same test were given to 100 second graders, the average correct would be less than one. If we administered it to 100 psychometricians, the average correct would be close to four. The test would most likely obtain the highest reliability with the fifth graders; reliability values would be low for the second graders and the psychometricians. Why? Because Cronbach's α depends on variances and covariances. If very few people get the items correct or few get them wrong, there will not be much variance to share and α will be low. It will be greatest when the average correct score is about 50%. For personality inventories, attitude questionnaires, and other measures not having "right" and "wrong" answers, we shift from "correct" to "endorsed" and get the same effect. If the instrument has items rated on a scale such as a Likert-style scale, the effect is the same if people tend to pile up on one or two rating points and there is little variance.

Cronbach's α is a form of statistics known as an intraclass correlation. Being a type of correlation, it can theoretically range from zero to one or even take on negative values, but in practice, this does not happen (Streiner, 2003a). Occasionally, an α can be calculated to be less than zero. This is due to some artifact in scale construction (Streiner, 2003a) and indicates that the test developer needs to revise the measure before gathering more data.

Another caution about internal consistency reliability concerns the type of measure. Streiner (2003b) points out that some instruments are scales while others are indexes. In a scale, it is assumed that responses to the items are caused by some underlying trait. A fifth grader with high numerical aptitude would score well on the multiplication test, whereas a fifth grader with very little numerical aptitude would not score well. The "cause" of the score is the aptitude for working with numbers. Domain sampling theory applies to scales. In an index, the items are chosen because together they define the trait. Personality disorders would be measured through an index. Antisocial personality does not cause the person to have narcissism, a lack of empathy,