John R. Rossiter

Measurement for the Social Sciences

The C-OAR-SE Method and Why It Must Replace Psychometrics

Springer

Measurement for the Social Sciences

John R. Rossiter

Measurement for the Social Sciences

The C-OAR-SE Method and Why It Must Replace Psychometrics

Springer

John R. Rossiter
Institute for Innovation in Business
and Social Research
University of Wollongong
Wollongong, NSW 2522, Australia
john_rossiter@uow.edu.au

ISBN 978-1-4419-7157-9 e-ISBN 978-1-4419-7158-6
DOI 10.1007/978-1-4419-7158-6
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010938285

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

I dedicate this book to the memories of James Lumsden, who taught me behavioral learning theory and psychometrics theory, and who taught me above all to be skeptical of anything in science that's too popular; my Father, who was my role model as an academic and my ideal, which I never lived up to, as a good man; and my Mother, who half-looked the other way while I learned and lived the qualitative truths of life. Also, to my two wonderful sons, B.J. and Stewart, who gave their Dad intellectual and moral support in this endeavor.

*Less direct dedications are to history's lone iconoclasts, especially Socrates, Voltaire, Nietzsche, Herman Hesse, Joyce, E.L. Doctorow, Bertrand Russell, Ayn Rand, Magritte, more recently the extraordinary Björk, and—a persistent thought as I write this book—to Jesus, who despaired about "Ye of little faith" (*Matthew *8: 26) but held to his own. Gadzooked I feel indeed, but I will rise again to rule via this book.*

Preface

C-OAR-SE rocks.

—Alan G. Sawyer (a realist)

[A]nd in many Cases one with amazement hears the Arguings, and is astonish'd at the Obstinacy of a worthy Man, who yields not to the Evidence of Reason, though laid before him as clear as Day-light.

—John Locke (1690), *An Essay Concerning Human Understanding*, Book II, Chapter XXXIII, para. 2.

Social scientist and would-be social scientists of all persuasions: this is the most important book you will ever read. I make this claim because social science *knowledge* is dependent—entirely—on valid *measurement*, and that is what we have lacked up until C-OAR-SE.

C-OAR-SE is my rational-realist measurement theory that I hope will drive out and replace psychometrics. It is an acronym for the essential aspects of the theory, which are Construct definition, Object representation, Attribute classification, Rater identification, Selection of item-type and answer scale, and Enumeration and scoring rule. Readers familiar with C-OAR-SE theory from the original article (Rossiter, *International Journal of Research in Marketing*, 2002a) should note that there are some very important improvements and updates of the theory in this book. The book newly includes a valuable chapter on *qualitative research* measurement. Also, the examples in the book come from *all* the social sciences, not just from marketing as in the journal article.

The German philosopher Arthur Schopenhauer said, "*reality* is a single will to existence and that *music* uniquely expresses that will" (Griffiths 2006, p. 710). When I want to get real—to get back in touch with the real mental and physical events that drive us—I listen to country music: Hank Williams, Patsy Kline, Dixie Chicks, or recent Australian-of-the-Year, Lee Kernaghan. However, I prefer to study and write while listening to wordless music: Brahms, Beethoven, or Schubert's *Impromptus*, or to post-Traditional and up-to-Coltrane jazz—music that is more ethereal, and in which the themes are deliberately difficult to detect so they don't interrupt you with reality. I am going to interrupt *you* with reality in this book. In my book, psychometrics is analogous to classical music or jazz, elegant but obscuring of reality. The latest derivation of psychometrics, structural equation modeling, I consider to be just a Bartokian or Ornette-Colemannish noisy distraction.

> Roamin' around, looking down on all I see.
> —Kings of Leon,
> "*Use Somebody*"

Those who know me as an academic know me as a fierce critic (attendees at Australian conferences call me "Dr. No"). My mentor in psychometrics—the late, great and delightfully radical Australian psychologist Jim Lumsden—said many times during research seminars when I was a psychology undergraduate (and see Lumsden 1978) that destructive criticism is sufficient. Jim used realistic anecdotes such as "It's sufficient if you let the driver know that his brake-light isn't working . . . you are not obliged to fix it." But I received my subsequent academic training in the U.S.A., where, if you're solely destructive you won't get published. That's why my C-OAR-SE theory ended up in a European journal instead of the leading U.S. journal, the *Journal of Marketing Research*, where the paper was first sent. I was in fact *very* constructive in that paper, but the conventional academics whose work I criticized couldn't see this for the red in their eyes! They still can't; I've had far more rejections of C-OAR-SE-based articles than on any other topic, and I've had more acceptances of journal articles than I've had years in my life. Never, never underestimate paradigm inertia if you're an innovator. And never get deflected by reviewers' discomfort if you're a critic. Just fight them, and if necessary send your article elsewhere—or write a book.

In this book, as I said, I attempt to destroy psychometrics, the dominant conventional approach to measurement in the social sciences, and to constructively substitute my C-OAR-SE theory of measurement. C-OAR-SE is a "rational realist" theory—totally opposite and opposed to the empirical and unreal "latent construct" approach that dominates social science measurement at present.

> You can't teach an old sheep new tricks.
> —Fabled Australian adage

The book is targeted in hope toward younger, more open-minded doctoral students, toward starting academics, and starting applied researchers in the social sciences. I have written off all my older colleagues who have built their careers on the conventional approach and who, quite obviously, are not about to adopt such a new and different approach. The only two people—both younger academics—who have *understood* C-OAR-SE are Lars Bergkvist and Tobias Langner, a Swede and a German, both free of U.S. academic indoctrination not surprisingly, and the only person who has *used* part of it (single-item theory, see Bergkvist and Rossiter, *Journal of Marketing Research*, 2007) is Lars. That the believers are but several is the reason for my allusion to the despair of one of the world's most famous prophets in my dedications ("Ye of little faith . . ."). I must add that Sara Dolnicar, my ex-European and now Aussie colleague at our research institute, has given constant support to my radical approach, though I don't think she fully agrees with it. But that's cool and I'm working on converting her.

My thanks to Marilyn Yatras (my administrative assistant) and Mary O'Sullivan (my lovely partner) for the typing, organization, and, in Mary's case, sanitary editing of my R-rated first draft of this book while leaving in my many tilts at political correctness. And to Wollongong's City Beach and Dancespace Dance Studio—my two escapes—and Litani's coffee shop, just around the corner from my home, for all those long-black coffees, *baklava*, and, I have to admit because I'm a realist and too many people know, cigarettes, that fuelled the writing. My smoking, incidentally, makes me a rationalist gambler—see Chapter 7. I am rewarding myself with a few of the excellent new *Bluetongue Aussie Pilsener* beers as I write this preface to usher in the New Year and, hopefully, a new era of rational-realist measurement.

It's proof-reading time now and the book is about to go to press. I add my thanks to Nick Philipson at Springer for his welcome encouragement throughout the entire project. I've now switched to *Fat Yak* ale and *Camels* (see p. 145).

Wollongong, NSW John R. Rossiter
October 15, 2010

Contents

About the Author

John R. Rossiter (B. Psych. Hons., W.A.; M.Sc. Marketing, U.C.L.A.; Ph.D. Communications, Penn.) is Research Professor in the Institute for Innovative Business and Social Research at the University of Wollongong, Australia, and is also on the staff as a Visiting Professor of Marketing in the Institute for Brand and Communication Research, Bergische Universität Wuppertal, Germany. For over 20 years, John Rossiter has been the most-cited Australian marketing author and has the distinction of being the only Australian listed in the inaugural Penguin Dictionary of Marketing. In parallel with his academic career, John Rossiter works as a marketing and advertising research consultant to private companies and government organizations.

Chapter 1
Rationale of C-OAR-SE

Statistics . . . lends itself only to the pinchbeck tarradiddle to which advertising is by nature prone.
—M.J. Moroney, Facts from Figures, 1951

Amaze your colleagues. Learn coarse language!
Step right up!
—Barnum-style advertisement for this book

This introductory chapter explains why I invented C-OAR-SE theory, what it is, how it differs radically from conventional psychometric theory, and where I've updated it.

After reading this chapter you should be able to:

- Remember that C-OAR-SE is an acronym for the six aspects of the theory
- Understand the fundamental difference between C-OAR-SE theory and psychometric theory
- Question the catchall notion of "random errors of measurement"
- Understand the difference between fully researcher-defined (psychological) constructs and partly rater-defined (perceptual) constructs
- Appreciate that rational expert judgment is always needed to design a new measure (or to critique an old one)
- And why statistics are *not* needed

1.1 Why C-OAR-SE?

I invented a new theory of measurement because I became completely disillusioned with conventional psychometric theory—the measurement theory that is universally employed in the social sciences. This is due mainly to the wide influence of books on psychometric theory by Guilford (1936, 1950), Nunnally (1967, 1978), and the article by Churchill (1979). The influence is likely to continue if the two new books, *Handbook of Psychological Testing* by Kline (2000), and *Psychological Constructs* edited by Embretson (2010), become popular. The blatant aim of the present book

J.R. Rossiter, *Measurement for the Social Sciences*,
DOI 10.1007/978-1-4419-7158-6_1, © Springer Science+Business Media, LLC 2011

is to stop this from happening and to replace psychometrics with something that is scientifically much better.

Psychometrics has been defined as "the integration of psychological theory with formal statistical models" (slightly paraphrased from a definition by Millsap, 1994, Editor of *Psychometrika* at the time of writing) and this is where its problems originate. Psychometric theory, I allege (see Rossiter 2002a), by its reliance on statistics for validation, has produced and continues to produce many erroneous empirical results and therefore leads researchers to wrongly accept and reject hypotheses and entire theories. In this book, I will provide numerous arguments and examples to support this very serious allegation and explain what should be done about it.

C-OAR-SE is first and foremost a *theory* (it is secondly a *procedure*). It is a *rational* theory, testable only by the "evidence" of logical argument. It is *not* testable by any empirical means—other than, I hope, the eventual empirical outcome of better social science knowledge. In philosophical terms (see especially the works of the German philosopher Gottlob Frege and the British philosopher Bertrand Russell), the principles of C-OAR-SE are "analytic"—true by definition—not "synthetic" and reliant on empirical proof.

"C-OAR-SE" is an acronym but it is also a pun. The punnish name was chosen to punish the "overrefined" statistical approach to developing and validating measures. Overrefinement produces precise but wrong scores, and thus misleading data and erroneous conclusions. Researchers routinely use poorly chosen measures and yet treat the scores from these measures as impossibly precise—such as reporting to 3 or 4 decimal places to make results look more accurate than they really are, and obsessing about *p*-values in significance tests no matter how bad the measures involved, to name the most common overrefinement practices. Statistical overrefinement is responsible for the lack of realism in measurement. Most measures in the social sciences today lack realism because they do not measure what they are supposed to measure. Many examples of unrealistic measures are given in this book, taken from the leading journals.

C-OAR-SE is an acrostic acronym for the six aspects of the theory:

1. Construct definition
2. Object representation
3. Attribute classification
4. Rater-entity identification
5. Selection of item-type and answer scale
6. Enumeration and scoring rule

The "OAR" in C-OAR-SE signals the central theoretical idea that a *construct* consists of *three* elements: (1) O, the *object* to be rated, (2) A, the *attribute* on which it is to be rated, and (3) R, the *rater entity* who does the rating. The OAR conceptualization of a construct (and the O, A, and R classification scheme, which is explained in detail in this book) is shown in Fig. 1.1.

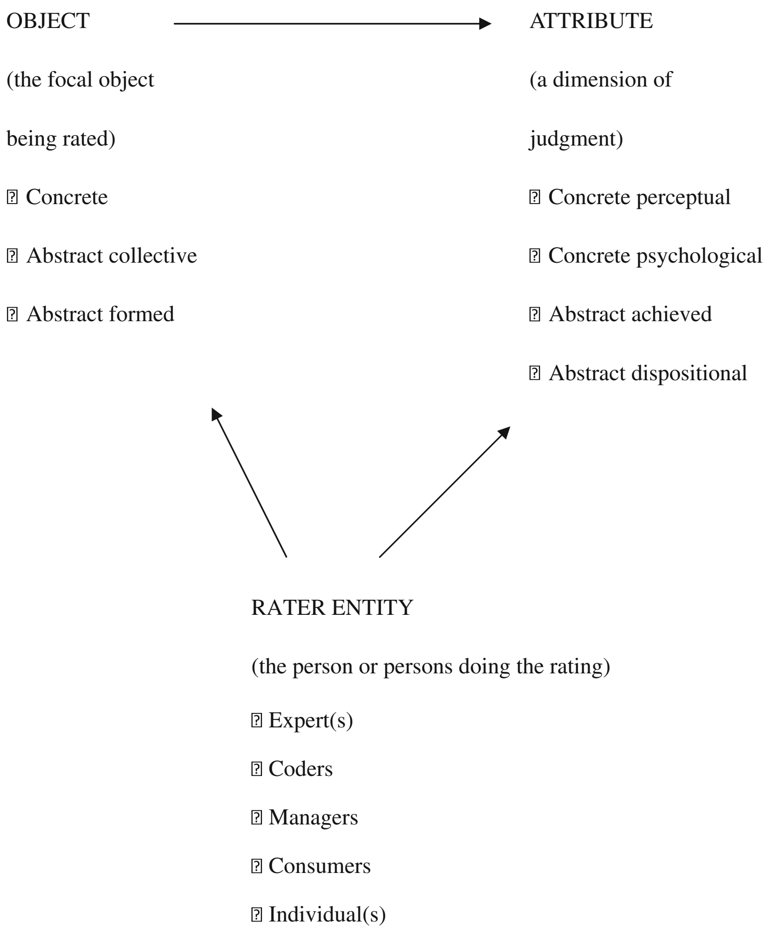

Fig. 1.1 The elements of a construct (object, attribute, rater entity) and a preview of the classifications of each element. In the rating, the object is projected onto the attribute (hence the *one-directional arrow*). The rater has to apprehend both the object and the attribute (hence the two arrows from the rater entity)

As we will see, conventional psychometric theory, to its detriment, focuses entirely on the *attribute* element. Psychometricians—invariably—refer to the attribute as a "construct" when it is only *one part* of a construct.

1.2 The C → M → S Structure of Measurement

The fundamental difference between C-OAR-SE and conventional psychometric theory is best understood by considering that *measurement* involves three stages, as depicted in Fig. 1.2.

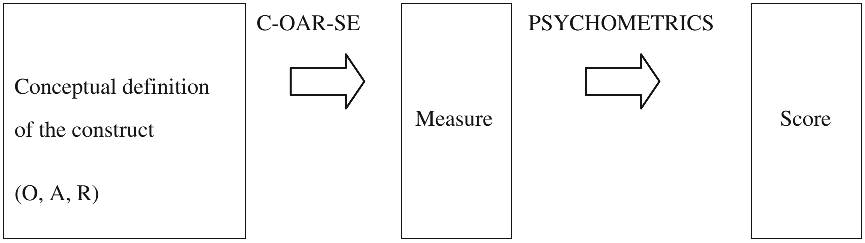

Fig. 1.2 The Construct → Measure → Score structure of measurement

C-OAR-SE theory, as a rationalist theory, is "front-ended" on the first two stages (C→M). These two stages are the domain of *content validity*. Content validity is the primary and *sine qua non* ("without which nothing") form of validity. Content validity (which I will sometimes abbreviate as CV) addresses the basic question of validity attributed to Kelley (1927), which is, slightly paraphrased, "Does the measure measure what it is supposed to measure?" Content validity refers to the *semantic correspondence* between the conceptual definition of the construct, C, and the measure, M. "Fully content valid" means *semantic identity* of C and M; that is, that the measure exactly represents the construct as defined. For "abstract" constructs (as explained later) the most realistic aim is for a *highly* content valid measure but for "doubly concrete" constructs *full* content validity is the aim.

Psychometric theory, in contrast, is "back-ended" on the last two stages, which are the measure itself and its scores. Psychometric theory tries to establish the validity of the construct, C, by examining the relationship between the measure, M, and the scores it produces, S, via the inference S → M. Totally illogically, the psychometric approach ignores C, the construct itself! As one of my rare colleagues who understands C-OAR-SE, Tobias Langner, remarked, this is equivalent to "trying to evaluate whether a cake was made correctly according to recipe by tasting it" (Langner, 2008, personal communication). I have provided a detailed comparison of C-OAR-SE with the Nunnally–Churchill version of psychometric theory in the appendix (Appendix A), mainly so that young researchers will have ammunition to fight off conventional reviewers.

I am going to now give a typical example of the psychometric fallacy of evaluating a measure's validity from the scores it produces (S → M) rather than by looking at the semantic correspondence between the construct definition and the content of the measure (C → M). My field's leading journal, the *Journal of Marketing*, published a study of travel agencies (in the March, 2009, issue). The constructs are in all caps and the items purported to measure them are as follows:

- PERCEIVED RELIABILITY OF THE TRAVEL AGENCY

1. "I make sure that the booked travel meets my expectations."
2. "I make sure to get what I want for my money."

- PERCEIVED QUALITY OF THE BRANDED SERVICES OFFERED BY THE TRAVEL AGENCY

  1. "I can trust to get the products of well-known travel companies."
  2. "I get branded products of established tour operators."

If you cannot see the almost total lack of correspondence between the construct and the content of the items in both cases, then don't bother to read further, because you won't get the main message of this book. The reviewers must have looked only at the "psychometric properties" of the measures, which are impressive: the coefficient alphas (actually just bivariate correlations) for the two-item measures were $\alpha = .93$ and .94. This example is typical of the rubbish that passes muster in our leading journals, in every issue. I pick on other equally low-validity examples from the other social sciences throughout this book, and also I use them in many of the end-of-chapter questions.

1.3 New True-Score Model: $O = T + D_m + E_r$, Where D_m Is Distortion Caused by the Measure and E_r Is Error Caused by the Rater

All measures aim to produce *true scores*, that is, ratings that fully represent the construct.

Conventional psychometric theory is based on the classic "true score" model made famous long ago by Lord and Novick (1968) and invented earlier by Spearman (1904). This model is usually symbolized as $O = T + E$, where O is the observed score, T is the true score that we intend to measure, and E is "random errors of measurement." These random errors are assumed to occur during the act of measurement and to be caused entirely by the respondent or *rater* (fatigue-induced wrong entries, guessing of answers, or malicious responding such as checking the midpoints of a series of bipolar answer scales or omitting ratings when the rater knows the answers). The classic true-score model further assumes that rater errors, being random, will *cancel* one another over a large number of items so that the E term becomes zero and the *average of the observed scores* on the items is the true score.

However, the much larger source of error is not rater error (which I call E_r), but rather *distortion* caused by the *measure* (D_m). Distorting departures from the true score are caused specifically by *low content-validity* of the question part of the measure *or* the answer part. (Content validity, as explained in Chapter 2, comprises item-content validity, meaning high semantic correspondence of construct and measure, and answer-scale validity, meaning that the answer alternatives allow the rater to express the true score and only the true score.) Measure-induced distortion, D_m, *systematically* biases the observed score away from the true score; it is *not* random, and so averaging distortion across items does *not* cancel it to zero. The true score can never be observed by averaging the scores of inaccurate (i.e., non-content-valid) multiple items, which is what the "partial interpretation philosophy" adored by the psychometricians assumes (e.g., Steenkamp and Baumgartner 1998). Nor can it be observed in "purifying" collections of items by deleting those whose scores do not correlate highly with the total score, as in factor analysis, a classic "statistical crutch" psychometric procedure.

The new C-OAR-SE true-score model is $O = T + D_{\mathrm{m}} + E_{\mathrm{r}}$, where D_{m} is *measure-induced distortion* of the true score caused by low content-validity of the measure—its item(s) *and* its answer scale—and E_{r} is rater error caused by individual raters' mental state-induced or motivation-induced careless or malicious answering.

The main way that the *measure* causes distortions of the true score can be found in the answer scale used for the items. As documented in Chapter 6 on item types, answer scales that do not allow the rater to easily give true intended answers can drag observed scores away from the true score. This common problem is *not* correctable statistically after the fact. Psychometric techniques based on the classic true-score model that attempt statistical corrections after the fact are useless. The measure—the items *and* the answer scale—have to be highly content-valid in the first place.

The *new* true-score model is, therefore, a much better guide for researchers because measure-induced distortions, not rater errors, are the main problem to guard against. Guiding and guarding is what C-OAR-SE theory does (as does the DROAVR checklist, which I employ as a summary device in the final chapter).

1.4 Researcher-Defined (Psychological) Versus Partly Rater-Defined (Perceptual) Constructs

The psychometric approach in effect allows the *raters* to select the items for the measure. Therefore—in another instance of its "backwards" procedure (i.e., an S→M→C reversal of the sequence in Fig. 1.2 earlier)—it allows the *raters*, instead of the researcher, to *define* the construct! Specifically, it allows raters to define the *attribute* of the construct, which is arguably the most important part.

Consider, for example, the attribute labeled as SERVICE QUALITY. (In this book, as I have done in all my journal articles on C-OAR-SE, I will denote the main OBJECT, ATTRIBUTE, and RATER ENTITY of the construct in ALL CAPS; Constituents and Components of OBJECTS and of ATTRIBUTES in Upper and lower case; and first-order, that is, lowest-order, object-and-attribute *items* in "Quotation marks.") The most popular academic measure of SERVICE QUALITY is the multiple-item SERVQUAL instrument (Parasuraman, Zeithaml, and Berry 1988). The 22 items in SERVQUAL were essentially chosen by the raters (by CONSUMERS). This is because Parasuraman et al. (in their 1985 article, the forerunner of SERVQUAL) asked CONSUMERS to rate a much larger set of items that they, the researchers, gathered from open-ended interviews with a previous sample of consumers; the researchers then subjected these ratings to factor analysis from which, from the item intercorrelations, they pulled out five artificial "factors" or "dimensions" which they labeled as Empathy, Assurance, Reliability, Responsiveness, and Tangibles. Other items—even if CONSUMERS had *told* the researchers that the item was an important and for them an *essential* component of SERVICE QUALITY—were simply discarded because they didn't "load" significantly (weren't correlated with) one of these artificial factors! Parasuraman and colleagues (2005) later repeated the mistaken psychometric approach in developing their measure of E-RETAILING SERVICE QUALITY, called E-S-QUAL, which I have criticized in a recent article in *Service Science* (Rossiter 2009a) and in which I constructed a much more valid measure.

Many examples of essentially rater-defined constructs plague the social sciences literature, as we shall see. I will just add here an example from organizational behavior (O.B.) research to avoid the reader getting the impression that my criticisms are all about marketing (they're not: all the social sciences are similarly and equally at fault). A 2006 study in the *Journal of Business Research* included the construct—or rather the attribute—of RELATIONSHIP COMMITMENT. However, the researchers never bothered to define what they meant by "relationship commitment." Instead, they assumed that all readers (and raters) *know* what it means, and in the theory part of the article they merely talked about the role of this assumed self-evident construct in the theory they were proposing. When it came to *measuring* RELATIONSHIP COMMITMENT, the researchers—as is all too typical—simply borrowed another researcher's items and neglected to check the content of these items against the conceptual definition of the construct (which they couldn't anyway because they didn't provide one). Here are the items (there are only two) that they used to measure RELATIONSHIP COMMITMENT. The first item was "The distributor actively works together to carry out its responsibilities and commitment in this relationship." Ignore the awkward grammar in this item and note that this item does not represent a component of the COMMITMENT attribute, but rather mentions "commitment" directly, thus assuming that the raters know what it means and that all have the same meaning in mind. The content of this item is, in any event, confounded by reference to "commitment" *and* "responsibilities." The second item was "The distributor invests considerable resources and time to make the relationship a success." Presumably, a Likert-type "agree" answer to this question implies "commitment," but the item refers to a particular *consequence* of commitment rather than being another meaning (a *component*) of it. The researchers then justified use of this paltry two-item measure by the fact that observed scores on the two items were highly correlated ($r = .77$), a mere statistic that says nothing at all about the items' content validity. I'm sure you see the illogic in this typical psychometric approach. Content validity is completely ignored and, therefore, so is the construct ignored.

It is the responsibility of the *researcher* to conceptually define the construct or constructs to be measured (this is the first aspect of C-OAR-SE theory and the first step in the DROAVR checklist in the final chapter). However, this does not mean that rater input is always unnecessary, as I will now explain.

Some constructs are *psychological* constructs, such as the Freudian constructs of subconscious PROJECTION and subconscious REPRESSION (see Freud 1911, and observe the work of a master theoretician whose ideas have never been more relevant than they are today, what with pedophilia, incest, and adultery hogging the headlines, all tips of a giant, ever lurking sexual iceberg). Researchers cannot ask the sufferers to define these constructs for them! This is like accepting teenagers' meaning of PERSONALITY, as in "He's (or she's) got no personality," in place of the scientific definition of the construct (see Siegel and Castellan 1988, p. 1, for this and similar examples). A topical example of a psychological construct relevant to marketing, not just to psychology, is IMPLICIT ATTITUDE (see the important review book edited by Petty, Fazio, and Briñol 2009). IMPLICIT ATTITUDE is topical because of the current fascination with "nonverbal" measures and also with neuroscience, or "brain-imaging" research. (I discuss the construct and measurement of IMPLICIT ATTITUDE in Chapter 4.) In the case of psychological constructs, the *researcher* alone has to define the construct—the object and possible constituents or components and the attribute and possible components—and then the *researcher* must choose the item or items to measure it. Rater input is not necessary.

Other constructs are *perceptual* constructs. A clear example is THE COMPONENTIAL SERVICE QUALITY OF COMPANY Z AS PERCEIVED BY ITS CUSTOMERS (Rossiter 2009a). Here the researcher wants to find out what the company's CUSTOMERS regard as constituting "good" and "bad" service quality. Therefore, CUSTOMERS, as *raters*, must be asked for their perceptions before the measure can be designed. However, it is the *researcher* who must decide on the *final* set of items in the measure, even of a perceptual construct such as this. (The researcher should do this by judgmentally eliminating redundantly paraphrased perceptions and then performing a simple frequency count to identify the most prevalent perceptions, which then become the defining *components* of the SERVICE QUALITY attribute in the construct.)

1.5 Ultimate Dependence on Rational Expert Judgment for Defining Constructs

Researchers are timid creatures when it comes to defining the constructs that they want to use in their research. They are like sheep (to invoke the DROAVR simile from this book's final chapter). Nearly all of them will look for an "established" measure and accept *its* definition as correct. The more researchers who join the flock, the more that "precedent" can be cited as "validation" of the measure. I call this the "sheep fallacy" in choosing measures and it is rife in the social sciences.

The social sciences could be saved *as* science if young researchers were required (for their doctorates, for instance) to think up new constructs or at the very least to propose better definitions of old ones. All they would need to do is adopt the C-OAR-SE procedure and define (1) the *object class* in the construct (e.g., LOW-INCOME HOUSEHOLDS; TELECOMMUTERS; LUXURY PRODUCTS; WEB BANNER ADS), (2) the *general attribute* on which the objects are to be rated (e.g., attributes fitting the above objects might be: CONSUMPTION OF A HEALTHY DIET; PRODUCTIVITY; MINIMUM PRICE POINT IN THE CATEGORY; ATTENTION-GETTING ADVERTISING EXECUTION), and (3) the relevant *rater entity* (e.g., to fit the above, respectively: CODERS in a content analysis, or should that be "contents" analysis, of household garbage—see the classic work by Webb, Campbell, Schwartz, and Sechrest 1966, which marketing and social researchers would do well to read because it is a fine example of realist measurement; TELECOMMUTERS by self-report; CONSUMERS as the first rater entity and then possibly a survey of retail prices in which the second rater entity is CODERS; and, once more, CODERS). Young researchers could then make a genuine contribution to knowledge by designing their own *measure* of the construct. All they need to do to *validate* their measure is conduct a *content-validity check* for semantic correspondence between their conceptual definition (elaborated to include object constituents or components and attribute components if the construct is "abstract," as in the above cases) and the measurement items, and then devise a valid answer scale for the items.

All of the above, like C-OAR-SE, is an exercise in *rationality*, not empiricism. The only expertise required of the young researcher is good knowledge of colloquial English language or whatever language the measure is going to be worded in. For instance, I'm an expert in English but not, unfortunately, in German, and especially not colloquial German. I'm a visiting professor in Germany at my colleague Tobias Langner's research institute, but cannot do content-validity checks of the measures we use there and nor would I attempt to.

1.6 The Construct Definition Depends on the Role of the Construct in the Theory

The first theoretical aspect of C-OAR-SE is *construct definition*. The construct must be defined in terms of its (1) object, (2) attribute, and (3) rater entity, and an abstract construct's definition must include specification of object constituents or components (if the object is abstract) and attribute components (if the attribute is abstract).

A very important principle that should be introduced at the outset (in this first chapter) is that the definition of the construct can differ depending on its role in the overall *theory*, which is being investigated and tested. The best and most central illustration of this principle is the *perceptual* (self-reportable) construct of EXPLICIT ATTITUDE (TOWARD AN OBJECT WITH THE SELF AS RATER). As pointed out in my original article on C-OAR-SE (Rossiter 2002a), the theory may be about attitude *formation* (e.g., expectancy-value or multiattribute attitude theory), in which case COMPONENTIAL ATTITUDE must be defined—and measured—by its *causal components*, which are beliefs and emotions (see Rossiter and Percy 1997). Or the theory may be about an already formed *existing* attitude, in which case OVERALL ATTITUDE is defined as a *concrete* attribute and measured with one good—that is, fully content-valid—item and answer scale (see Bergkvist and Rossiter 2009, Rossiter and Bergkvist 2009). The conceptual definition of the construct, therefore, should always be preceded by a statement of how the construct will be used *theoretically* in the researcher's theory or model.

Acknowledgment of the construct's role in the theory is another reason why the researcher must define the construct *fully* in terms of its object, attribute, and rater entity (and constituents and components of the first two elements if the construct is abstract). A widespread—and sheepishly followed—example of this failure is marketing researchers' attribute-only definition of MARKET ORIENTATION (e.g., Narver and Slater 1990). The definition should be MARKET ORIENTATION OF THE COMPANY OR ORGANIZATION AS RATED BY (E.G.) ITS MANAGERS, thus clearly acknowledging that it is a construct within *management* theory. It can be distinguished, in management theory, from an ENTREPRENEURIAL ORIENTATION (see Miller 1983, for the first apparent use of this construct and see Merlo and Auh 2009, for an application of it).

The eminent statistician—or rather measurement theorist—Clyde Coombs, inventor of conjoint measurement and the earliest academic advocate of the practitioners' "pick any" measure of BELIEFS (see Coombs 1964), is reported to have said, in an astounding reversal of the usual practice, "If the data don't fit the theory, throw out the *data*." International bureaux of standards in physics and chemistry regularly do just this if new data don't closely fit the accepted standard estimate (see Hedges 1987). I would amend Coombs' advice from the perspective of C-OAR-SE theory. My advice is: "If the data don't fit the theory, *examine the measures*." A theory cannot be accepted (or rejected) if its constructs and their causal relationships have not been measured properly.

It is easy to find examples of bad science caused by bad measurement in the *leading* journals in *every* field of social science—and I do so zealously in this book. The bad science problem will remain unchallenged and uncorrected unless researchers understand my C-OAR-SE theory of construct measurement. Then, it would help if they adopted my DROAVR application checklist given in the final chapter when planning, reporting—and also when reviewing and evaluating—social science research.

1.7 End-of-Chapter Questions

(1.1) You have been asked to explain the *main* difference between the C-OAR-SE approach and the psychometric approach to designing measures. Based only on this chapter, and without paraphrasing what's written in the chapter—that is, in your own words—write an explanation of up to 500 words. You may find it helpful to include a small diagram additional to your verbal answer. (7 points)

(1.2) Think up one or two more examples of rater errors (E_r) other than the examples in the chapter. For each, discuss whether or not you believe the scores resulting from the errors would be random (a) across raters on any one item and (b) across items for an individual rater in a multiple-item scale—in other words, would the rater's error scores average out to zero in both cases? (10 points: 5 points maximum for discussing one example and 10 points maximum for discussing two examples)

(1.3) Download the article by Deligonul, Kim, Roath, and Cavusgil from the *Journal of Business Research* (see References near the end of this book). Read their definition of MANUFACTURER SATISFACTION on p. 805, then look at the researchers' two-item measure of this construct in Table 1 on p. 806, where for some reason they've relabeled it. First, based on the "OAR" core of C-OAR-SE, explain which is the better label and why, and then suggest a highly accurate label. Second, evaluate the measure from the standpoint of C-OAR-SE theory as you have understood it from this first chapter. Do not look further into the book, yet. (8 points: 3 points maximum for the first part and 5 points maximum for the second)

(1.4) Discuss the order and relative contributions of rater input and expert judgment in designing measures of (a) the psychological construct of NEED FOR COGNITION (see Cacioppo and Petty 1982) and (b) the perceptual construct of POLITICAL CORRECTNESS (no reference needed) and then (c) summarize the differences. (15 points: maximum 5 points each for a, b, c)

(1.5) Why should Clyde Coombs' (reported) advice be amended as I amended it in Section 1.6 of the chapter? More broadly, what would be your answer to the dumb question that I often get asked, which is "How can you prove *empirically* that C-OAR-SE is a better approach?" You may find it helpful to read the second chapter before answering this question although if you think hard about what I said in this first chapter, you should get it. (12 points: 5 points maximum for the answer to the first question and 7 points maximum for the second)

Chapter 2
Validity and Reliability

> *Valid: soundly reasoned, logical.*
> *Reliable: dependable, safe.*
> —Collins Dictionary & Thesaurus (2002)

The concepts of *validity* and *reliability* of measures are defined (and also assessed) differently in C-OAR-SE than in conventional psychometric theory. Acceptance of the new definitions of validity and reliability is essential if you want to apply C-OAR-SE.

After reading this chapter you should be able to:

- See that "construct validity" is a misnomer
- Learn that content validity, which is the only essential type of validity, consists of item-content validity and answer-scale validity
- Understand the logic problem with the standard multitrait-multimethod (MTMM) approach to validating measures—and be able to reproduce the arguments for *not* using MTMM when you switch to C-OAR-SE in your research
- See why predictive validity is desirable but not essential for a measure
- Distinguish the only two important types of reliability, which are stability-of-scores reliability and precision-of-scores reliability

2.1 Content Validity (CV) Not "Construct Validity"

According to the Construct → Measure → Score structure-of-measurement model introduced in Chapter 1, only *content validity* matters. Nothing more than *content validity* is required to "validate" a measure. This is because content validity completely covers the C→M relationship. Validity has nothing to do with the M→S relationship that is the focus of psychometric theory. The M→S relationship *excludes* the construct, C.

What has not been realized by anyone is that the high-sounding term "construct validity" is nonsense. To "validate" means "to establish the truth of." But a "construct" is a *definition*. A definition can be judged as reasonable or unreasonable but *not* as true or false.

J.R. Rossiter, *Measurement for the Social Sciences*,
DOI 10.1007/978-1-4419-7158-6_2, © Springer Science+Business Media, LLC 2011

Only a *measure* can be validated—in relation to the construct as defined. This is *content validity*. Content validity asks the question "how truthfully does the measure represent the construct?"

Content validity is *not* another name for *face validity*; the former is inescapable and the latter incapable. With face validity, which is basically just appraising the measure after the fact, you can only see the items *retained* for the measure, not the ones that were missed altogether or deleted for erroneous statistical reasons. Even with these *fait accompli* items, the judges will not know how to assess the content validity of those items unless, first, the researcher has provided a detailed conceptual definition of the construct and second, the judges have read and understood C-OAR-SE theory. Why is an understanding of C-OAR-SE necessary? Because without C-OAR-SE, the judges of face validity won't know to look for the "deeper" content: the object, the attribute, and the attribute and its *level* (see Chapter 6) in each item. (In the author's long experience, reviewers of academic studies *never* look at the questionnaire items *or* at the items' answer format to assess validity. Instead they look for supportive numbers from the *scores* from the M→S "back end" as represented by convergent and discriminant correlations or, even more irrelevant, by coefficient alpha.) After using sloppy measures, researchers report mean scores and correlations to three or more decimal places and apply statistical tests to the nth degree ($p < .001$, for instance, or even the impossible $p < .0000$ in push-button statistical software!) as though a veneer of precision makes the measures more valid.

To be *content valid*, each item in the measure must have *both* of the following properties:

(a) *High item-content validity*. This means that the semantic content of the question part of the item corresponds closely with a constituent or component of the object in the conceptual definition of the construct *and* with a component of the attribute in the conceptual definition *unless* the attribute is in the answer part––see (b) below. For basic "doubly concrete" constructs (clear single object and clear single attribute) a single item, only, is necessary because there are no constituents or components, but even the most "abstract" of constructs is ultimately represented in the measure by single items measuring the first-order components. As Rossiter and Bergkvist (2009, p. 8) point out, "all measures are, or are aggregations of, single items." It is vital that each and every item be highly content-valid. This truism is often overlooked by psychometricians; they believe that numerous sloppy items when averaged can somehow compensatingly arrive at the true score.

(b) *High answer-scale validity*. This means that the semantic content of the *answer part* of the item allows the rater to see only the main answer alternatives that he or she has in mind, and to easily choose *one* answer that fits his or her true score. The answer part of the item is always an *attribute*—either a *second* attribute (disagreement–agreement) in the case of the popularly used Likert measure and also in the DLF IIST Binary measure (see Chapter 6), or else the main or component attribute in the case of *all other* measures.

A content-validity check is a two-step process carried out by the researcher. *Open-ended interviews* with a sample of three experts (if EXPERTS are the rater entity), five of the least-educated managers (if MANAGERS are the rater entity), or ten of the least-educated consumers (if CONSUMERS are the rater entity—see Chapter 5) are advisable as a *pretest* of the *researcher's* initial choice of item content and answer format. These interviews are semi-structured with open-ended answers that are themselves content-analyzed by the researcher. The *researcher* then finalizes the item or set of items and their answer scales (which will be the same answer scale for all items for a given construct unless "behavioral categories" answer scales are used—see Chapter 6). No further pretesting is needed.

Content validity in C-OARSE is, therefore, much more sophisticated than measurement theorists realize; it is much more difficult to achieve than they realize; and it is the primary and *sine qua non* ("without which nothing") form of validity.

2.2 Why MTMM Is Wrong

The reigning theory of the validity of measures is *multitrait-multimethod theory*—commonly abbreviated as MTMM—introduced into the social sciences by the psychologists Campbell and Fiske (1959). Every social science researcher should learn this theory and then read here (and in Rossiter 2002a, for the same arguments paraphrased) why it is logically wrong. MTMM is the "more-the-merrier mistake." Multitrait-multimethod theory is a back-end and *backward* theory (it argues from scores to measure, i.e., S → M), which contends mistakenly that the validity of the *construct* can be established empirically by comparing its *measure* with other measures. "Construct validity"—and Campbell and Fiske *meant* "measure validity"—is said to be demonstrated empirically if the *scores* on the measure exhibit both "convergent" validity and "discriminant" (divergent) validity with scores on *other* measures.

Convergent validity of a new measure M_1 is said to be demonstrated if scores on this measure, S_1, correlate highly with scores S_2 on an "established" measure M_2 of allegedly the same construct, C_1. (The more it correlates, the merrier the fool of a researcher.) To give an example, consumer behavior researchers are typically happy with shortened versions of Zaichkowsky's (1994) lengthy (20 items) measure of PERSONAL INVOLVEMENT (OF THE INDIVIDUAL WITH AN AD OR A PRODUCT CATEGORY). If scores on the short measure correlate highly (say $r = .7$ or higher, which would be about 50% or more "shared variance") with scores on the original 20-item measure when both are administered to the same respondents, or when the scores on the short measure are extracted from scores on the long measure by factor analysis, which is the *usual* way of shortening measures, then the new measure is said to be valid in the *convergent* sense. But convergence assumes that the "old" measure, M_2, is *content-valid* to begin with! And note that the old measure, being the only one in existence at the time, could not *itself* have had a convergent validity test! What really matters is the intrinsic content validity of the new measure, M_1, not any crutch-like correlation with scores from *another* measure. Both measures, M_1 and M_2, could have *low* content validity with regard to the presumed common construct C_1—thereby making them *both* unusable—while their scores, S_1 and S_2, spuriously "converge" due to both measures sharing a content-validity error such as common methods bias in the answer scale. Convergent correlation, therefore, provides no proof whatsoever of the validity of the measure.

Discriminant (or divergent) validity, which is the other less frequently invoked "half" of MTMM, has the same logical flaw. Discriminant validity requires that scores S_1 on the new measure M_1 of construct C_1 (the original construct) do *not* correlate highly with scores S_3 on measure M_3 of a *different* construct (call it C_3). For example, scores on the new shortened measure of PERSONAL INVOLVEMENT OF THE INDIVIDUAL WITH HIP-HOP MUSIC might be shown to correlate only $r = .25$ (a "small" correlation according to Cohen's, 1977, "effect size" rules-of-thumb—and see also Appendix C in the present book, which gives binary effect sizes in percentage terms) with scores on a measure of PURCHASE FREQUENCY OF HIP-HOP CDs AS REPORTED BY THE INDIVIDUAL. Following the mantra of MTMM, the researcher then concludes from this small correlation that the two measures are measuring "distinct" constructs, namely C_1 and C_3. To be fair, the researcher is usually obliged to nominate for comparison a construct that is distinct but within the same overall theory rather than a construct from a different theory, which would be too easy a test of distinctiveness or discrimination. But here's where it gets *really* illogical because the researcher will then want to use the *same* small correlation used to prove they are different, to show that C_1 and C_3 are *related* (e.g., that PERSONAL INVOLVEMENT WITH HIP-HOP MUSIC is one *cause* of PURCHASE FREQUENCY OF HIP-HOP CDs). The fact that scores on a new measure are only weakly correlated with scores on another measure implies nothing about the validity of *either* measure. Discriminant validity, like convergent validity, is not validity.

MTMM—the "more-the-merrier mistake"—is yet another instance of the psychometric approach leading sheepish researchers astray. MTMM theorists try to prove that M represents C by looking only at S in the C→ M → S framework given earlier; the construct itself, C, never comes into it!

C-OAR-SE theory postulates that *content validity* is all that is required to demonstrate the validity of the measure (in relation to the construct). Content validity (CV) in turn consists of item-content validity (CV_{item}) and answer-scale validity ($\text{CV}_{\text{answer}}$), as explained in the next two sections of the chapter.

2.3 Item-Content Validity (CV_{item}) and How to Establish It

Establishing item-content validity (CV_{item}) is different for psychological and perceptual constructs. The two types of construct were distinguished in Chapter 1 and become relevant again here.

Psychological constructs. Psychological constructs are invented constructs—invented and defined by social science researchers—and cannot be observed directly (see the classic article by Nisbett and Wilson 1977, and the updated review article by Wilson 2009). Instead, the existence of a psychological construct is inferred from its manifestation(s) or *effect(s)*. This effect or these effects must follow from *theory* and be represented in the conceptual definition of the construct.

With an *abstract psychological* construct, which has multiple meanings and is the most difficult type of psychological construct to validly measure, the semantic content of the definition is likely to be technical. By "technical" is meant that the object (e.g., LIBERTARIAN) or the attribute (e.g., INDIVIDUALISM-COLLECTIVISM) or both is not in everyday language. However, the definition of an abstract psychological construct must be expanded to include everyday language descriptions of the *components* of the abstract object or abstract attribute. Moreover, the components must be *concrete*—having a *single* meaning—otherwise the researcher cannot select items to measure them. Another way of putting this is that the components must be *real*. For example, the LIBERTARIAN researcher should go back to J.S. Mills' writings to see how he described the components of this abstract psychological *object*. The researcher would find that the object involves particular concrete and clearly understandable Beliefs (or Attitudes in the traditional sense) as *components* of the object. Or take the abstract psychological construct that incorporates the attribute, INDIVIDUALISM-COLLECTIVISM. This construct, originally a group-level or "cultural" construct, has more recently been redefined as the individual-level personality trait—more correctly, the *learned disposition*—called INDEPENDENT VERSUS INTERDEPENDENT SELF-CONSTRUAL (see Brewer and Chen 2007). For this personal dispositional construct, the *object* is the SELF because the rater is rating his or her own disposition, and the *rater entity* is the INDIVIDUAL. The items are *mental or behavioral activities* that represent real-world manifestations—self-observable, self-reportable effects—of the disposition. While the attribute is abstract and psychological, the *items* are *concrete*, so they must be written in everyday language because this is what raters have to respond to on the questionnaire. The items refer to thoughts and behaviors that clearly signify INDEPENDENCE or else INTERDEPENDENCE. In many SELF-CONSTRUAL inventories, these are separate items, but since INDEPENDENCE and INTERDEPENDENCE are opposing ends of a single theoretical attribute, I believe the forced-choice type of item where the rater must answer one way or the other proves a more valid measure (more on the binary answer format in Chapter 6). A good item might be

"I would say that *most* of the time
(CHOOSE ONE ANSWER):
□ I prefer to be on my own
□ I prefer the company of others"

This example assumes that in the expanded definition of the construct—which should be given in the theory part of the article or research report—there is a component of the overall attribute of SELF-CONSTRUAL (for short) that refers to the everyday language term of Sociability (or a similarly understandable label). Other *defining* component attributes for this construct might be Group decision preference, Respect of group rights over individual rights, and Seeking advice from others before making big decisions. Actually, now that I think about them, these component attributes aren't all that concrete (specific) and should better be considered as second-order, with the overall SELF-CONSTRUAL attribute moving up to *third-order* in the conceptual hierarchy of the construct definition. Then, several first-order (lowest level) items can be written for each component attribute, like the item above for Sociability.

Item-content validity for an *abstract psychological* construct then becomes a fairly simple matter of checking that the item wording accurately conveys the meaning—in plain-English *dialect* or whatever the language of the questionnaire—of the relevant *component* object, if the object is abstract, and of the *component* attribute, if the attribute is abstract, and is a concrete (single-meaning) statement of it. Timid researchers may want to engage a couple of literate colleagues or acquaintances to "verify" their selections (especially if the researcher is an INTERDEPENDENT!).

Perceptual constructs. Perceptual constructs are much easier to establish item-content validity for. Perceptual constructs, as the name suggests, *can* be observed directly; they are the observations made *by raters* about the object. The two leading examples of perceptual constructs in the social sciences are BELIEFS (ABOUT ATTRIBUTES OF OBJECTS) and OVERALL ATTITUDE (TOWARD AN OBJECT). The "object" may be animate, such as a group or person, or inanimate, such as a company, product, brand, or advertisement. Establishing item-content validity in these cases is easy because the belief or attitude is defined concretely and is measured the same way, thus approaching semantic *identity* between the construct and the measure. For example, the belief that AUSTRALIANS ARE FRIENDLY can be highly (and probably fully) validly measured by the item

"Australians, in general, are
(CHECK ONE ANSWER):
□ Very friendly
□ Friendly
□ Unfriendly
□ Very unfriendly"

And the overall attitude of LIKING OF THE BENETTON "NEWBORN BABY" AD (which all advertising researchers would recall, as it has been reproduced in many advertising textbooks) can be highly and possibly fully validly measured by the item

"[Picture of the ad]
How much do you like or dislike this ad?
(CIRCLE A NUMBER FROM –2 TO +2 TO
INDICATE YOUR ANSWER):
Dislike extremely –2 –1 0 +1 +2 Like extremely"

Note that I have also provided answer scales for these exemplifying items and it is to this second part of content validity that I turn shortly.

Before doing so, I need to discuss a complex case concerning whether a construct is psychological or perceptual. A very practically important construct is REACTANCE (see Brehm 1966). It is practically important—vital even—because it is the major cause of the failure of most health- and safety-promotion campaigns among the most at-risk audiences (see Rossiter and Bellman 2005, ch. 18). SELF-REACTANCE TO A RECOMMENDED BEHAVIOR, to give this construct its full label, is generally thought to be *perceptual*, in that people can self-report its presence—see, for instance, the 11-item (!) self-report measure of a "reactance disposition" in *Educational and Psychological Measurement* (Hong and Faedda 1996). However, I think REACTANCE can only be validly measured as a *psychological* construct—that is, not validly self-reported but validly inferrable by a *qualitative research* interviewer (see Chapter 8) using open-ended questions. Support for my assessment comes from the unbelievable findings in a study reported in one of my field's top journals, *Marketing Science*, by Fitzsimons and Lehmann (2004). In an experiment conducted with smart University of Pennsylvania undergrad students—I know because I taught at Penn for 5 years and these pre-med, pre-law, or pre-MBA students were the most savvy I have taught in 35 years of teaching—these researchers found that 92% of self-rated "high reactance" participants (self-rated on Hong and Faedda's 11-item measure) reacted against an expert's recommendation to *not* buy an evidently good-performing model of subcompact car. That is, 92% *chose* the car in contradiction of the recommendation (in a simulated choice against two other subcompact cars). What these researchers measured was more likely savvy students' "reactance" against a silly experiment that had all-too-obvious demand characteristics, or "transparency," to use this misused word. Much as I favor reactance theory as an explanation of counterattitudinal-message rejection, I would never cite this study in support.

With a *psychological* construct, the researcher will be misled by using a simple perceptual measure. This is one reason why I am such a strong advocate of *qualitative* research (see Chapter 8). There are some debatable "gray area" constructs but note that psychometrics don't help at all. Worse, psychometrics mislead. In Fitzsimon and Lehmann's study, Hong and Faedda's statistically "refined" *perceptual* measure of "psychological reactance" had an "impressive" coefficient alpha of .8 and thus was naively accepted by the researchers—and by the reviewers—as a valid measure of a *psychological* construct.

2.4 Answer-Scale Validity (CV$_{\text{answer}}$) and How to Establish It

The *answer scale* for an item in a measure is the other locus of content validity (the first locus is the item itself, as just explained). Content validity can, therefore, be expressed as $\text{CV} = \text{CV}_{\text{item}} \times \text{CV}_{\text{answer}}$. The two content validity terms, CV$_{\text{item}}$ and CV$_{\text{answer}}$, are multiplicative to indicate their complementarity; if either is zero there is *no* content validity overall for the measure and, indeed, both should *ideally* be 1.0, that is, both fully content-valid, which gives CV = 1.0 or 100%. In realist terms, however, especially for an abstract and, therefore, multiple-item construct, CV can only *approach* 100% (the adjective "high" to most people means "at least 80%"—see Mosteller and Youtz 1990).

Answer-scale validity ($\text{CV}_{\text{answer}}$) means that the answer part of the item allows, realistically, nearly all or, ideally, all raters to easily and quickly recognize that the answer alternatives fit the main possible answers that they could make. The answer alternatives provided should neither *underfit* (too few alternatives to allow a precise answer) nor *overfit* (too many, so that the rater will waver and cannot choose an answer that exactly fits). These complementary properties may jointly be called the "expressability" of the answer scale (a description coined by Dolnicar and Grün 2007).

As examples of $\text{CV}_{\text{answer}}$, consider the answer scales given in the preceding section on CV_{item}. The first example was the item for measuring (one component of the attribute of) the personal *disposition* of INDEPENDENT VERSUS INTERDEPENDENT SELF-CONSTRUAL, which was:

"I would say that *most* of the time
(CHOOSE ONE ANSWER):
□ I prefer to be on my own
□ I prefer the company of others"

Because of the attribute-qualifying *level* in the item ("... *most* of the time"), these are the only two possible answers. The answer scale, therefore, has perfect expressability. This is a "2-point behavioral categories" answer scale (see Chapter 6).

The second example was the *belief* that AUSTRALIANS ARE FRIENDLY, measured by the single item:

"Australians, in general, are
(CHECK ONE ANSWER):
□ Very friendly
□ Friendly
□ Unfriendly
□ Very unfriendly"

The answer scale in this case is "4-point bipolar verbal" (again further explained in Chapter 6). The answer alternatives are verbal because I hold to the theory that BELIEFS are mentally represented as verbal statements (Collins and Quillian 1969) whereas many researchers wrongly use *numbers* in belief-rating answer scales. Moreover, there is deliberately no middle answer category because an answer of "Average" is most unlikely and might also encourage evasion of a considered answer (see Cronbach 1946, 1950, for discussion of evasion and see Rossiter, Dolnicar, and Grün 2010, for evidence of it with "pick any" and midpoint-inclusive answer scales). The four *verbal* answer alternatives represent the most likely responses that raters are likely to think of in answering this item, so the answer scale has good "expressability" or, in other words, it is highly content-valid.

The final example was an item measuring an OVERALL ATTITUDE, namely LIKING OF THE BENETTON "NEWBORN" AD

> "[Picture of the ad]
> How much do you like or dislike this ad?
> (CIRCLE A NUMBER FROM –3 TO +3 TO
> INDICATE YOUR ANSWER):
> Dislike extremely –2 –1 0 +1 +2 Like extremely"

In this case, the answer categories are *numerical*. This is because overall evaluative responses ("attitude" *singular* in the modern sense; see Fishbein 1963, Fishbein and Ajzen 1975, and Fishbein and Ajzen 2010) are almost certainly represented mentally (and possibly physiologically, felt in the "gut") as a *quantitative bipolar continuum*. *Evaluative* responses such as OVERALL ATTITUDE are *conditioned responses* elicited automatically on encountering the stimulus object. They are quite unlike BELIEFS, which have to be actively "retrieved from verbal memory" or actively "formed on the spot" if a new belief. Moreover, and this could easily be tested, it is likely that most people discriminate only a couple of levels of "like" and a couple of levels of "dislike" for objects such as ADS, although for *important* objects, such as OTHER PEOPLE or, for many individuals, NEW CARS, I would use five levels each of like and dislike for valid "expressability" (i.e., –5 to +5). And on these numerical answer scales there *is* a midpoint, because some people can genuinely feel neutral about the attitude object, or have no conditioned evaluative response to it yet, and, being a single-item measure, there is little likelihood that raters would use the midpoint to evade answering.

In all three examples, the researcher has made a thorough attempt to think through the possible answers and to provide an answer scale whose alternatives match as closely as possible what's in the typical rater's mind after he or she reads the item. This is "expressability," or answer-scale validity, and the researcher should aim for a *fully* content-valid answer scale, although slight individual differences will inevitably make it only *highly* content-valid.

Answer-scale validity can be established practically in two ways.

The best method—especially when designing a new measure—is to look for the alternative answers during the *open-ended pretesting* of item content for clarity of meaning to the least educated in the sample of target raters. Simply present each new item alone and ask individual raters what answers they can think of for the item if verbal answers are planned, or how *they* would put numbers on the answers if numerical answers are planned (this will be rather simple quantification, as befits the realist nature of C-OAR-SE). Some very revealing findings about people's interpretation of answer categories can be found in the important article by Viswanathan, Sudman, and Johnson (2004) in the *Journal of Business Research*, where it will be seen that *most* answer scales "overdiscriminate" and thus *cause* rater errors.

The other method of finding a valid answer scale is to *study Chapter 6 in this book*, which discusses item types for the main constructs in the social sciences, and where “item type” consists of the question part and the answer part (i.e., the answer scale). Under no circumstances should you unthinkingly accept the answer scale from an “established” measure (this, again, is the “sheep’s way” of doing research). It is near certain that item-content validity is *not* satisfactory for the “established” measure. And I’ll bet my last dollar that the answer scale has not even been noticed, let alone properly validated (using the open-ended interviewing method outlined above).

2.5 The Desirability of Predictive Validity (PV) and the True Population Correlation (R_{pop})

Content validity, which I have abbreviated as CV (in a deliberate allusion to *curriculum vitae*, or credentials), is the only *necessary* property of a measure of any construct.

Only after the CV of a measure has been established—as fully or at least *highly* valid—can predictive validity (which I’ll abbreviate as PV) be considered. Although any old measure might by luck or coincidence turn out to be a good predictor of some valued outcome or criterion, social science researchers should be interested only in *causal* relationships between constructs—relationships that are predicted and explained by theory. To prove causality, it is necessary that both the predictor measure *and* the criterion measure be highly *content-valid*.

Most outcomes in the social sciences have *multiple causes* and this means that any one cause should not be expected to predict an effect at more than about $r = .5$.

Many predictive relations in the *health sciences*, such as the correlation between cigarette smoking and lung cancer, which is about $r = .18$, are far lower than this, and none exceeds $r = .40$ nor approaches the $r = 1.0$ assumed by many health researchers and the general public (see Meyer, Finn, Eyde, Kay, Moreland, Dies, Eisman, Kubiszyn, and Reed 2001, for an interesting, and eye-opening, review of medical research findings). Most of the causal correlations between medical treatments and successful cures are below $r = .30$ (an r of .30 means a binary 60% chance of success—see Appendix C). Treatments for obesity, for instance, are pessimistic indeed: only 28% success for surgery, 11% for lifestyle modification programs, and 8% for drugs (Creswell 2010).

Interestingly, and not entirely unrelatedly, the average correlation between ATTITUDE and BEHAVIOR (toward the same OBJECT) is the same as the computer’s answer in *Hitchhiker’s Guide to the Galaxy*, namely “42,” or correlationally speaking $r = .42$ (see Kraus 1995, for a meta-analysis that arrived spookily close to this number and see Rossiter and Percy 1997, p. 271, for the qualification of his overall average of $r = .38$ that makes it $r = .42$). All BEHAVIORS have multiple causes and ATTITUDE is just one of them.

So forget about touting very high correlations as “evidence” of a predictor measure’s validity. If the observed PV is greater than $r = .5$, you should be suspicious about the circularity of the predictor and criterion constructs or about measure distortion (D_{m} in the new true-score model of Chapter 1) in both measures causing spurious inflation. The *sole exception* is GENERAL INTELLIGENCE, also known as GENERAL MENTAL ABILITY—measured as I.Q.—which is the most powerful predictor in all of the social sciences (see Table 2.1) and frequently produces cause–effect correlations that are greater than .5.

Most researchers don't realize that predictive validity is *not* a matter of trying to *maximize* the correlation between scores on the predictor measure and scores on the criterion measure, but rather to come as close as possible to the *estimated population correlation* (R_{pop}) between the two constructs (actually, between the *scores* obtained from content-valid measures of those constructs). For examples of how to estimate R_{pop} from meta-analyses, see Ouellette and Wood (1998) and Rossiter and Bergkvist (2009) but be *wary* of meta-analyses because they include studies with low content-valid measures. Some important R_{pop} estimates are given in Table 2.1 from a compilation by Follman (1984). Which do you think might be *causal* correlations? This is not an easy question!

Table 2.1 Some interesting R_{pop} estimates (from Follman 1984)

Predictor	Criterion	R_{pop}
I.Q. at age 6 or 7	Grade 1 school achievement	.88
I.Q. at end of high school	College (university) achievement	.53
Own I.Q.	Spouse's I.Q.	.50
Own I.Q.	Children's I.Q.	.50
Physical appearance	Spouse's physical appearance	.40
I.Q.	Creativity	.35 (much higher below I.Q. 120 and much *lower* above 120)

If no appropriate meta-analysis (or large-scale representative study) is available, as would be the situation for a new construct and, therefore, a new measure—which could be a measure of either a predictor variable or a criterion variable (or both in a sequential theory)—then the researcher still has to *make* an estimate of R_{pop} and justify it. The researcher cannot simply claim that the *highest* observed correlation between the measures is the *true* correlation, which is a thoughtless empirical decision rule invoked so widely in the social sciences.

So-called *nomological* validity (Bagozzi 1994) is simply another instance of *predictive* validity. In nomological validation, a measure is evaluated by the size of its correlations with antecedent and consequent variables in a "theoretical network." However, the network should use estimates of R_{pop}, which in the case of multiple determinants will be *partial* $R_{\text{pop}}s$, controlling for the effects of other determinant variables. Without these *true* R_{pop} or partial R_{pop} estimates as guides, nomological validity interpreted on the *observed* correlations (or "fit statistics") is meaningless. It becomes in effect just another aimless application of the convergent validity principle of MTMM, which I have said is logically worthless.

In sum, a measure *must* be argued to be either highly or preferably fully content valid (CV) and this is *sufficient* because the validity of a measure must be established in its own right and not by the relationships of its scores with other measures' scores. Then it is *desirable* for the measure to also predict well (PV) within reason (within the approximate 95% confidence interval of R_{pop}), or to be "on the end" of a reasonably accurate *causal* prediction if the measure is a criterion measure. R_{pop} is sometimes written in statistics textbooks as R_{XY}, where X is the predictor construct and Y the criterion construct, but R_{pop}—"pop!"—more dramatically expresses the importance of chasing down or making this estimate so as to properly interpret the predictive validity of a predictor measure.

2.6 Why Coefficient Alpha Is Wrong

Coefficient alpha (Cronbach 1951) is, without doubt, the main statistic used by psychometricians to justify multiple-item measures. It is thought to indicate "reliability," and many researchers report coefficient alpha as an implied claim of *validity* for the measure—see, for instance, Robinson, Shaver, and Wrightsman's *Measures of Personality and Social Psychological Attitudes* book and especially Bearden and Netemeyer's *Handbook of Marketing Scales*.

Ironically enough, I was possibly responsible for introducing coefficient alpha into marketing in an early and well-cited article in which I developed a multiple-item measure of CHILDREN'S ATTITUDES TOWARD TV ADVERTISING (Rossiter 1977). This was the topic of my Ph.D. thesis at the University of Pennsylvania back in 1974, supervised by a great guy and avid cognitive psychologist, Larry Gross, at the Annenberg School of Communications, and mentored by another great guy, Tom Robertson, now Dean of Wharton, where I was fortunate to get my first academic appointment.

However, I have since changed my opinion about alpha—twice. In my first article on C-OAR-SE (Rossiter 2002a) I recommended using coefficient alpha (preceded by Revelle's (1979), coefficient beta, which only my Australian colleague, Geoff Soutar, has picked up on and used) for *one* of the six cells of scale types in the 2002 version of C-OAR-SE: when there is a "concrete" object and an "abstract eliciting" attribute. The construct of CHILDREN'S ATTITUDES TOWARD TV ADVERTISING does *not* fit this cell (in hindsight, it is obvious to me now that the attitudes *form* children's *overall* attitude toward TV ads, so alpha does not apply). But in 1977, I had yet to invent C-OAR-SE!

Now—in this book—I have changed my opinion about alpha again, this time much more radically, scuttling even the limited role I ascribed to alpha in the 2002 version of C-OAR-SE. I thought hard about my central proposition in C-OAR-SE: that *content validity* is the only essential requirement of a measure (C → M in the Construct → Measure → Score model of Chapter 1). Coefficient alpha, or α, is a measure of the "internal consistency" of *scores*, S, on a multiple-item measure of a construct. Alpha, therefore, falls into the same logical trap that all of psychometrics falls into. This is the trap of assuming that you can validate a measure of a construct by examining the *scores* obtained with the measure—that is, by a backward S → M inference according to my C → M → S model. So, forget coefficient alpha. It signifies nothing about the validity of the measure.

Nor does coefficient alpha indicate “reliability” in any useful sense of the term—contrary to prevailing psychometric theory. It is not the “savior statistic” that everyone thinks it is and even its inventor, Lee Cronbach, later abandoned it!

There are only two meaningful (and useful) interpretations of *reliability*: stability-of-scores reliability, $R_{\text{stability}}$, and precision-of-scores reliability, $R_{\text{precision}}$. The concepts of $R_{\text{stability}}$ and $R_{\text{precision}}$ are defined and discussed in the next and final sections of this chapter.

2.7 Stability-of-Scores Reliability ($R_{\text{stability}}$)

Highly content-valid measures should produce stable scores on a short-interval retest. This is “test-retest” reliability, which I dismissed in the original C-OAR-SE article (Rossiter 2002a) as uninformative because a very poor measure (with low or even zero content validity) could produce highly repeatable (stable) scores. This was pointed out in Nunnally’s (1967, 1978) classic textbook on psychometric theory. An interesting example of very high stability with very low *predictive* validity is one’s astrological STAR SIGN (mine is Aries), which is regarded by many as a good measure of, and even a determinant of, one’s “PERSONALITY” (which is the constellation, in an apt metaphor, of one’s PERSONALITY TRAITS). STAR SIGN is not a zero predictor of PERSONALITY as most scientists believe: it is a very weak but statistically significant predictor (see the review by Dean, Nias, and French 1997), and it is of course 100% stable, and over an infinite interval. My Australian Aries birth symbol is the Crocodile, which happens to be the focal symbol on the Rossiter family coat-of-arms. I believe that the Aries Crocodile personality profile, which I came across only a year ago, fits me well and I believe those who know me well would agree, and would especially agree with the “argumentative” trait! My egotistical self can’t resist including this profile—which, like all of them, errs on the flattering side

> Crocodile people are natural born leaders, charming, intelligent and strong-willed. They court success, are assertive and quick-witted. Being independent and competitive by nature, when challenged they can become argumentative and impatient and may need to practice seeking peaceful outcomes by negotiating. They are self-confident, dynamic, passionate, and big-hearted.

Of course, there are many Aries who *don’t* have all these traits, hence the low predictive validity of STAR SIGN despite perfect stability.

What I had failed to acknowledge in the 2002 article was the “reverse” case. That is, a measure *cannot* be a good predictor unless it produces highly *stable* scores on a short-interval retest (“short interval” means 1–2 weeks—what Cattell, Eber, and Tastuoka (1970), in their *Handbook for the 16PF*, p. 30, identified as “the lapse of time . . . insufficient for people themselves to change with respect to what is being measured”). This is (now, to me) logically obvious: if a measure produces different scores at the individual-rater level each time it is used, it can hardly be recommended as a predictor measure!

The insight that stability of scores is due to—in fact, is an essential property of—the *measure* came later, during new research I was doing, and am continuing to do, with Sara Dolnicar, my excellent, and I hope almost converted, psychometrician colleague (her work has been published in the journal, *Psychometrika*) at the Marketing Research Innovation Centre in the Institute for Business and Social Research at the University of Wollongong. Sara and I (in an article in review as I write, together with expert statistician Bettina Grün) found that rating measures commonly used in the social sciences, such as "Semantic Differential" measures and "Likert" measures, often produce too-low stability of scores ($R_{\text{stability}}$). We believe, and have stated and tested in the forthcoming article, that this is mainly due to the measures' differing *answer-scale validity* ($\text{CV}_{\text{answer}}$). If the answer mode (words or numbers) and answer alternatives (polarity and number of scale points) do not match the main alternative answers that the *rater* has in mind, then this property of the *measure* will lead to individually inconsistent—low stability—scores.

Stability-of-scores reliability ($R_{\text{stability}}$), therefore, *does* say something about the *predictive* validity of the measure. Just as "necessity is the mother of invention," stability is the "mother" of *predictive validity*. Empirical proof of this oracular-sounding pronouncement—for those who demand empirical proof beyond plain logic—is given in my hopefully forthcoming article with Sara and Bettina, which examines the stability of measures of BELIEFS predicting OVERALL ATTITUDE. Also, in Chapter 8 in this book, where it will be seen that predictive validity is the only way to validate *qualitative research* measurement, I point out that qualitative research conclusions must be *stably inferred* by the qualitative researcher before they are put forward *as* conclusions.

Psychologists should note that what I am calling "stability" is what Cattell (Cattell et al. 1970) called "dependability." I don't use his term—other than in this chapter's opening quotation—because I think it could ambiguously refer also to the second type of reliability, discussed next.

2.8 Precision-of-Scores Reliability ($R_{\text{precision}}$)

Precision-of-scores reliability ($R_{\text{precision}}$) is a statistic that is important to report for each *use* of the measure so that users can see how *accurate* an absolute estimate is (e.g., a percentage or proportion) or an average estimate is (e.g., a mean or median). High accuracy, or good precision, however, doesn't mean that the measure is *valid*. In the major ongoing debate about "climate change," for instance, there has been little or no publicity questioning the content validity of the measures that go into the projections. The projections themselves are assumed to be *precise* simply because they are based on "computer modeling." The public is being misled on both counts. While I realize that policymakers have to consider the "worst case scenario," possible low-validity measures and definite low precision due to a small sample of recent large changes in climate make the "worst case" speculative in the extreme. The first and necessary step toward resolution is to put pressure on climate scientists to justify that their measures have very *high content validity*. The next necessary step is to import cause-detecting methods of analysis beyond correlation (see West, Duan, Pequegnat, Galst, and others 2008). Only then will *precision* become relevant.

As implied by the acronym "C-OAR-SE," I am against overrefinement in the reporting of scores because the scores are usually based on measures that are less than highly content-valid. The modern procedure of estimating precision by computing the "confidence interval" around the absolute or average score by first calculating the *sample* standard deviation or standard error (see most statistics texts) is an example of overrefinement. I don't know about you, but I find it almost impossible to decipher the results presented in this fashion, and so too the "odds ratios" and their confidence intervals that have crept into health-science reporting.

Sufficient accuracy for users (especially managers) to make decisions, I suggest, is given by simple "look-up" tables that base the standard error majorly on *sample size*, commonly symbolized by N, or by n_1 and n_2 in the case of a comparison between samples, and minorly on an average standard deviation computed over thousands of surveys. It is obvious that the larger the sample(s)—assuming random selection or at least, practically speaking, "representative sampling"—the more accurate (precise) any absolute or average estimated score will be—though with diminishing returns since it is the *square root* of sample size(s) that matters.

How much does sample size matter? In Tables B.1 and B.2 in the Appendix B, I have reproduced two useful look-up tables from my advertising textbooks (Rossiter and Percy 1987, 1997, Rossiter and Bellman 2005), acknowledging their source from the (United States) Newspaper Advertising Bureau. The first is for estimating the accuracy of a single average score and the second for estimating the difference between two average scores needed to be *reasonably* confident that they are in reality different (e.g., the superiority of a new ad over the previous ad for a brand, a difference that I have many times had to put my scientific reputation on the line for as an advertising research consultant). Table B.1 is widely used by the better U.S. newspapers and now by some European and Australian newspapers in conjunction with the reporting of public opinion survey results. Table B.2 is mainly useful for managers when evaluating whether to change strategy in any area of business (or politics or public health). It is also useful for preventing people (and politicians) from becoming too excited about small percentage differences even when they come from quite large samples in surveys!

To disclose a personal anecdote about their usefulness, these tables spared me from a lawsuit threatened by a major advertising research company. This company implied that ads' scores on one of their measures were exact (i.e., perfect $R_{\text{precision}}$). I claimed the scores could not be exact because they were based on sample sizes of only 100, and, therefore, could be as much as 6 percentage points lower or 6 points higher if the study were repeated (roughly in 5% of the repeats, they could be expected to be even more deviant because of this 95% confidence interval). I showed the company Table B.1 and the would-be litigants backed off. But still in their literature and to clients they imply the unjustifiable precision. I had neither the time nor resources to fight further but I did make sure *my* clients—some very large advertisers—"knew the score," so to speak.

Be warned, however, as eminent management guru Peter Drucker observed in several of his books: it is more important to be vaguely right (with a highly content-valid measure) than precisely wrong (with a low content-valid measure). Only with a large and representative sample and a highly content-valid measure can you be *precisely right*.

2.9 End-of-Chapter Questions

(2.1) What is "construct validity?" Explain it in terms of the C → M → S structure of measurement model in the chapter, without "parroting" what I've written. (5 points)

(2.2) Write out a logical argument, as much as possible in your own words, against (a) convergent validity and (b) divergent or discriminant validity, which together constitute the multitrait-multimethod (MTMM) approach. You won't find any help in the original source or in research textbooks! (4 points maximum for each and 2 bonus points for adding (c), a convincing summary of what's wrong with MTMM)

(2.3) Of what does content validity (CV) consist; how does it differ from "face validity;" and why is it the only essential type of validity? (7 points)

(2.4) Discuss, as much as possible in your own words and in no more than about 500 of them, the importance, desirability, and nature of predictive validity (PV), defining it first. (5 points) Advanced second question: Look up the study by Quińones, Ford, and Teachout in *Personnel Psychology*, 1995, 48(4), 887–910, in which these researchers estimated the population correlation for WORK EXPERIENCE predicting JOB PERFORMANCE as $R_{\text{pop}} = .27$. Write a detailed critique of their estimate and explain, from their meta-analysis studies, what your revised—if any—R_{pop} estimate would be. (10 points)

(2.5) What is "reliability" and what should it be, according to C-OAR-SE theory? Include a critique of coefficient alpha reliability (and a defense if you're up to it) and clearly explain the two useful types of reliability, $R_{\text{stability}}$ and $R_{\text{precision}}$. (7 points, with a maximum of 3 points for a convincing epitaph for alpha and 2 points for each of the answers about the two useful types of reliability)

Chapter 3
Object Classification and Measures

UFO: acronym for unidentified flying object
—Oxford Dictionary

Applied to measurement theory, "UFO," in Aussie Bluetongue beer-speak could well-stand for "unidentified *friggin*' object" because this frequent measurement mistake is so very annoying. Half the problem with measures in the social sciences is due to misrepresentation of the *object*. The object of the construct—the element to be rated or otherwise evaluated—is carelessly represented in nearly every measure. For example, a common measure-distorting mistake (D_{m} in the new true-score model of Chapter 1) in marketing research is to represent consumer products in the measure by their verbal brand names when the actual object of choice is the physical product in its visual (and sometimes tactile) brand package as it is presented in the store. Non content-valid object representation in the measure can reduce the validity of the whole measure drastically and lead to grossly misleading scores. Another example of low content-valid object representation in the other social sciences is the important set of abstract objects called VALUES. These are often poorly represented (e.g., in Rokeach's 1968 Values Survey and in Kahle's 1983 List of Values items) and the result is almost laughably superficial and erroneous readings of people' or an organization's values. Any qualitatively oriented psychologist would tell you this (see, e.g., my 2007b article criticizing the Australian government's definition of values and giving my nominations of real "Aussie" values).

After reading this chapter, you should be able to:

- See why correct object representation in a measure is vital for the measure's validity
- Understand the three classifications of objects and how each should be represented in the measure
- For multi-constituent or multi-component objects, realize that a formed index is required, just as for multi-component attributes

DOI 10.1007/978-1-4419-7158-6_3, © Springer Science+Business Media, LLC 2011

3.1 The Importance of Valid Object Representation in the Measure

Neglect of object representation—the "UFO error"—is largely the result of social scientists' myopia in conceptualizing constructs entirely in terms of the attribute. The essential elements of a construct are *threefold*: the object, the attribute, and the rater entity. I could find only two articles in addition to my own 2002 C-OAR-SE article that seriously considered object representation. These were two recent articles in the same volume of the *Annual Review of Psychology* (by Martin 2007, and Peissig and Tarr 2007), which focused on brain representations of objects—without considering the obvious implications for measures.

The idea of focusing on the object representation as well as attribute representation came from my background in S–R psychology (and was reinforced in a 1989 article by the late, great psychologist William McGuire). As psychology students, we were always taught to carefully analyze and describe the *stimulus* (the object) before evaluating the response (on whatever attribute was of interest). I have applied this lesson consistently over the years (for instance, my explanation of advertising's communication effects is in terms of S–R theory or, actually, S–O–R mediation theory; see the Rossiter and Percy 1987 and 1997 advertising textbooks and also the Hull-Spence theory of consumer behavior proposed by Rossiter and Foxall 2008). Stimulus focus is similarly vital in designing—or evaluating—measures.

There are two serious consequences of failure to focus on the stimulus object.

The first consequence is that the observed scores will represent beliefs about, or an attitude toward, the *wrong object*. And note that this is a measure distortion error, D_{m}, not a rater error, E_{r} (see Chapter 1's new true-score model). A good example of this would be rating BMW (the car brand) as an object. The BMW name (and logo) carries very high prestige among car drivers worldwide. However, the BMW company, like many prestige car manufacturers nowadays, manufactures a number of "lower end" models that are indistinguishable from everyone else's cars. Take a close look at BMW's website for its lower-numbered model series and compare the pictures with those on Ford's or Toyota's website and you'll see what I mean. If these lower-end models were rated from the pictures alone, without their names and BMW logo visible, I am sure the researcher would observe much lower ratings of the car's attributes (beliefs or perceptions) and a much lower overall evaluation (attitude) than if they were identified as BMWs. My suspicion is verified by the fact that BMW, in its ads, for whatever particular model, always features its blue and white, black-circled logo prominently. And it's also why the logo is so easy to spot on its cars on the street.

A clinical example of the necessity of defining the stimulus object of the construct emerged in a study of the attribute of DISGUST by Tybur, Lieberman, and Griskevicius (2009). By varying the classes of objects rated, these researchers found that there are three different types of disgust: PATHOGEN DISGUST, SEXUAL DISGUST, and MORAL DISGUST. That these are three different *constructs* was suggested by their different correlations—zero, small-negative, and moderate-negative, respectively—with SELF-REPORTED PRIMARY PSYCHOPATHY (behaviors that represent a lack of concern for others' welfare and willingness to lie and cheat—the latter tendency disturbingly characteristic of students today, according to some newspaper reports, as well as of young and not so young academics, as I've heard anecdotally and seen as a reviewer). Getting back to the study, I would be interested to learn whether these are actually three different *attributes* representing complex composites of specific avoidance responses due to different aversive feelings or whether it is only the *objects* that differ.

The other consequence of failure to focus on the stimulus object is that, when measuring beliefs, perceptions, or associations, failure to specify and clearly represent the *object* of the construct often means that the *wrong attributes* will be selected. A glaring example of this in the marketing literature is the SERVQUAL measure (Parasuraman et al. 1988). The component attributes of "service quality" depend entirely on the service quality *object*. Parasuraman et al. selected attributes that apply to COMMERCIAL RETAIL ESTABLISHMENTS such as insurance companies and banks (and even here wrongly pared down the set of attributes by using factor analysis, as I pointed out earlier, and see Armstrong and Soelberg's devastating critique of factor analysis published in the *Psychological Bulletin* in 1968). Smith (1999) tried to apply SERVQUAL's attributes to a very different object category—HEALTHCARE ESTABLISHMENTS (hospitals, in her study)—and found that SERVQUAL's attributes missed many essential attributes that would apply only to healthcare-provider objects. Apparently oblivious to Smith's problem, researchers Etgar and Fuchs (2009) used the SERVQUAL attributes to attempt to measure the consumer-perceived service quality of SPECIALIST PHYSICIANS. Not surprisingly, they found that hardly any of SERVQUAL's five attribute "dimensions" predicted important outcomes such as SATISFACTION, BEHAVIORAL LOYALTY, or PATIENTS' RECOMMENDATION OF THE PHYSICIAN. Thus, as I said, SERVQUAL should have been defined by its originators as a measure of the construct of THE SERVICE QUALITY OF COMMERCIAL RETAIL ESTABLISHMENTS AS RATED BY (E.G.) POTENTIAL CUSTOMERS. The object must be defined and clearly represented in the measure, otherwise the attributes in the measure will be wrong.

C-OAR-SE theory postulates that there are three classifications or types of object: a concrete object (CONCOB), an abstract collective object (COLLOB), or an abstract formed object (FORMOB). These object classifications are defined and exemplified in the remaining sections of the chapter and summarized in Table 3.1.

Table 3.1 Object classification definitions

Classification	Explanation
Concrete object (CONCOB)	• Unambiguous single object
Abstract collective object (COLLOB)	• Collection of constituent concrete objects
Abstract formed Object (FORMOB)	• Composite of the main meanings (components) of an ambiguous object • Each component must be concrete

3.2 Concrete Object (CONCOB): Iconic Representation

The first category of classification of the object is a *concrete object*, abbreviated CONCOB, and meaning that the object has only one meaning which is clear to all raters. In constructs where the object is a CONCOB, the object must be *iconically represented* in the measure. This means, for example, that if you want to measure consumers' liking of the famous BENETTON "NEWBORN BABY" AD (see Chapter 2), then you have to include a full color reproduction of the ad in the measure. "Concrete" means unambiguous—as opposed to abstract and ambiguous—so the real stimulus, or a faithful reproduction of it, must be made the first part of the measure to which the rater is asked to respond.

I will give two further examples, which illustrate how the incorrect representation of a CONCOB can produce a serious content-validity mistake, with the low-validity measure then producing erroneous scores. The first example concerns the relatively new soft drink beverage, Coke Zero, which now has a larger market share than the virtually identical earlier product, Pepsi Max. When Coke Zero was first advertised, consumers were encouraged by the ads to form beliefs between the *verbal* stimulus "Coke Zero" and relevant verbal attributes—more precisely, *levels* of attributes—such as "no calories" or "no sugar" or "no both." Belief formation was made easier by the clever choice of the name for the new product, Coke *Zero*. Contrast the situation for Pepsi Max, where the "Max" in the name refers to "maximum taste" (if you listen to the commercials closely). For Pepsi Max, consumers were required to learn a more complicated message, namely, that its "maximum taste" was achieved with a zero amount of sugar, and thus near-zero calories (in the metric units of kilojoules used in Australia, Pepsi Max has 1.2 kJ per 100 ml container versus 1.4 kJ for Coke Zero and a whopping—if you're a "calorie counter"—180 kJ for regular Coke). The objects for these BELIEFS are correctly—iconically—represented by the *verbal* stimuli, "Coke Zero" and "Pepsi Max." But what about at the point of purchase, where brand ATTITUDE, rather than a recalled brand-attribute BELIEF, is more likely to be the cause of PURCHASE and especially of REPEAT PURCHASE. This time, the CONCOB is the *visual* brand package: the Coke Zero can or bottle, if in a vending machine or freezer case, or the Coke Zero six-pack wrapper or bottle label, if in a supermarket display. The object in this situation is an imposing, predominantly black, visual stimulus with semantic properties quite unlike those of the verbal stimulus "Coke Zero," and in fact there's a *different* verbal stimulus on the label, specifically the full company name *Coca-Cola* (in the company's familiar, patented script) with the word "Zero" in lower-case, and on the can these two words are written vertically. A fully content-valid measure of ATTITUDE TOWARD COKE ZERO to fit the point-of-purchase choice situation could, therefore, only be made by presenting the consumer with a visual icon of the actual object encountered at the point of purchase.

The other example comes from the Nestlé company in Australia, which several years ago reformulated its package labels for its "gourmet" line of instant coffees, such as my favorite instant brew, Nescafé Alta Rica. BRAND ATTITUDE measured toward the *verbal* stimulus, "Nescafé Alta Rica," would remain as before. But, with the new label, this measure is no longer fully valid. As companies such as Nestlé have found, sales often decline, often for a considerable period during which consumers may switch to another brand, because regular buyers fail to recognize the relabeled product *or* because, for regular buyers and for new buyers who find the product, the attitude formed—on the spot—toward the *new visual stimulus* may be less positive than it was toward the old visual stimulus.

With SERVICES, in contrast with products, the use of a *verbal* brand stimulus is usually iconically correct, because the service provider's *name* has to be verbally recalled. Examples are "Fedex" (an abbreviation "borrowed" from the U.S. stock exchange) and "DHL," if you need a courier service. However, a verbal stimulus is *not* iconic for services that are chosen by many consumers on a "walk-in" basis—by *visually* recognizing the retail service *logo*. An everyday example would be when looking for a specific bank's free-standing ATM (card-operated cash dispenser) in a shopping mall.

Failure to represent the concrete object (CONCOB) iconically and thus in the right *modality* in the measure is a content-validity error made also by psychologists, who, I am happy to notice, are increasingly using marketing (consumer behavior) applications in testing their theories. A recent example of the object representation error described above appeared in a very good journal, the APA Division 23's *Journal of Consumer Psychology* (Hung and Wyer 2008) and so the scores and findings from that study can't be trusted. Measure-designers in all the social sciences need to understand the iconic representation principle when measuring constructs that incorporate a CONCOB.

3.3 Abstract Collective Object (COLLOB): Representative Sample of Constituents

The second category of classification of the object is an *abstract collective object*, abbreviated COLLOB. In constructs where the object is a COLLOB, the principle is to use a representative sample of *constituent* objects. A topical example in organizational behavior theory would be testing the hypothesis that UNIVERSITIES are becoming more ENTREPRENEURIAL. (This is a hypothesis that most academics of my generation know to be sadly true. Indeed, as I write, there is a serious proposal on the table in Britain to ban Ph.D. research at 80 of the country's 120 universities because they are "too small" or "too low quality" in research training; see Hurst 2009. But shouldn't the quality of doctoral research training be judged in the marketplace of journal publications? Is a university without doctoral students a UNIVERSITY?) The abstract collective object of the construct is UNIVERSITIES, the constituent objects of which are *particular* universities. To safely conclude anything about whether universities nowadays are "entrepreneurial" requires that a representative sample of universities be included in the measure.

Also note that even though this classification of the object is called *abstract* collective, the constituent objects are *concrete*. Items in a measure have to have a concrete object (and also a concrete component attribute, as we'll see later: items have to be "doubly concrete"). An example of a COLLOB widely studied in psychology, and in marketing, is FEAR APPEALS. In most studies of the effectiveness of FEAR APPEALS, this abstract collective object is hopelessly variously represented, making conclusions very hard to draw. The concrete fear-appeal stimuli represented in the experiments vary in their capacity from eliciting no fear to mild fear—often labeled ambitiously by the researcher as "high fear." But how much can you scare people with a print ad, for example, or with a short video, the typical stimuli in these experiments? Furthermore, researchers rarely properly classify the fear-appeal stimuli as being either sudden-fear "shock" appeals, "rising-fear" appeals (with these two forms of fear appeal offering no relief), or "fear-relief" appeals, which also vary in the degree of fear-reduction that they offer, from partial to total (for further details of this classification of fear appeals, see Rossiter and Thornton 2004).

Various selections of *constituents* of COLLOBs completely change the observed responses. This prompts a "mixed" conclusion to be drawn by researchers in their literature reviews (or meta-analyses) and "mixed" conclusions are a scientifically unacceptable state of affairs.

As a reviewer for a number of leading social science journals, as well as several lesser journals that tend to compensate by having more interesting topics, one of the most frequent problems I encounter is the testing of theories or hypotheses about abstract collective objects represented by only a single constituent object or perhaps just two constituent objects (most often these are companies, stores, or ads). If I believe that the object or objects used in the survey or experiment are *reasonably typical* of the abstract object class—and that there is not much important *variation* among objects in this class—then I would usually point this out in the review and conclude that the test results can be quite safely generalized. Much more frequently, however, my recommendation to the researcher is to go back to the laboratory and test the theories or hypotheses on a more representative sample of constituent objects. I can't help noticing how many tenuous theories make it into the literature as "proven" when the researchers have misrepresented or underrepresented a COLLOB in their survey or experiment.

On the other hand, what is not often realized by researchers—and reviewers—is that it is legitimate to *disprove* a theory or hypothesis by measuring just *one* constituent object of a class of collective objects, no matter how atypical of the class it is. This is because disproof requires only one failure of the theory or hypothesis (K.R. Popper's "falsification principle"). Witness the famous—in philosophy—"problem of induction," where, for instance, we only need to find one black swan to disprove the hypothesis that "All swans are white" (the philosophers evidently weren't born in Western Australia, the Australian state where I was born, where black swans not only exist but dominate; as a kid, I held the inductive hypothesis that "All swans are *black*"). Experiments or surveys are usually set up with the purpose of *proving* a given theory or hypothesis, not disproving it, but it is legitimate to argue *ex post* for *disproof* based on a sample consisting of just one constituent object. But you'd better be sure that the rest of the measure (the attribute part) is content-valid.

It is extremely easy to "lose" representativeness of the abstract collective object in *multiple-item* measures. This is because researchers do not scrutinize the question part of their items for content-valid object representation. For a criticism of this in a multiple-item organizational behavior measure, see my article in the *British Journal of Management* (Rossiter 2008a). This example is so typical of the "UFO error" that I discuss it in more detail later in this chapter.

A frequent special case of an abstract collective object is THE SELF. (The SELF can be a rater entity—the INDIVIDUAL—and also the object of the construct, as in, e.g., SELF-RATED PERSONALITY TRAITS.) In most theories about the "self concept" in psychology, and also in consumer behavior, it matters greatly whether the rater (and the researcher) is referring to the Actual Self (the "real me"), the Ideal Self (me as I would "like to be"), or the Social Self (me as I believe I am seen by "others"), as identified by one of the pioneering psychologists, James (1892). These distinctions are very important because they have different projections on attributes, such as PERSONALITY TRAITS. In fact, it is virtually meaningless to consider THE SELF as an abstract collective object in the first place (that the three types of SELF are "constituents" of an abstract collective object). Rather, ACTUAL SELF, IDEAL SELF, and SOCIAL SELF are best conceptualized as separate concrete objects—that is, as CONCOBs. They should be identified as such in the *construct definition*.

3.4 Abstract Formed Object (FORMOB): Set of Main Meanings

The third and final classification of objects in C-OAR-SE theory is an *abstract formed object*, abbreviated FORMOB. This is conceptually the most complex type of object. A FORMOB has *components*, not constituents, and these are the set of *main meanings* of the object. These main meanings must be *concrete*.

One of the most important abstract formed objects in the social sciences is a VALUE, which may be defined more clearly as an enduring GOAL in one's life (see below), either to work toward in one's daily conduct (a positive value) or to avoid (a negative value). Note that the VALUE is the *object* in this construct; positive–negative EVALUATION is the *attribute*; and THE INDIVIDUAL ACTING AS THE ACTUAL SELF is the *rater entity*. VALUES are prominent constructs in cross-cultural studies (CULTURAL VALUES), organizational behavior studies (ORGANIZATIONAL VALUES or what is often called "CORPORATE CULTURE"), social psychology (HUMAN VALUES, made famous by Rokeach's 1968 and 1973 books), and marketing (CONSUMER VALUES).

A VALUE is an *abstract object* in that it obviously has various meanings. The main meanings must be identified by the *researcher*, since this is a psychological not a perceptual construct (see Chapter 1 and Chapter 2), and the main meanings must be included, as object components, in the *measure*. Scores on these main meanings then *form* the overall object score. The famous Rokeach Value Survey items (Rokeach 1968) include serious object mistakes. In an attempt to clarify the meaning(s) of his theorized values, Rokeach added "concrete" explanations of them in parentheses. Sample items from his set of end-state or "terminal" values are A COMFORTABLE LIFE (i.e., A PROSPEROUS LIFE), two human states which I would argue are *not* synonymous; MATURE LOVE (i.e., SEXUAL AND SPIRITUAL INTIMACY), which to my mind doesn't explain this value at all; and WISDOM (i.e., A MATURE UNDERSTANDING OF LIFE), in which the explanation is as vague as the value. Alarmingly, in one study reported in Rokeach (1973), he found that the rankings of the 18 values were unaffected by the omission of these "explanations" (which should have been separate items, by the way), a finding that suggests they were *not* the main alternative meanings of the values. Rokeach's VALUES remain (very) abstract objects and I suggest they produce largely nonsense rankings (or ratings).

What researchers should do to define VALUES is employ qualitative research conducted by a skilled psychological analyst (see Chapter 8) asking people to give open-ended answers describing the GOALS that they strive for in everyday life. They should not be asked what "values" they hold, which is the procedure used in all values surveys, because a VALUE is far too abstract an object. An up-to-date *thesaurus* (dictionary of synonyms) is a great, and greatly overlooked, aid for researchers when choosing the components of a FORMOB because it lists the main everyday meanings of abstract nouns. (This idea occurred to me because, being a crossword addict like my partner, Mary, I dip into a thesaurus regularly, e.g., Collins 2002, or Harringman 1990.) The *thesaurii* reveal that the common synonyms for the noun VALUE—Account, Cost, Quality, Utility, Worth—do not represent the meaning of the object as it is intended in social science theory. On the other hand, the common synonyms for the noun GOAL are obviously suitable as concrete, everyday language meanings of the object, VALUES, in the measure—Aim, Ambition, Objective, Purpose. It is obviously valid to ask people to describe their aims in life, their ambitions, their objectives, and what they see as their purpose in life. Their answers will also inevitably reveal the "goals" that they *don't* want to end up "scoring" in life. In other words, this concrete questioning about GOALS will reveal their positively evaluated *and* negatively evaluated VALUES. Of course, the researcher must ultimately categorize and label the VALUES (preferably aided by a couple of other expert content-analysts—a procedure that should be but is rarely followed by psychometrics researchers when labeling "factors" in factor analysis, I note, but factor analysis, as I pointed out earlier, is not a valid or validating procedure). I guarantee the researcher would come up with a much more valid set of VALUES—and also the *defining components* of each VALUE—than those in the standard measures (invented by Rokeach, Schwartz, Hofstede and others). These concrete components could then be included as the items in a new VALUES questionnaire.

There is a giant opportunity with this *qualitative* approach, by the way, to validly represent CONFUCIAN VALUES, which are completely missed in questionnaires that ask about purported WESTERN VALUES (see Chan and Rossiter 1998). Most interesting in Chan and Rossiter's study was that Chinese students born during or slightly after Mao's Cultural Revolution (which attempted to replace Confucian values with Communist values) gave exactly the same importance ratings, overall, to *Confucian* values as did English-background Australian students. This suggests, first, that Communist values did not take hold among the young Chinese and, second, that allegedly unique Confucian values such as HARMONY, RECIPROCATION, and FILIAL PIETY may be more universal than sociologists believe. Either the latter or the measures aren't "deep enough" and thus are not highly content-valid, which I strongly suspect may be the case.

While I'm harping on about conventional measures (and definitions) of VALUES, which don't adequately represent the objects (FORMOBs), I additionally wish to point out that the Rokeach measure and also the more recent Schwartz measure (e.g., Schwartz 1992) *must* have low content validity. VALUES are supposed to be "enduring," but survey results using Rokeach-type or the similar Schwartz-type measures show they are evidently not. Specific VALUES are far from 100% consistently ranked, rated—or even pair-comparisoned—on a brief-interval (2 weeks later) retest; see the retest studies by Munson and McIntyre (1979) and Reynolds and Jolly (1980). Yet, researchers continue to use these shaky measures in the best social science journals (e.g., Bardi, Lee, Hoffmann-Towfigh, and Soutar 2009) as though they were capable of measuring VALUE true-scores!

There are two "shortcut" methods of representing abstract formed objects that should be mentioned in order to dismiss them. One method is to scatter object components across items and the other is the opposite error, which is to jam all the components into a single item.

Scattering error. In an organizational behavior study, Diamantopolous and Sigauw (2006) attempted to represent the abstract formed object, THE EXPORT COMPANY, by scattering its component objects, which were different DEPARTMENTS of the company (Export Department, R&D Department, Manufacturing Department, Finance Department, Sales Department) across items in the questionnaire. The attribute in the items (and in the construct) was COORDINATION between departments, an abstract attribute that was itself represented by concrete components (which were examples of "coordinated" behaviors). The researchers ultimately wanted to measure COORDINATION between all possible pairs of DEPARTMENTS but their items were worded variously to include arbitrary pairs of departments (e.g., Export, R&D), misnamed combinations (e.g., Marketing/Sales; Finance/Accounting), or sometimes no pairs at all (e.g., the vague reference to Other Functional Areas). This arbitrary, unsystematic representation of object components must result in a measure with too-low validity (see the detailed critique in Rossiter 2008a). Object components must be fully and equally represented across the items if a valid measure of a FORMOB is to be made.

Jamming error. The opposite type of measure error when measuring an abstract formed object is to jam all the components into the question part of a single item. A construct frequently the focus of organizational behavior and management research is SELF-RATED JOB SATISFACTION. (The object is one's JOB, the attribute is SATISFACTION, and the rater entity is the INDIVIDUAL.) Researchers Gardner, Cummings, Dunham, and Pierce (1998) thought OVERALL JOB too complex an object for employees to rate and attempted to include the main object components in a single question, as follows: "Job factors. These are the things directly related to the work you perform. This includes all of your job duties, the challenge of your job, the activities you engage in, the decisions you make as part of your job, and your job responsibilities" (p. 901). Wow! The researcher will get an answer (a score or rating on SATISFACTION) but what can it possibly mean? If the researchers had wanted to measure SATISFACTION WITH ONE'S OVERALL JOB, where OVERALL JOB is conceptualized as a CONCOB (see, e.g., the study by Boswell, Shipp, Payne, and Culbertson (2009)), not a FORMOB, then there was no need, and it's wrong, to "prime" the judgment with object components. If, on the other hand, their theory is about what I call COMPONENTIAL JOB SATISFACTION, then the researchers should have measured satisfaction with each object component *separately*. Separate items were used, for example, in the study of COMPONENTIAL JOB SATISFACTION (my label for this construct) by Law and Wong (1999).

The "jamming" shortcut is a quick path to severe loss of content validity in the measurement of a FORMOB. For a highly content-valid measure, the main components of an abstract formed object must be represented in concrete separate items.

3.5 End-of-Chapter Questions

(3.1) Look up online or in your university library the *Journal of Consumer Research* (a leading interdisciplinary journal in the social sciences). Find three studies that measure the construct of OVERALL ATTITUDE TOWARD THE BRAND AS RATED BY CONSUMERS (this shouldn't be too difficult as "attitude" is the most widely studied construct in the social sciences). Scrutinize the "Measures" section of each of the articles and find out how the researchers represented the *object*, that is, the brand or branded product, in the attitude measure. Do you consider the object representation in the attitude measure to be content-valid? Write a half-page discussing why or why not for each of the three articles you found. (7 points)

(3.2) In C-OAR-SE theory, objects must be classified as either a CONCOB, a COLLOB, or a FORMOB. Classify each of the following objects accordingly and briefly explain why you chose this classification. I have deliberately chosen examples not mentioned in the chapter so that you will have to demonstrate that you fully understand the classification principles. (a) The object in any version of H. Triandis's measure of INDIVIDUALISM-COLLECTIVISM. (b) The COMPETENCE OF THE UNITED NATIONS (THE U.N.) ORGANIZATION. (c) The GENERAL MENTAL ABILITY (OLD NAME "I.Q.") OF AFRO-AMERICANS. (d) R. Petty and J. Cacciopo's NEED FOR COGNITION measure. (e) The two related constructs of PATRIOTISM and WOWSERISM. (1 point for each correct answer and an additional 1 point for a correct explanation of whichever classification you choose; maximum 10 total points)

(3.3) What are the two bad consequences that can result from misrepresentation of the object in the measure? Illustrate each with an example that is clearly different from those in the chapter. (5 points maximum; 2 points each for correct answers and 0.5 bonus points each for selecting examples that are really different from those in the chapter)

(3.4) Explain in your own words what's wrong with the "scattering" and "jamming" shortcuts when measuring a FORMOB. (4 points; and double points if you can locate a new example of each)

(3.5) I have alleged a number of times in this book and in the original C-OAR-SE article that the unacceptably low content validity of (one or both) measures can change correlations that test hypotheses from significance to insignificance, and *vice versa*. In a recent study in the *Journal of Applied Psychology*, researchers Judge, Hurst, and Simon (2009) reported the practically important finding that one's PHYSICAL ATTRACTIVENESS significantly determines one's INCOME (the correlation of scores between their two measures was $r = 0.13$, which is statistically significant, two-tailed, at $p < 0.05$). Look up and read their article—this and their other findings have major real-world importance. Then answer the following two questions. (a) What *object* did they use in their INCOME measure; would this change the correlation with PA, and if so, how? (5 points) Then (b) closely examine the *formed* object (FORMOB) components in their measure of PHYSICAL ATTRACTIVENESS and suggest how you might weight these differently to compute a more real-world relevant or "externally valid" measure. (5 points, thus 10 maximum points total) There are many more construct definition and measure mistakes in this article, and instructors may want to assign it for criticism later in the course.

Chapter 4
Attribute Classification and Measures

> *'(These Are a Few of) My Favorite Things.'*
> —Compare John Coltrane's version with Julie Andrews*'*
>
> *Good works.*
> —Paranomasic (punned) motto of the religious charity, the St. Vincent de Paul Society

The purpose of this chapter is to introduce a comprehensive classification of the *attribute* part of the construct. This chapter is especially important because, as mentioned in Chapter 1, most researchers define the construct *only* in terms of the attribute, making its correct classification crucial for the measure.

After reading this chapter, you should

- Be able to correctly classify the attribute of the construct you wish to measure as either a concrete perceptual attribute (CONCPERC), a concrete psychological attribute (CONCPSY), an abstract achieved attribute (ABACHD), or an abstract dispositional attribute (ABDISP)
- See how to generate one highly content-valid item for a CONCPERC attribute or a CONCPSY attribute; one highly content-valid item for each *first-order component* of an ABACHD attribute; and *several* highly content-valid single items for each *second-order* component of an ABDISP attribute
- Understand that component scores for an abstract attribute—both an ABACHD attribute and an ABDISP attribute—*form* the total attribute score (and that the psychometric "reflective" model, so often assumed, is wrong)
- Realize that all measures are composed of good (highly content-valid) *concrete* single items and sometimes only one such item

4.1 New Fourfold Classification of Attributes

The original version of C-OAR-SE (Rossiter 2002a) identified only three types of attribute, but here, in this book, I propose a different, better, and fourfold classification. Concrete attributes are now divided into two classes: concrete *perceptual*

J.R. Rossiter, *Measurement for the Social Sciences*,
DOI 10.1007/978-1-4419-7158-6_4, © Springer Science+Business Media, LLC 2011

attributes (CONCPERCs) and concrete *psychological* attributes (CONCPSYs). Abstract attributes are now labeled more clearly as abstract *achieved* attributes (ABACHDs) and abstract *dispositional* attributes (ABDISPs). Abstract—or "multiple meaning"—attributes are now shown all to be *psychological*, in that they are constructed by the researcher, although their first-order components (the items to be rated) are *perceptual* (self-reportable by the rater) and must be concrete. Abstract attributes, ABACHDs sometimes and ABDISPs always, have *second-order* components formed from the first-order components (the items). All abstract attributes, which have multiple components and, therefore, always require multiple items, are now conceptualized as being *formed* from the component items' scores, which means that *no* attributes follow the "reflective" model of psychometric theory.

The four types of attribute are previewed in Table 4.1.

Table 4.1 Attribute classification definitions

Classification	Explanation
Concrete perceptual (CONCPERC)	• Unambiguous to raters • Self-reportable (single item)
Concrete psychological (CONCPSY)	• Unambiguous to researcher • Not self-reportable but rated by expert (single item)
Abstract achieved (ABACHD)	• Accumulation of component attributes • Self-reportable as concrete perceptual achievements or performances (multiple items)
Abstract dispositional (ABDISP)	• Manifestation of component attributes • Self-reportable as concrete perceptual mental or behavioral activities (multiple items)

4.2 Concrete Perceptual Attribute (CONCPERC): One Good Self-Rated Item

CONCPERC attributes are the most common type of attribute in marketing, by far. They are probably the most common type of attribute in sociology and psychology, too, when it is realized that BELIEFS and ATTITUDE are concrete perceptual attributes, as are, surprising to many, COMPLEX HUMAN EMOTIONS. Also, the two most common attributes of "memory"—RECOGNITION and RECALL—are CONCPERCs.

Each of the previously mentioned attributes is, first, "concrete" because it has only one meaning—both in the mind of the researcher and in the mind of the rater—and is, second, "perceptual" in that it can be consciously responded to by the rater (unlike a "psychological" attribute, which is responded to, but not consciously and directly); in other words, concrete perceptual attributes are *self-reportable*.

In the social sciences the attributes known as BELIEFS (or PERCEPTIONS or ASSOCIATIONS) are each *unquestioningly* measured with a single item. A single item is also *unquestioningly* selected to measure the attributes of RECOGNITION and RECALL. (I say *unquestioningly* because measurement theorists, such as Nunnally 1978, and Churchill 1979, claim that multiple items are *always* necessary and that a single item can never do the job.) The single-item decisions are correct decisions according to C-OAR-SE theory because the constructs are "doubly concrete" (they have a concrete object and a concrete attribute).

For the unacceptable reason of convention, however, researchers seem to think that OVERALL ATTITUDE and specific COMPLEX HUMAN EMOTIONS must be measured with multiple items. (In this sense they are acting like John Coltrane during his famous variations on the *Sound of Music* song "*My Favorite Things*" when they should keep it simple and real like Julie Andrews did.) I will discuss the case of OVERALL ATTITUDE here and leave the case of COMPLEX HUMAN EMOTIONS until the later, specialized chapter (Chapter 6 on item types).

The construct of OVERALL ATTITUDE most clearly illustrates the fallacy of using multiple items to measure a CONCPERC attribute. The multiple-item fallacy can largely be blamed on the eminent psychometrician Nunnally (1967, 1978), who publicized the multiple-item approach in the social sciences, and the eminent marketing researcher Churchill (1979), who imported the multiple-item fallacy into marketing. Nunnally pronounced that any single item *must* be a fallible measure of an attribute because all single items contain "measurement error;" the idea that follows from this is that a large number of such fallible items must be used so that their "measurement errors," when averaged across the items' scores, will cancel out to zero. This idea is based on the classic "true score" model, which I have already shown to be wanting (in Chapter 1).

The fallacy in the "multiple items are always necessary" idea was brought home to me originally by analogy a couple of years ago, when I revisited my favorite museum, France's Musée d'Orsay. (The Coltrane-Andrews analogy came later.) The museum is famous for its collection of Impressionist paintings—readers will most likely recall the works of the artists Renoir and Monet among the Impressionists—and also for its collection of Neo-Impressionists, of whom Vincent Van Gogh is the most famous. The following descriptions are partly mine but owe a debt to the unnamed authors of the museum's excellent undated guidebook titled *Orsay: Masterpieces of 19th Century Art*. Van Gogh created an "impression" of the object he was painting by using "thick dabs" of different colored paint applied in broad strokes, a striking and memorable example of which is his 1887 work, *Portrait of the Artist*. Monet, whose later work is often classified as Neo-Impressionist, employed the application of paint in a "powdery haze" to deliberately fuzz the object in the painting, such as in his 1904 painting, *London, the Houses of Parliament, Sunlight breaking through the Mist*. In France, Seurat invented the so-called Divisionist technique, taken up in some of the paintings by the remarkable Austrian painter Klimt, in which the paint was applied in "roughly circular dots," also with the purpose of partially obscuring the object in the painting. However, in the Musée d'Orsay you will come across the earlier work of the Naturalism school, of which Manet's work is possibly the best known because people confuse his name with Monet's, and the Realist school of the 1850 s, an art movement generally attributed to Courbet, whose graphic work, *The Origin of the World*, will not be forgotten by those who have seen it and which was probably the inspiration, knowingly or not, for the much later but similarly confronting Benetton ads such as *Newborn Baby*. Well, multiple-item theorists are Impressionists, or an even better analogy would be Neo-Impressionists, whereas I am a single-item Realist. That is, *the best single item is the one that most directly represents the attribute and does not try to obfuscate it with multiple "fallible" items*. A multiple-item measure of a concrete attribute might well be regarded as a more sophisticated "work of art" (with the sophistication justified by the spurious means of coefficient alpha) but, as a "work of science," it's fallacious.

Let me illustrate the "multiple items for a *concrete* attribute" fallacy in another way in case the foregoing analogies don't grab you. Almost all social scientists believe that multiple items must be used to measure an OVERALL ATTITUDE ("attitude" in the sense of the rater's overall evaluation of the attitude object, *à la* Fishbein 1963, and Fishbein and Ajzen 1975, and most recently Fishbein and Ajzen 2010, although in my opinion the late Marty Fishbein rather "lost the plot" in this book by reverting to Ajzen's nondistinction between OVERALL ATTITUDE and ATTITUDES as BELIEFS, see Ajzen 1988, Chapter 1, which was the key distinction in their excellent 1975 book). In marketing, one of the most frequent applications is when measuring ATTITUDE TOWARD THE BRAND, which academics often symbolize as A_b. The attribute, nominally, is EVALUATION. However, in C-OAR-SE content-validity theory, I maintain that the researcher must first figure out and specify the exact meaning of "evaluation" that he or she wants to measure. In the case of consumer evaluation of brands, it is most likely that the researcher wants to find out how much the consumer believes it to be a *good or bad* brand in the sense of *performance*. If so, then the best—the most content-valid—single item will be "Good-Bad," measured on a bipolar (positive through negative) answer scale. Researchers nevertheless feel compelled by convention—the "sheep" phenomenon—to add other items. Several of the most commonly added items to measure A_b are "Like-Dislike," "Useful-Useless," and "Pleasant-Unpleasant" (sources for these commonly used items in multiple-item measures of A_b can be found in the articles by Bergkvist and Rossiter 2007, and Rossiter and Bergkvist 2009). The problem is that these other items pick up *other attributes* that are correlated with the target attribute of PERFORMANCE but are not isomorphic with it (they don't have identical meaning). "Like-Dislike," for example, represents a more "affective" evaluation. "Useful-Useless," on the other hand, signifies a more "utilitarian" evaluation. "Pleasant-Unpleasant," the third alternative item, is simply "off-base" because it can be meaningfully applied to describe only certain types of objects and in the case of brands, perhaps only food or beverage brands (the attribute of PLEASANTNESS is used in Osgood, Suci, and Tannenbaum's classic 1957 book on the "semantic differential" as an example of an attribute that, when rated for different concepts, produces "concept-scale" *interaction*, which in C-OAR-SE terminology is object-attribute interaction).

Another, more complex example of wrongly using multiple items to measure a CONPERC attribute can be found in the recent study by Breivik and Thorbjørnsen (2008) in the lead article in a major journal in marketing. This example is only more complex than the previous example because it requires the reader to *look* *closely at the content of the items*—something I said editors and reviewers rarely bother to do. The example parallels my criticism in the original C-OAR-SE article (2002a, p. 313) of Taylor and Baker's (1994) three-item measure of the BUYING INTENTION construct. In the 2009 study, the researchers labeled the attribute as BEHAVIORAL FREQUENCY, by which they actually meant USAGE FREQUENCY. The researchers unnecessarily (but typically) chose three items to measure the construct that I would label SELF-REPORTED USAGE FREQUENCY OF THE BRAND: (1) "I often use this brand," (2) "I have used this brand for a long time," and (3) "I seldom use other brands in this product category" (see their article, p. 462). They also chose ambiguous Likert answer scales ("Strongly disagree ... Strongly agree") for these items, but that's not the mistake I am illustrating here. The first item, "I often use this brand," is alone sufficient. It's also reasonably content-valid (although it escapes me why the researchers didn't simply ask respondents directly to report their recalled frequency of usage—daily, weekly, monthly, etc.). The second item, "I have used this brand for a long time," measures a *different attribute*, namely PAST DURATION OF USAGE, not CURRENT USAGE FREQUENCY. The third item, "I seldom use other brands in this category," also measures a *different attribute*, which seems to be FREQUENCY OF USAGE OF *OTHER* BRANDS, which is only indirectly relevant to the USAGE FREQUENCY OF *THIS* BRAND. Adding the other two items *must* lower the content validity of the measure and produce wrong (and pretty much uninterpretable) scores.

> There's your trouble, there's your trouble,
> seein' double with the wrong one.
>
> – Dixie Chicks,
> "*There's Your Trouble*"

The fallacy of adding other items to the one best item when measuring a concrete attribute can be demonstrated by the simple graph in Fig. 4.1. Because the items' loadings on the target attribute differ, the multiple-item (average) score will differ. In the example, the best single-item score = 7.0, but the multi-item average score is (5.0 + 7.0)/2 = 6.0. This deviation from the true score occurs with just two items.

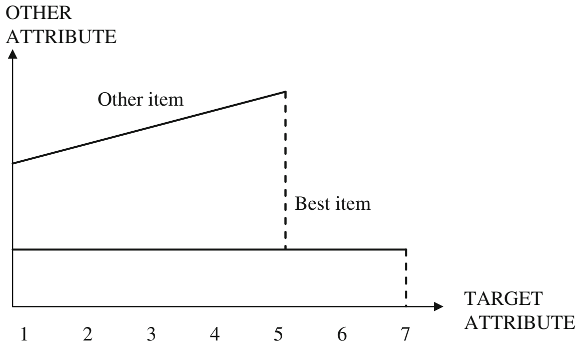

Fig. 4.1 Demonstration of the "multiple items are always necessary" fallacy

Typically three or four items are used and the deviation of the (averaged) observed score from the true score *worsens* with *every additional item.*

The addition of multiple items to the one highly content-valid item when measuring a basic or doubly concrete construct can radically change empirical results and thereby wrongly accept or reject hypotheses and prove or disprove theories. A recent example of this occurred with a manuscript I received when I was editor of a special issue of a marketing journal. The researchers had used a good single-item measure of BRAND ATTITUDE (the 7-point, bipolar-rated item, “Bad-Good;” see Rossiter and Bergkvist 2009) in their first experiment and obtained support for a very interesting, and theoretically predicted, interaction effect from an “informational” advertisement. In their second experiment, they used a “transformational” advertisement and for some reason switched to a *multiple-item* measure of BRAND ATTITUDE, which used four other items in addition to “Bad-Good.” In the second experiment, they again observed their predicted significant interaction effect. However, my advertising grid theory (see Rossiter and Percy 1987, 1997, or Rossiter and Bellman 2005) predicts that this interaction effect should not have occurred for “transformational” advertisements. Also, my C-OAR-SE theory (see Rossiter 2002a) states that they should not have used a multiple-item measure of this doubly concrete construct. I, therefore, asked the authors to re-run the analysis in the second experiment using only the single item “Bad-Good,” same as in the first experiment, to measure the dependent variable. This time, the interaction was nonsignificant and only a significant main effect emerged, as predicted by my theory. I had a bit of trouble persuading the authors to report my version of the results, because they thought that, if anything, the single-item measure was *less* valid according to conventional psychometrics theory. I was the editor, so I won.

There’s a related parable here: you cannot pick and choose measures to suit the results that you want. This is a scientifically incorrect—and unethical—procedure that must tempt many researchers (especially with the easy item-deletion statistics available in software programs such as SPSS, SAS, and more recently R). For a doubly concrete construct, properly defined, there is only “one good item” and it alone must be used. As the St. Vincent de Paul motto says, “Good works.” The Dixie Chicks were also on the ball, as usual (“… seein’ double with the wrong one”).

I thought I might have disposed of the “multiple items are always necessary” fallacy when my article with Lars Bergkvist—one of my very few colleagues who understand C-OAR-SE, as I acknowledged in the Preface—was published as the lead article in the leading research journal in marketing (Bergkvist and Rossiter 2007). This article proved that—for measuring a basic (doubly concrete) construct—a single-item measure is equally as *predictively* valid as a multiple-item measure. Since its publication in May 2007, up until the time that I wrote the first draft of this chapter, August 2009, our article had become the most-cited article in all of marketing (according to the Social Sciences Citation Index, now owned by the ISI Web of Knowledge). Whereas this would normally be cause for celebration, the fact is that this article threatens to become the most *mis-cited* marketing research article ever published! I’ve checked *how* our article has been cited and at least three-quarters of the researchers have used it to "postrationalize" their use of a single-item measure that they had *incidentally* included among their multiple-item measures—or, perhaps, that they *retained* only because other items didn't produce "a high-enough alpha." They did not preselect the single-item measure for the right reason, which is that the construct is doubly concrete.

The "one good item" principle applies when measuring concrete perceptual attributes (CONCPERCs), as in the foregoing examples, and it also applies when measuring *concrete psychological* attributes (CONCPSYs), as we'll see in the next section. A theoretical conclusion not made clear in the Bergkvist and Rossiter (2007) article, but hammered into a clear message in the follow-up article by Rossiter and Bergkvist (2009, see pp. 16–17) is that the "one good item" principle—meaning a single highly content-valid item—applies to the measurement of *all attributes*, because, at the item or "questionnaire" level, *every* item must be "doubly concrete." Scores on abstract attributes are, therefore, an aggregation of scores on single items, each measuring a concrete attribute.

4.3 Concrete Psychological Attribute (CONCPSY): One Good Researcher-Rated Item

Some attributes are concrete and *psychological* rather than concrete and perceptual. As also defined in Chapter 1, a "psychological" attribute is an attribute that is inferred by the researcher and is not directly perceivable by the rater (i.e., it is not a "perceptual" attribute). Psychometricians should also note that a CONCPSY attribute is not a "latent" attribute (it's a *real* attribute—it truly exists in the researcher's mind and subconsciously exists in the rater's mind). *Concrete* psychological attributes mostly occur in psychology rather than in the other social sciences, but several such attributes have been picked up by researchers in consumer behavior. I will discuss two examples of CONCPSYs—IMPLICIT ATTITUDE and FREUDIAN SEXUAL AFFECT—to illustrate how the "one good item" principle applies.

Toward some objects, particularly objects that represent socially or personally sensitive topics, people undoubtedly hold an "explicit" attitude (which is a CONCPERC attribute) and an "implicit" attitude that they are unable to report (which is a CONCPSY attribute). As examples of implicit attitudes, REAL ATTITUDE TOWARD A CULTURAL OUTGROUP is socially sensitive, obviously, and REAL ATTITUDE TOWARD TOPICS TO DO WITH SEX is personally sensitive, as Sigmund Freud (see, e.g., Freud 1949) and many other psychoanalysts have hypothesized. Such IMPLICIT ATTITUDES can be most validly measured by *reaction time* tests (also called *response latency* tests) notable among which is the Implicit Association Test (the IAT; see Greenwald, McGhee, and Schwartz 1998, De Houwer, Teige-Mocigemba, Sprout, and Moors 2009). Reaction time is critical because genuinely existing implicit attitudes should emerge "without thinking." Reaction time is concrete, even though it is measuring a *psychological* attribute that respondents can neither perceive nor report directly. It is a single-item measure.

This measure may be repeated over multiple IAT trials but it is still a single-item measure.

FREUDIAN SEXUAL AFFECT has long interested marketers (especially advertisers). This interest peaked with the publication of the popular book, *Subliminal Seduction*, by Key (1974), who also discussed the less well-known Freudian concept of DEATH-WISH AFFECT. Readers who doubt the veracity of Freud's theories should read the article by cognitive psychologist Westen (1998), in the *Psychological Bulletin*, in which he argues impressively that most of Freud's theoretical constructs have been verified—under different labels—by modern cognitive, social, developmental, and personality psychologists and, I add, more recently by neuroscientists.

Freud's most famous idea, of course, is that SEXUAL AFFECT can be experienced *subconsciously* in response to certain everyday objects that symbolize sexual stimuli of either a phallic, vaginal, or coital nature (see especially Freud 1911). I and a then-Ph.D. student, Janette Overstead (Rossiter and Overstead 1999), attempted to measure the result of subconscious sexual affect in response to Freudian symbolic advertisements; this followed promising but confounded results in an experiment by Ruth, Mosatche and Kramer (1989). However, we used a measure of EXPLICIT ATTITUDE (liking of the ad) instead of, for what is in theory subconscious, a measure of IMPLICIT ATTITUDE, and we obtained null results. But in an unpublished pilot study that I'm hoping to replicate soon with Tobias Langner—another C-OAR-SE *cogniscentum* whom I acknowledged in the Preface—we found that ads which contained Freudian sexual symbols, in contrast with otherwise identical ads for the same brands which did not, produced a second, smaller, but reliable, peak in electrical skin conductance response (SCR) after the usual big SCR peak, which is known as the "orienting response" and is made to any new stimulus. The SCR (also known as GSR and EDR for those who have followed the literature in psychophysiology) is measuring a nonreportable *concrete psychological* attribute—in this case, SUBCONSCIOUS SEXUAL AFFECT—with a single item.

4.4 Abstract Achieved Attribute (ABACHD): One Good Item per Defined First-Order Component

An *abstract* attribute has multiple meanings—from the *researcher's* perspective. In other words, an abstract attribute has *components* that "form" it or make it up. These component attributes must be specified by the researcher and included in the detailed *construct definition*. (They need not be included in the shorter *label* of the construct, which need only mention the main object, the main attribute, and the rater entity.)

Because abstract attributes are defined by the researcher, they are necessarily *psychological* rather than perceptual. Their *first-order (lowest level) components* can be perceived by the rater (and thus the first-order components are *concrete perceptual*, or CONCPERC, attributes), but the *second-order attribute* that these components *form*—in many instances of abstract attributes—*cannot* be perceived by the rater.

The second-order attribute is nonetheless *real* in the *psyche* of the rater rather than a "latent" artifact of the psychometrically trained researcher's statistical imagination.

In my new version of C-OAR-SE theory, presented in this book, there are two types of abstract psychological attribute, now labeled ACHIEVED and DISPOSITIONAL, respectively. This section (Section 4.4) explains the first type, an abstract *achieved* attribute, abbreviated ABACHD. (In the section after this, Section 4.5, I will discuss the other type of abstract psychological attribute, abbreviated ABDISP.) This type of attribute is "achieved" by the individual and it is measured by *essential components*' scores added up by the researcher (or in some cases multiplied, as we'll see in Chapter 7 on *scoring rules*). ABACHDs always consist of component attributes.

I will discuss two important examples of an ABACHD attribute. The first is taken from cognitive psychology and is KNOWLEDGE or rather KNOWLEDGE IN A PARTICULAR FIELD. KNOWLEDGE (in any field) is obviously achieved—learned and accumulated—not inborn. But researchers devising knowledge tests (include the leading examples of knowledge tests in psychology and in marketing) invariably conceptualize the KNOWLEDGE attribute as "reflective" (in new C-OAR-SE theory terms, "dispositional") when it's not. I made the same mistake some years ago when devising a test of MARKETING KNOWLEDGE. I carefully selected 30 items from the excellent—or at least those from the early editions were excellent—Test Bank of the highly acclaimed textbook on marketing management written by marketing management guru Kotler (1978). I was teaching marketing management at the University of Pennsylvania's prestigious Wharton School at the time and I am therefore sure that I had a very good idea of the item content that would represent the important principles and procedures in marketing management. To my dismay, when I applied the conventional psychometric procedure of factor analysis to students' scores on the 30 items, expecting a single "unidimensional" factor to emerge in line with the assumption that KNOWLEDGE is a "reflective" attribute, I found that the test results split into approximately 13 separate factors! I then administered the test to another sample of MBA students (similar to the first sample) and also to a large sample of first-year undergraduate business students. I observed the same finding of multiple, indeed atomized, factors. But ignoring the "nonunidimensionality" of the test, I found some intriguing results with the sum-scores on MARKETING KNOWLEDGE. The results implied that undergraduate students pick up no marketing knowledge *whatsoever* from an introductory marketing management course; their mean score was 44% correct answers on day one of the course and 47% on the final exam, when the test was readministered, which is a nonsignificant difference. MBA students, on the other hand, all of whom had several years of work experience, not only showed more marketing knowledge *going in to* their introductory marketing management course, scoring on average 57% correct, but also demonstrated a *16% absolute gain* in knowledge as a result of the course, scoring 73% on the final exam retest. The important implication is that marketing courses—and perhaps all business courses—should be taught only at the graduate level. I published these results "only" at a conference because I was sure the "psychometrics" of the test would prevent their publication in a good journal. It was only later, when I was formulating C-OAR-SE theory, that I realized that KNOWLEDGE is a *formed* (achieved) attribute and that the atomized "factors" was a result that made perfect sense! Conceptualized and scored as an *achieved* attribute (an ABACHD attribute), my 30-item test of MARKETING KNOWLEDGE, which I had dubbed K-Test in tribute to Kotler, proved to be highly predictively valid for predicting MBA GRADE-POINT AVERAGES, though not undergraduate GPAs, because the "freshmen" picked up no knowledge in their first-year marketing course. K-Test would have been consigned to the dustbin or "file drawer" of research failures had I not realized its true nature. I am currently updating the test, which is an easy job because the important marketing principles and procedures have not changed appreciably in four decades (only the applications—notably Internet marketing—have changed)—and I'm hoping to attract a good doctoral student to take on this portentous research topic as a Ph.D. thesis. It is portentous because, if they are replicated, the results will support the controversial hypothesis that undergraduate business education is a waste of time and effort, a hypothesis believed by a number of leading U.S. universities, including Harvard, Chicago, and Columbia, who teach business management only at the graduate level and to students who have had 3 or 4 years of real-world management experience. I hasten to add that I exclude undergraduate *accountancy* teaching from this hypothesis as accountancy is arguably *not* management and does not need to be taught in a management school—except maybe incidentally to MBA students so they can read and, some would allege "adjust," balance sheets and quantitatively "decode" Panglossy company prospectuses.

The second important example of an abstract achieved attribute (ABACHD) is SOCIAL CLASS, a prominent construct in sociology which *was* prominent in social psychology and in marketing (e.g., Martineau 1957). SOCIAL CLASS is a very powerful predictive (and causal) variable in all the social sciences with the exception of organizational behavior, where the object is not the person or household but the organization. It has become *politically incorrect* in the ever-egalitarian U.S.A. and Australia to include SOCIAL CLASS in surveys; it is still widely used in surveys in the U.K. and Western Europe. But the main problem is due to the devolution of the SOCIAL CLASS construct into what has become known as SOCIOECONOMIC STATUS, or SES, coupled with the unconscionable use of "any old indicators" to measure SES—see the thorough meta-analysis of SES measures by White (1982), and see the use of only parents' education level and family income to measure SES (in a study published recently in the *Psychological Bulletin*, no less!) by Sackett, Kunal, Arneson, Cooper, and Waters (2009). This issue has become tremendously important in Australia, where government educational policymakers naively follow the erroneous belief (fuelled by the 1996 Coleman Report in the U.S.) that educational achievement is almost completely determined—a correlation to the order of .5 but assumed to be 1.0—by parental SES or, in the Australian government's case, by the average SES of the area in which the family resides! I'll discuss this stupid belief shortly.

Young readers will appreciate neither the existence nor major social influence of SOCIAL CLASS, because most, as I did, probably grew up in an upper-middle-class neighborhood, with upper-middle-class friends, and with upper-middle-class colleagues at a local university; this “cocooned” existence, I believe, blinds you, as it did me, to the severe realities of the social divisions that exist within every known society. The blinders abruptly come off if you happen to get a job in real-world marketing research where you have to qualitatively interview “real people,” a job which I was lucky enough to have in the United States for several years between my Masters and Ph.D. degrees and which I continued to do part-time as a consultant when I became an academic. Even if you’re not fortunate enough to get such a job, go visit India, China, or even the “modern” country of Japan and your eyes will rapidly open to the reality of SOCIAL CLASS—and “REVERSE” RACISM.

The operative attribute in the construct of SOCIAL CLASS is not CLASS, except in a minor meaning of that word, but rather *PRESTIGE*. The actual construct is the SOCIAL PRESTIGE OF THE HEAD OF HOUSEHOLD AS PERCEIVED BY THE PUBLIC. This is made clear in sociologist W. Lloyd Warner’s original measure of social class (see his 1949 book) which, as summarized by White (1982, p. 462, emphasis added) to highlight the prestige attribute, consisted of “(a) Occupation of principal breadwinner, (b) *Source* of income (i.e., inheritance, investments, salary, wages, or welfare), (c) *Quality* of housing, and (d) *Status* of dwelling area.”

Table 4.2 is a measure I made up approximately 25 years ago for a consulting project in Australia, where we have the same seven SOCIAL CLASS strata as the U.S.A., the U.K., and Western Europe. This measure is not based on SOCIODEMOGRAPHIC PRESTIGE as above but instead is based on VALUES derived from Warner’s pioneering theory and also on some later insights contributed by a leading sociologist, Coleman (1983), and especially marketing sociologist Levy (1968). The codes on the left are: X = “X” class, the sort of Maslowian “self-actualized” class that is completely independent of income (see Fussell 1983); UU = Upper-upper class; LU = Lower-upper class; UM = Upper-middle class; LM = Lower-middle class; UL = Upper-lower class; and LL = Lower-lower class. (The items, of course, should be put in a *random order* in the questionnaire.) These seven *classes* are *second-order* components of the overall ABACHD attribute of SOCIAL CLASS. Checking now from a C-OAR-SE perspective, I would still vouch for the high-content validity (items and binary answer scale) of this measure.

The desire for a “shortcut” measure has led to widespread use of the substitute and perhaps more politically correct term SOCIO*ECONOMIC* STATUS, or SES, to replace SOCIAL CLASS—naively implying that high social class “ain’t classy” any more! The SES measure is always a *demographic* measure—usually a composite of the head of household’s Occupation (ranked in terms of social prestige, ironically) and Education level; a third indicator, total family Income, is sometimes added but Income actually *lowers* the correlation of SES with CHILDREN’S SCHOOL ACHIEVEMENT because it is confounded with the other two indicators and, anyway, Income is often refused or lied about in surveys (see White 1982). The Occupation + Education index of SES correlates about $r = .33$ (a “moderate” effect size) with the CHILD’S SCHOOL ACHIEVEMENT. However, with the student’s I.Q. taken into account—controlled for, or “partialed out” statistically—the relationship between SES and school achievement drops to an almost negligible (“small”)

Table 4.2 A values-based measure of social class (Rossiter 1984). Worded for Australia but easily adaptable for other countries

Respondent (rater entity): primary head of household			
X1	Can speak or understand several languages	Yes	No
X2	Enjoy classical music	Yes	No
X3	Can choose own working hours in full-time profession	Yes	No
X4	Regard a car as just something that gets you from A to B when necessary	Yes	No
X5 (neg)	Have in the home several collectors' series, for example, book series, special plate series, ornament series, etc.	Yes	No
UU1	Inherited most of our money	Yes	No
UU2 (neg for LU)	Come from a well-known, established British or Australian family	Yes	No
UU3	Our children will never have to worry about money	Yes	No
LU1 (neg for UU)	Own or plan to buy a Mercedes, Jaguar, or Rolls Royce automobile	Yes	No
LU2	Now earn a lot of money	Yes	No
LU3	Believe it is important for career reasons to belong to the best social circles	Yes	No
UM1	Professional success is more important than salary or income	Yes	No
UM2	Very important that my children do well at school	Yes	No
UM3	Husband and wife should be equals in deciding money matters	Yes	No
UM4	Active in community where we live	Yes	No
UM5	Eat evening meal later than most people	Yes	No
UM6	Have morning newspaper delivered to home	Yes	No
UM7	Like to watch news and documentaries on television	Yes	No
LM1	Like to spend most evenings and weekends at home with family	Yes	No
LM2	More important that my children are good and well-behaved rather than brilliant academically	Yes	No
LM3	In our household, everything has to be done on a regular schedule each day	Yes	No
LM4	A family is not a family without a good husband and father	Yes	No
LM5	We frequently watch movies on TV	Yes	No
UL1	Prefer to wear work clothes rather than shirt and tie, or dress, to work	Yes	No
UL2	Very important for our family to be part of the larger family that encompasses our parents and relatives	Yes	No
UL3	Children should be able to look after themselves from a relatively early age, the sooner the better	Yes	No
UL4	Most important that the husband holds a regular job and is never on welfare or out of work	Yes	No
UL5	Often buy an evening paper on way home from work	Yes	No
UL6	Our favorite shows on TV include "real life" serials and game shows	Yes	No
UL7	We usually spend holidays or long weekends visiting relatives	Yes	No
LL1	Keeping up with the bills and payments is a constant daily battle that I feel I'm always losing	Yes	No
LL2	We couldn't get by without some financial help from the government to meet living costs	Yes	No
LL3	Even if it means money that should be going to the family, I feel it's necessary to take some of it for myself, to have a drink with friends or to buy something once in a while to make me feel good about myself	Yes	No
LL4	We can rarely afford to eat at restaurants, even the less expensive ones	Yes	No
LL5	Our social life is very limited compared with most people's	Yes	No

effect size of $r = .16$, according to White's U.S. meta-analysis. This weak relationship would undoubtedly be unchanged with modern data in *any* industrialized country.

Regarding income, which is often wrongly used by market researchers to measure SES, researchers Plug and Vijverberg (2005) have demonstrated that children of higher-income parents do better in school because—on average—they inherit superior GENERAL INTELLIGENCE (see my Chapter 2, Table 2.1). School achievement is *not* the result of higher-income parents "being able to afford to buy their children a better education." In my home state of New South Wales in Australia, for instance, government-run (low-fee) high schools quite regularly outperform private (high-fee) schools in producing children with high graduating marks—especially government-run "selective" schools, which select students on I.Q. (though they don't call it that). The naive meta-analysis by, not surprisingly, U.S. researchers Sackett et al. (2009)—which, incidentally, supported the minor influence of SES on the SAT-College grades correlation—completely ignored students' I.Q. together with the extensive research on GENERAL INTELLIGENCE (now a totally taboo topic in the U.S.A.). What *was* surprising was that this research team comes from the University of Minnesota, home of one of the most famous genetic researchers in the world—David Lykken. This is another sad case of the interdisciplinary ignorance characterizing universities today.

The demographic (SES) interpretation conceptualization of SOCIAL CLASS was used by Bollen and Lennox in their influential article (1991) as the main example of "formative indicators" (earlier called "cause indicators" by Blalock 1964). In the original C-OAR-SE article (Rossiter 2002a, p. 314, note 6) I argued against this conceptualization because the term "formative indicators" refers to the *items*, not the attribute. The attribute—which should be referred to only as SOCIOECONOMIC STATUS—is a *formed*, or what I now call, hopefully more precisely, an *achieved* attribute. (It is also *second-order in total*, with the demographic indicators being lowest, or *first-order*, component attributes.)

I now argue, in this updated version of C-OAR-SE theory, that even abstract attributes which have "reflective indicators" (Bollen and Lennox's term, which Blalock called "effects indicators") are *formed*—from a C-OAR-SE measurement standpoint. This type of abstract attribute was called *eliciting* in my 2002 article. For a reason given shortly, the new term for this type of attribute is *dispositional*.

4.5 Abstract Dispositional Attribute (ABDISP): Several Good Items per Defined Second-Order Component

The other type of abstract attribute is an *abstract dispositional* (ABDISP) attribute. (Two prominent social science examples discussed below are specific ABILITIES and PERSONALITY TRAITS.) Dispositional attributes are *abstract* in that they have multiple meanings for the researcher. Like all abstract attributes, they are also *psychological* because they are inferred by the researcher in defining the construct and are not directly perceivable and reportable by the rater. (Only the *first-order components* of an abstract attribute are self-reportable.) ABDISPs always have (abstract) second-order components and so in total the attribute itself is *third-order*.

In the original (2002) version of C-OAR-SE, I called this type of attribute *eliciting* but I no longer believe the label to be appropriate. This is because I realize now that it may lead to confusion with Skinner's much earlier (1938) use of the term "elicited" to refer to responses "brought out by a stimulus," as distinct from responses automatically "emitted" by the organism, and it is the *latter* that I had in mind when describing this type of attribute. I am, therefore, changing the label to the long-established definition of a *disposition*, which is "an internal readiness of the organism to act in a certain way" (paraphrased from English and English's excellent 1958 dictionary of psychological terms) and now I quote directly (from p. 158): "Specific dispositions are named for the kind of behavior effects produced . . . ." Thus, specific PERSONALITY TRAITS and ABILITIES are *dispositional* attributes. (In neo-Hullian, stimulus–organism–response, or S–O–R, theory which I advocate as a general paradigm for the social sciences—see Rossiter and Foxall 2008—dispositional attributes are "O" variables.)

I used to think that eliciting (now dispositional) attributes followed the classical psychometric "reflective" model because an eliciting attribute, as an internal disposition, "reflects" its manifestations. In the original version of C-OAR-SE theory (Rossiter 2002a) and also in my subsequent journal articles on C-OAR-SE theory (Rossiter 2005, 2007a, 2008a, Rossiter and Bergkvist 2009), I also believed that the manifestations' scores (the items' scores) should be very highly correlated and thus "unidimensional"—that is, the scores should load highly on a single vector or "factor" in the factor-analytic sense. But I now realize that the unidimensionality requirement contradicts one of my main C-OAR-SE principles! This is the principle that the validity—the content validity—of a measure *cannot* be established by looking at its *scores* (see the Construct → Measure → Score, or C → M → S, model in Chapter 1, Fig. 1.2). Yet, this is precisely what factor analysis does and so too coefficient alpha! *All statistics, including statistical tests of "unidimensionality," are irrelevant for evaluating measures*.

The "reflective is actually formative" argument is reasonably easy to demonstrate in the case of PERSONALITY TRAITS, so I will discuss that example first. Take the personality trait of EXTRAVERSION-INTROVERSION. This personality trait is most strongly identified with the psychologist Hans Eysenck (though Freud's student, the psychoanalyst Carl Jung, is generally credited with its label; see Eysenck 1981). Eysenck originally defined EXTRAVERSION (with negative scores signifying INTROVERSION) as consisting of *seven* sub-traits, which he called Sociability, Risk-Taking, Expressiveness, Lack of Reflection, Activity, Impulsiveness, and Lack of Responsibility (see Eysenck and Wilson 1976). However, the authoritative *Dictionary of Psychological and Psychoanalytic Terms* by English and English (1958) defines EXTRAVERSION as consisting of only *three* sub-traits, which correspond with Sociability, Expressiveness, and Lack of Reflection. This means that four of Eysenck's theorized sub-traits are omitted: Risk-Taking, Activity, Impulsiveness, and Lack of Responsibility. Indeed, Eysenck himself, while retaining the Risk-Taking and Activity sub-traits, later moved the last two sub-traits, Impulsiveness and Lack of Responsibility, to become the main sub-traits of a new personality trait that he called PSYCHOTICISM (forming the third trait in his Extraversion–Neuroticism–Psychoticism, or E–N–P, sometimes called P–E–N, theory). This means that some items in his "old" measure of personality traits no longer are content-valid in that they do not represent the new constructs.

What's the outcome of these shifts and comparisons? Well it's a vital one as far as the content (and hence construct) validity of measures is concerned. EXTRAVERSION is typically measured by selecting items "reflectively" (following Nunnally's "domain sampling" theory of measurement) that load on a single factor which is presumed to be the "latent" (i.e., unobservable) trait of "EXTRAVERSION:" see the measures developed by, for example, Norman (1963) and the "Big Five" personality-trait theorists Goldberg (1992) and Back, Schmukle, and Egloff (2009). Their "random sampling" procedure completely ignores the defining sub-traits! For example, "reverse engineering" (by simple content analysis) of Norman's (1963) "EXTRAVERSION" items reveals that he has *inadvertently* represented just four sub-traits, which are Sociability, Risk-Taking, Expressiveness, and Lack of Reflection. Statistical analysis of the scores (S) has led to the measure (M) and, in a clear example of the "psychometrics fallacy," the measure has now defined the construct (C)! The researcher (not the raters) must first define—and argue for—the components (here sub-traits) that are going to be included in the definition of the construct (here the attribute of EXTRAVERSION) and then, by expert judgment—read "extensive colloquial vocabulary"—generate *several* items sufficient to represent each sub-trait in terms of the typical mental or behavioral activities manifested in theory by the sub-trait (see my Chapter 6). There are absolutely no statistics involved and the items' scores need not be "reflective" of anything. Rather, the items are first-order (and "doubly concrete" CONCOB-CONCPERC) components of the defining sub-traits. The sub-traits are *second-order* constructs—and several, say four to six, items are needed, for each, to *precisely* classify individuals ($R_{\text{precision}}$, see Chapter 2). Their scores are added up (then averaged) in a "formative" sense—in my terminology, *formed*—to form the third-order construct which is the TRAIT itself.

A similar problem arises with measures of the trait that Eysenck called NEUROTICISM. In his writings, Eysenck often called this trait EMOTIONAL INSTABILITY (low scores which would signify EMOTIONAL STABILITY or "adjustment"; see Eysenck and Wilson 1976). A better—and less pejorative—term for EMOTIONAL INSTABILITY would be EMOTIONAL *LABILITY* because this trait really means that the person's overall affective state *fluctuates*, or goes up and down, without notice—for example, literally "happy one minute and sad the next." In its severe and long-duration form, this fluctuation is akin to manic depression, or "bipolar disorder," but thankfully nowhere near as dysfunctional—in fact the LABILITY is often quite *functional* because both the mania and the depression quickly dissipate and do not linger to bother others, much like "catharsis" in the colloquial notion of "getting it out of your system." (For socially important evidence of CATHARSIS—a CONCPSY attribute, by the way—see *The New York*

Times of January 8, 2009, for a writeup of a correlational study implying that "slasher" movies may be responsible for an estimated *reduction* of about 1,000 assaults per weekend in the U.S. over the last decade! Also, see the amazing graph in the *Psychological Bulletin* article by Ferguson and Kilburn (2010, p. 176), showing clearly an almost perfect, $r = -.95$, *negative* correlation between video-game sales and youth violence!) However, many U.S. researchers, who now pointedly do not reference Eysenck because of his copious production of "nonegalitarian" evidence of genetic determinants of intelligence, personality, mental illness, and criminal behavior (I met the man and have read as much of his prodigious work as I possibly could, including his fascinating and erudite autobiography, *Rebel With a Cause*, 1997). U.S. psychologists are being bigoted—and, worse, unscientific—and have, probably without realizing, drifted away from measuring the essential attribute of EMOTIONAL LABILITY. What they are measuring is a very different attribute, namely CHRONIC TRAIT ANXIETY, which misses the fluctuation aspect entirely. This content-validity error—induced solely by factor analysis—has slipped unnoticed into popular "short" versions of measures of the "Big Five" traits. For example, Saucier's (1994) brief version of Goldberg's Big-Five trait markers includes as items purportedly measuring the trait of NEUROTICISM the completely "off-attribute" items "Relaxed," "Jealous," and an obviously redundant synonym for the latter item, "Envious." Another example is Lahey's (2009) definition of NEUROTICISM, in the *American Psychologist* no less, as "relatively stable tendencies to respond with negative emotion . . ." (p. 241). This definition is so far off Eysenck's definition that it's a scientific scandal.

The second example of an ABDISP attribute that I wish to discuss is ABILITIES. Take as the most important case in all of psychology the attribute now known as GENERAL MENTAL ABILITY or GMA (previously known politically incorrectly as GENERAL INTELLIGENCE, or I.Q., terms which I will continue to use). GMA always has been, and continues to be, the single best predictor of JOB PERFORMANCE (for good evidence, see Hunter and Hunter 1984, and Schmidt and Hunter 1998, their table 2). To put this relationship in concrete terms, the average I.Q. of white-collar professionals in the U.S. is 124, of blue-collar "skilled" workers such as electricians or plumbers 109, and of "semi-skilled" workers such as truck drivers or hairdressers about 95, and of "unskilled" workers about 80, where the total population average for White Americans is 100 (see Harrell and Harrell 1945, and do not entertain the naive *tabula rasa* notion that these differences have diminished or disappeared since then).

Nearly every intelligence theorist (the notable exceptions are the poly-ability theorists J.P. Guilford and Robert Sternberg) conceptualizes GMA as consisting of two sub-abilities, Verbal and Mathematical Ability. For example, the very widely used U.S. Scholastic Aptitude Test, or SAT, gives separate scores for these two sub-abilities and the items on the SAT are accordingly of two types, Verbal and Mathematical. (The Australian government annually tests kids for what it calls "Literacy" and "Numeracy," refusing to see that these are the two fundamental components of I.Q., and blaming low performance on teachers!) Here, a good argument can be made that item scores on Verbal Ability and, separately, Mathematical

Ability *should* be quite highly correlated, and thus appear to be "unidimensional." This is because, unlike in the case of a PERSONALITY TRAIT, sub-sub-abilities of Verbal and Mathematical Ability are necessary components of the sub-ability itself (e.g., Verbal Fluency is necessary for Verbal Ability, and Abstract Reasoning is necessary for Mathematical Ability). The trick in selecting items for measures of Verbal Ability and Mathematical Ability is not so much in selecting correlated items, because their scores will almost inevitably be correlated (and produce a high coefficient alpha if there are enough of them), but in selecting items that represent a wide range of *difficulty* so that the VA and MA scores of individual test-takers can be reliably arrayed ($R_{\text{precision}}$) at the individual level. The need for high $R_{\text{precision}}$ is also why a *large number of items* is needed for testing GMA, and the items should be several for each second-order component, which are the sub-sub-abilities of the sub-abilities of VA and MA. (This makes GMA a *fourth-order* construct, note.) I will resume this discussion of I.Q. measurement in Chapter 6.

Another example of an ABDISP is CREATIVE ABILITY (discussed in Chapter 6). In terms of importance for humankind, CREATIVE ABILITY is a very close second to GMA.

ABACHD or ABDISP? I wish to emphasize here that abstract attributes cannot be *postclassified* as an ABACHD or an ABDISP attribute (or as "formative" versus "reflective" in the deficient *old* terminology of psychometrics). Many researchers appear to believe that postclassification is both possible and legitimate (e.g., in the "method of tetrads," see Coltman, Devinney, Midgley, and Venaik 2008, or in the now ubiquitous comparison of the results of "structural equation models" for "best fit," and I'll be polite and not use the word that rhymes with "fit!"). I railed against the postclassification fallacy in the original C-OAR-SE article (Rossiter 2002a, p. 315, note 8) and I am still constantly protesting it in reviewing new manuscripts, particularly those in which the researchers have jumped like sheep onto the structural equation modeling, which I call "silly empirical meddling," bandwagon.

Abstract attributes—just like concrete attributes—have to be classified *beforehand*, using *theory* and not statistics. In other words—in "oldspeak"—you must not label an attribute (or construct) as "formative" or "reflective" after the fact, that is, after you've used the measure and looked at the "alpha" of its item scores. This mistake is especially prevalent in management and marketing journals of late.

4.6 The Serious Problem of Mislabeling Abstract Attributes

I want to call attention to another serious problem with researchers' use of factor analysis to select and delete items for measuring *abstract attributes*. In C-OAR-SE, I advise against using factor analysis—or indeed any statistics whatsoever—in the design of measures, but I have no doubt that this erroneous psychometric practice will continue because factor analysis is the social science sheep's favorite fodder and sheep would rather eat than think! That FA will continue to be used is evidenced by Fabrigar, Wegener, MacCallum, and

Strahan's (1999) review in the journal *Psychological Methods* in which they cite but then ignore Armstrong and Soelberg's (1968) convincing dismissal of this statistical technique in the *Psychological Bulletin*. If you *are* going to use factor analysis, you should be made aware that expert factor analysts—notably J.P. Guilford and William Stephenson, two pioneers of psychometrics—emphasize that extracted factors should be named (labeled) by *several language-literate judges*; the judges should scrutinize the *content* of the several highest-loading items to come up with a semantically common-content description to label the factor. This "consensus" step is vital because the factor label—an abbreviation, of course—will *become* the *construct* (actually it will become the *attribute* of the construct if you follow the three-part C-OAR-SE definition of a "construct"). But consensus labeling of empirically derived factors is *never* practiced these days—check the method section of *any* journal for studies that employ factor analysis—either "exploratory factor analysis," EFA, or "confirmatory factor analysis," CFA, as it has been hyperbolically tagged. It should be dubbed "SFA."

I'll give just one concrete example of the consequences of this all-too-regular oversight and I'm sure you can find many others if you search the journals. (I'm not going to include this search as an end-of-chapter question because I don't want to encourage the use of *factor analysis* in any way, shape or form, misleadingly "confirmatory" or timidly "exploratory." Read Armstrong 1967, and Armstrong and Soelberg 1968, and you'll see why. Random numbers produce meaningful "factors"—a classic case of GIGO! Even with real numbers, factor analysis is unnecessary and misleading, producing psychometric illusions called "latent" constructs.) In a recent article, the researcher carried out the usual factor-analysis procedure (it was Cattell's R-type factoring, but most researchers don't know what this is or how it differs from other modes of factor analysis) on a set of items intended to measure a SERVQUAL derivative for INTERNET RETAILING, or E-RETAILING (see Rossiter 2007a, 2009a). A group of six items was found whose ratings produced a factor that the researcher labeled *CUSTOMER SERVICE*. Inspection of the items for this factor, however, reveals that what was measured was not CUSTOMER SERVICE in general, but rather customer contact with the e-retailer *other* than the mechanical online interaction with the e-retailer's website (the three highest-loading of the six items were "Communications with this firm seemed personal," "Contacting customer service staff was easy," and "The company was happy to fix any problems"). Anyone casually reading this article and not looking at the items would doubtless see the label CUSTOMER SERVICE and think that it referred to a really important factor (attribute) in any form of retailing. After all, "Customer Service" is what retailing is all about! The researcher then used the *same label*, CUSTOMER SERVICE, to refer to a 4-item *subset* of the six items. In one regression of LOYALTY INTENTIONS (plural where it should have been singular and should have been measured with a good single item, since INTENTION is a CONPERC attribute) on the service quality factors, the researcher found that the earlier *6-item* measure of CUSTOMER SERVICE was an important (it had a statistically significant regression coefficient) though not the *most* important predictor. But in a second regression with the *4-item* measure of

CUSTOMER SERVICE it *wasn't* significant (the actual *p*-value was .057, so I'd say "marginally significant" although I would never attribute this degree of precision to such sloppy measures). These are patently implausible results: how on Earth can CUSTOMER-PERCEIVED SERVICE QUALITY not be the major predictor (and cause) of LOYAL PATRONAGE of a service provider? The ill-chosen *label* CUSTOMER SERVICE was scientifically misleading—for both measures.

The mislabeling problem is a further demonstration of (a) the paramount role of the *content validity* of measures, (b) the importance of having the *researcher* define the construct, and (c) the need for "sheepdip-blinded" reviewers to read the actual items *and* check the summary labels. (On a lighter note illustrating the importance of reading—and writing—carefully, here is a news item I spotted in *The Australian* newspaper, December 14, 2009, p. 31: "A Russian-born airport shuttle driver who dropped off passengers with a blood-alcohol level four times the limit—after drinking vodka—has avoided jail." *The Australian*'s "Talking Turkeys" Editor's comment: "But wasn't he right to drop them off?")

4.7 Attributes Can Change Status According to Their Role in the Theory

The change in status (classification) of the attribute is, most often, a change from a *componential* definition (from either an ABACHD or an ABDISP attribute) to an *overall* definition (to either a CONCPERC or a CONCPSY attribute).

Two examples of different status involve the attributes SERVICE QUALITY and JOB SATISFACTION. Both of these, in their construct definitions, should be preceded by the label COMPONENTIAL or by the label OVERALL, depending on what the researcher's theory is about (see my new measure called ER-SERVCOMPSQUAL in Rossiter 2009a).

Another extremely common example in the social sciences is the attribute ATTITUDE (discussed in Rossiter 2002a, and again in several places in this book). Researchers (and brand managers) studying the *antecedents* of ATTITUDE—that is, trying to find out what BENEFIT BELIEFS and EMOTION-STATES it is caused by (see Rossiter and Percy 1997)—are studying COMPONENTIAL ATTITUDE. Researchers studying how well an ATTITUDE predicts a BEHAVIOR are studying OVERALL ATTITUDE—an attribute that functions in a *concrete* manner as *one cause* of the subsequent object-specific BEHAVIOR (see Fishbein 1963, Fishbein and Ajzen 1975).

4.8 End-of-Chapter Questions

(4.1) In a recent study by Clark and Wegener (2009) published in the *Journal of Personality and Social Psychology*, acknowledged as the leading journal in social psychology, the dependent (outcome) variable was, using my label, OVERALL ATTITUDE TOWARD NUCLEAR POWER PLANTS (AS RATED BY COLLEGE STUDENTS AS INDIVIDUALS). The researchers used two different measures of this construct, one measure "pre" and the other measure "post." The premeasure was a single item in which the students rated the object NUCLEAR POWER PLANTS on a 9-point bipolar scale that was end-labeled "1 = Definitely opposed" and "9 = Definitely in favor" (p. 46). The postmeasure was a 5-item measure in which the students rated NUCLEAR POWER PLANTS on five 9-point "semantic differential" scales end-anchored by the word pairs "Bad-Good," "Harmful-Beneficial," "Negative-Positive," "Unnecessary-Necessary," and "Wise-Foolish" (p. 47). First, justify your C-OAR-SE classification of the *attribute* in this construct. Then explain which of the two measures, single-item or multiple-item, is the more valid, and why. (7 points)

(4.2) In a recent study in the *Journal of Consumer Psychology*, Sundie, Ward, Beal, and Chin (2009) attempted to measure the emotion of SCHADENFREUDE. Look this word up in a German-English or English dictionary and write down the definition. Then (a) explain how you would classify this attribute in terms of C-OAR-SE theory. The researchers measured SCHADENFREUDE by asking the student respondents to read a lengthy scenario describing another (fictional) student's misfortune—a very public mechanical failure of his expensive and ostentatious new car—and then to rate how much they personally felt the following *separate* emotional states in response to the event: "Happy," "Joyful," "Satisfied," "Glad." For part (b) of your answer, explain what is wrong with this measure. (3 points maximum for a correct answer to part a and 4 points maximum for a well-argued answer to part b)

(4.3) A heavily debated issue in marketing at present concerns the apparently small amount of influence that the marketing department has within companies and other organizations, including government and not-for-profit organizations. Find and read the study on this by Verhoef and Leeflang in the *Journal of Marketing* (March 2009, pp. 14–37). Scrutinize the measures of the constructs in their study (see their Appendix A, pp. 31–33). The researchers arbitrarily classified the 20 constructs (attributes) as "formative," "reflective," or "not classified." Classify the 20 attributes correctly according to the new *fourfold* classification in C-OAR-SE theory, briefly explaining your judgments. (This question is tricky for instructors to mark because the same misclassification errors are likely to be repeated across items. There are no CONCPSYs, so I suggest a maximum of 7 points each, or 21 total, for getting all of the CONCPERCs, ABACHDs, and ABDISPs correct, and a minimum of 4 points each, or 12 total, for a reasonable effort overall even if wrong.)

(4.4) Just for practice, identify and classify, in C-OAR-SE terms, the *object* in the construct in Q4.1 and Q4.2. (2 points each, 4 total)

(4.5) Why are *all* abstract attributes "formed" and not "reflective?" What does this say about virtually all multiple-item measures reported in the social science journals? Try to find two measures that you think escape this criticism and explain why. (7 points maximum for answers to the first and second questions and 5 points maximum for the last)

Chapter 5
Rater Entity Classification

'Everything is subjective,' you say; but even this is interpretation invented and projected behind what there is.
—Friedrich Nietzsche, *The Will to Power,* 1883, p. 481

Or, on a lighter note, Jimmy Barnes sings: "My girl is red hot." Chorus: "Your girl ain't doodley squat."
—Billy Riley & His Little Green Men, *"Red Hot"*

The third—and final—element of any construct is the *rater entity* (Rossiter 2002a). Just as almost all researchers leave out the object when defining a construct, so also do they neglect to include the rater entity. As we saw earlier, well-known examples of constructs in the social sciences that don't specify the rater entity are CORPORATE REPUTATION, RETAILER SERVICE QUALITY, INDIVIDUALISM-COLLECTIVISM, and less well-realized examples are BRAND RECOGNITION and BRAND RECALL. In none of these construct labels is the rater entity identified. Failure to account for different rater entities, like failure to identify the object in the construct (see Chapter 3), has led to confused findings—or "mixed results"—in the social sciences, particularly in *meta-analyses*. Inclusion of the rater entity as part of the construct sounds very "postmodern"; superficially it seems to follow the Postmodernists' belief that "Everyone's perspective is equally valid." This is an old idea in philosophy, as the opening quote from Nietzsche attests (and I could have gone back to Plato and other ancient Greek skeptics for this idea), and it deserves to be reinstated in measurement theory. And it is, in C-OAR-SE.

After reading this chapter you should:

- Remember always to specify the rater entity in your definition of the construct
- Be able to correctly classify the rater entity as either: experts (EXPRAT), coders (CODRAT), managers (MANRAT), consumers (CONRAT), or the individual (INDRAT)

J.R. Rossiter, *Measurement for the Social Sciences*,
DOI 10.1007/978-1-4419-7158-6_5, © Springer Science+Business Media, LLC 2011

5.1 Why the Rater Entity in the Construct Makes It a Different Construct Even If the Object and Attribute Are the Same

As I said in the introduction to this chapter, researchers' failure to specify the rater entity has led to much confusion and is one of the main reasons why researchers—in literature reviews and meta-analyses—get away with reporting that there are "mixed findings" regarding the theory or hypothesis of interest. The other main reason for "mixed findings," as should be evident from reading this book, is researchers' use of low-validity measures and often *different* low-validity measures of the same construct. If the theory is fully thought through *and* the constructs in the theory are properly defined and validly measured, there should never be "mixed findings." The usual excuse for contradictory findings is that there must be an undiscovered "contingency"—or worse, several undiscovered contingencies in the form of "moderator variables." This is an acknowledgment that the functional relationship between the constructs has not been fully thought through by the researcher. It happens a lot with functional relations empirically inferred from path-regression models or structural equation models. The moderator variables are usually *rater entity* variables.

Here are two examples that illustrate how "mixed findings" arise from failure to define the construct as including a particular type of rater entity. Organizational behavior researchers Hihouse, Broadfoot, Devendorf, and Yugo (2009), in a study published in the *Journal of Applied Psychology*, measured MAJOR AMERICAN COMPANIES' CORPORATE REPUTATIONS. The raters were professors of finance, marketing, and human-resources management. This purportedly is a case of EXPERTS as the rater entity (EXPRAT). Some experts the finance professors must have been! The researchers found that these three types of professor did not differ at all in their ratings of the CORPORATE REPUTATION of *any* of the nine COMPANIES that were the objects in their study (such as GENERAL MOTORS; DISNEY, which the researchers mistakenly referred to as the "WALT DISNEY" company; SONY; and McDONALD'S). All three groups of raters found "minimal variability" (p. 782) in these companies' REPUTATION. A slightly earlier study by Gerzema and Lebar (2008) using ratings by *true* experts—INDUSTRY ANALYSTS—revealed this finding to be nonsense. These companies in reality have very different CORPORATE REPUTATIONS and this is objectively reflected in their very different MARKET VALUES (the monetary value of their intangible assets). Also, the *JAP* researchers used an unnecessary and shamefully constructed multiple-item measure of CORPORATE REPUTATION (see p. 785 of their article). A single-item measure—as used by corporate market research practitioners—would suffice (see Rossiter and Bellman 2005, Chapter 16).

In another organizational behavior study, Ilies, Fulmer, Spitzmuller, and Johnson (2009) studied the construct of CITIZENSHIP BEHAVIOR. They correctly distinguished the two *objects* of the construct as the EMPLOYEE (which they called "THE INDIVIDUAL") and the ORGANIZATION—thus EMPLOYEE CITIZENSHIP BEHAVIORS and ORGANIZATIONAL CITIZENSHIP BEHAVIORS. The researchers used two different *rater entities*—SELF and OTHERS—to do the ratings of CITIZENSHIP BEHAVIORS (but did *not* distinguish the two rater entities in their construct definitions). Incredibly, the SELF versus OTHERS rater entity was interpreted as a "moderator variable!" However, OTHERS surely should be the only objective and relevant rater entity. No one would trust a SELF-report of these obviously socially desirable (and corporate-desirable) behaviors.

In C-OAR-SE theory, five main types of rater entity are identified: Experts (EXPRAT), coders (CODRAT), managers as a group (MANRAT), consumers as a group (CONRAT), and the individual (INDRAT). The classifications are defined in Table 5.1 and discussed in the chapter.

Table 5.1 Rater-entity classification definitions

Classification	Explanation
Experts (EXPRAT)	• Individuals highly qualified in the substantive field (e.g., financial market analysts, professional clinical psychologists, other social scientists who have contributed demonstrably to theory)
Coders (CODRAT)	• Trained content-analysis coders who have successfully completed a pretest of the coding scheme
Managers (MANRAT)	• A group rater entity (*level* of management should be specified in the construct definition)
Consumers (CONRAT)	• A group rater entity (usually subdivided into Prospective, Current, and Lapsed Customers of the brand)
Individual (INDRAT)	• An *individual* rater entity often nominally aggregated as a group for the analysis (the *object* is assumed to be the rater's Actual Self unless otherwise specified)

5.2 Experts as the Rater Entity (EXPRAT)

Experts are the rater entity for many important constructs in the real world. These include financial-strength ratings of companies, creditworthiness ratings of companies; experts' ratings of other individuals such as prospective employees or job applicants—and, of course, medical doctors' assessments of every patient they see! In marketing, there are also ratings of consumer products made by experts on the judging panels of consumer magazines such as *Consumer Reports* in the U.S.A. and *Choice* in the U.K. and Australia.

For the sake of accuracy—and also credibility—the experts must be qualified in the *substantive field*. Qualified experts would include, for example, financial market analysts with an MBA degree, certified professional clinical psychologists, and other social scientists who have contributed demonstrably to theory and knowledge in the field. In clinical psychology, there has been a debate raging for decades about whether *psychiatrists* are appropriately qualified to treat psychological disorders because, although they have medical degrees, psychiatrists typically have no more than 2 years training in clinical psychology. However, this debate has abated in recent years with the emergence of impressively effective pharmacological remedies for many mental illnesses, remedies, which have unfortunately been overgeneralized (and abused) as solutions to everyday coping problems.

I have a bone to pick with health reporters in the media passing themselves off as experts. I read a lot of these reports and I commented on health-promotion research in the final chapter of the marketing communications book by Rossiter and Bellman (2005). Every health reporter whose reports I have seen, read, or listened to has three problems: the health reporters are technically incompetent to evaluate social science reports; they never put the report in "context" (much as I hate that clichéd word, it's appropriate here); and they never follow up on any of their reports. A scandalous recent case of health misreporting concern breast-cancer screening. More attention is given to the opinions of celebrities than the opinions of scientists—witness, for instance, the prominence given to singer-actress Olivia Newton-John's enthusiastic endorsement of breast-cancer screening (*The Australian*, November 20, 2009, p. 3). In the same newspaper a few months later, the health reporter headlined a study with the announcement that breast-cancer screening—mammography—is useless (*The Australian*, March 25, 2010). Health reporters are incapable of evaluating the research they report on. They raise needless fears as well as false hopes. Health reporters are dangerous to public health. Health research *academics* have been deafeningly silent about the hazardous state of health reporting (see, for example, Chapman's benign review of health reporting on TV, which was published in the *Medical Journal of Australia*, December 11, 2009). Okay, Rossiter, settle down and get back to the text.

The C-OAR-SE approach to measurement relies on EXPERTS as the rater entity. The difference is that the experts using C-OAR-SE do *not* need to be qualified in the substantive field in which the measurement is being conducted. And, if you've read this far, you will see that they certainly should not be psychometricians! In C-OAR-SE, the validity of the measure is based entirely on a semantic content comparison, namely, semantic identity—or as close as possible to identity—between the construct, *C*, and the measure, *M* (see chapter 1). The experts need only be qualified in the *colloquial language* of the measure. (I read recently that English is the hardest language to learn to read, followed by French, both taking at least 2 years of study, versus several *months* for Italian or Spanish; see Smith 2009. It takes years of living in the country, however, to learn the colloquial language, especially that of people from other social classes.) As an avid reader of social science journals over the past 40 years, and also of innumerable proprietary practitioner reports, my expert opinion is that the pool of sufficiently highly literate experts who write in the journals is declining worldwide. The leading U.S. journals are an exception; in those, you almost need to have a *second Ph.D. in English* to get your manuscript published! Many foreign-born academics I know use an English expert ghostwriter to help them get papers into the top U.S. journals. Australian-born, I employed a U.S.-born freelance copy-editor for my 1987 and 1997 textbooks, mainly to capture the idioms of the books' largest market.

5.3 Coders as the Rater Entity (CODRAT)

CODERS are a vital part of most measures (and, therefore, of most data). In *qualitative research*—which is the most frequent type of *applied* research, by far, in organizational behavior, sociology, clinical psychology, and marketing research—the coding or interpretation should be performed by *one* expert (see chapter 8).

In contrast, in *content analysis*, multiple trained coders are required. Rust and Cooil (1994) provide a useful table for estimating *how many* coders will be needed to attain high inter-coder agreement.

For content analysis—and for coding of open-ended questions in surveys or experiments—the CODRAT rater entity must be *trained* content-analysis coders. They must have successfully completed a thorough pretest of the researcher's coding scheme. In academic research reports, including articles in the major journals, the coder-training phase is often carelessly handled or even skipped entirely. However, coder training *is* important—much more important than the statistical method used—because the data are at stake. (The coding scheme and its instructions is another essential element that is almost never reported. It should be reported *verbatim*, in an appendix.) With no or poor coder training, severe measure-distortion errors (D_{m} in the new true-score model of chapter 1) are likely to flow through to the coding measure and thus contaminate the scores.

5.4 Managers as Group Rater Entity (MANRAT)

MANAGERS are the rater entity for many important constructs in the fields of organizational behavior, management (of course), and marketing. Many theories in these fields compare the views of MANAGERS as the rater entity with those of EMPLOYEES or of CUSTOMERS as the rater entity. I gave examples in earlier chapters—SERVICE QUALITY was one.

As I've said many times throughout this book, an object rated on an attribute becomes a different *construct* with a change in rater entity. O.B. researchers frequently make the mistake of using EMPLOYEES to rate what should be MANAGERS' perceptions. This mistake was made recently, for example, by Edwards and Cable (2009) in a study published in the *Journal of Applied Psychology*. The researchers claimed to have measured the "congruence" of CORPORATE VALUES as rated by THE COMPANY'S TOP MANAGEMENT and as rated by THE EMPLOYEE by asking EMPLOYEES to rate both! They took no record of the actual CORPORATE VALUES the TOP MANAGERS were trying to instil. A similar mistake was made in a study by Liao, Toya, Lepak, and Hong (2009) reported in the *Journal of Applied Psychology*. These researchers asked EMPLOYEES to rate their own service performance—instead of asking the employees' SUPERVISORS to do so.

When MANAGERS are the rater entity, it is very important to specify the job position—the *level*—of managers that will become part of the construct. The researcher should aim to identify and sample the views of the ultimate *decision-maker* (or group of ultimate decision-makers). An interesting example of this, from my consulting experience, is the question of who is the ultimate decision-maker for approving the company's advertising campaigns—not just the company's advertising strategy but everything—down to the micro-details of its ads, including the slogan or headline, the creative idea dramatizing the company's or its products' key benefit (see Rossiter and Bellman 2005), and the selection of a spokesperson or indeed whether or not a spokesperson is used at all. One would think that it would be the marketing management team, or these days the marketing communications management team, who would make these decisions. But that is hardly ever true—it's the CEO who decides! Even at my university it is the Vice-Chancellor, Gerard Sutton (the university's President in U.S. terminology), who must approve all the main communications for the external public's consumption. Fortunately, Vice-Chancellor Sutton is a natural marketing wizard.

Researchers tend to interview *any* managers who make themselves available to answer what are often ridiculously long multiple-item questions. Top managers are hardly ever reached, except by commercial research firms that employ several highly skilled *executive interviewers*, who can get appointments denied to academics or student interviewers.

5.5 Consumers as Group Rater Entity (CONRAT)

I refer to this group rater-entity as CONSUMERS, but I also mean to include THE PUBLIC as in public-opinion polls reported in the media and in the leading realist journal on measurement, which is the *Public Opinion Quarterly*.

When CONSUMERS are the rater entity, the samples from which the measures are taken must be (a) *representative*—to ensure *content validity* (CV); and (b) as *large* as can be afforded—for *precision-of-scores reliability* ($R_{\text{precision}}$). Here's an interesting example of the importance of the representativeness requirement for content validity. In late 2009, the media reported that "the world's most influential man" was none other than Don Draper (the lead character in the cultish TV series *Mad Men*). This sensational news delighted me as I am an avid fan of *Mad Men*, mainly because of the TV program's script's devotion to the thinking that led to many classic ad campaigns (and the beautiful Grace Kelly lookalike actress, January Jones, may have kept me watching, too!). I was working in the U.S. as an advertising and marketing research consultant towards the end of the *Mad Men* era and was amazed at the accuracy of the stereotypes of advertising agency personnel and client characters in the program. Most valuable in the program's episodes, however, are the numerous examples of how major creative ideas in truly classic campaigns were generated and "sold" (most often by the Machiavellian Draper character, played by the also beautiful Jon Hamm) to clients. I require that all my advertising students, and my Ph.D. students working on advertising thesis topics, watch the DVDs of *Mad Men*. But as to Don Draper being "the world's most influential man," the validity of this measurement result obviously depends on the *rater entity*. The CONSUMER raters in the survey consisted of subscribers who link to the website AskMen.com. They are hardly a representative rater entity! This is further underscored by the fact that the number 2 "most influential man" was Usain Bolt (heard of him?) and number 3 was the more defensible choice for number 1, U.S. President Barack Obama, but whose public effectiveness ratings at the time of writing are slipping fast. The unfortunate Michael Jackson came in as the world's sixth-ranked most influential man—a ranking that most musicians will believe—but ahead of Steve Jobs? The construct should, therefore, have been *defined* as THE WORLD'S MOST INFLUENTIAL MAN AS PERCEIVED BY MOSTLY NAIVE SUBSCRIBERS TO THE WEBSITE ASKMEN.COM IN SEPTEMBER 2009.

When CONSUMERS are the rater entity—in their role *as* consumers—it is also worth making two further distinctions in the rater-entity definition. These two distinctions were pointed out earlier. The first is to distinguish Prospective Customers, Current Customers, and Lapsed Customers because their respective perspectives are practically—and theoretically—important. Second, for advertising research in particular, finer distinctions of target-audience types should be made in terms of their *brand loyalty* (see Rossiter and Percy 1987, 1997, Rossiter and Bellman 2005).

Another important distinction for the CONSUMER rater entity is between THE PUBLIC and COLLEGE STUDENTS. Every social scientist should read and carefully study the meta-analysis by Peterson, Albaum, and Beltrami (1985) of effect sizes of the direction and magnitude of the relationships between measures of common social science constructs when the rater entity is, respectively, COLLEGE (UNIVERSITY) STUDENTS and REAL-PEOPLE CONSUMERS—and I use the latter label advisedly having personally done heaps of research with both groups! In Peterson et al.'s (1985) meta-analysis of consumer behavior experiments, 52% of the 118 studies were based on college students and 48% on noncollege students, mainly members of the general public. (In psychology experiments the proportion of studies based on college students is astronomically higher, of course.) Two highlights of Peterson et al.'s findings are (a) the effect sizes (as measured by omega-squared, ω^2), on average, were approximately *44% larger* for *REAL-PEOPLE CONSUMERS* (.13) than for COLLEGE STUDENTS (.09); and (b) the effect sizes, on average, were approximately *30% larger* in *FIELD STUDIES*, based primarily on real consumers (.13), than for LABORATORY STUDIES, based almost entirely on college students as the rater entity (.10). The latter finding, in case you didn't realize it, means that the *same theory*—yours maybe—is roughly 30% more likely to be rejected if the rater entity is COLLEGE STUDENTS! For example—and this is relevant for researchers in psychology—a subsequent meta-analysis by Peterson (2001) revealed that:

- EMPATHY predicts PROSOCIAL BEHAVIOR for college students ($r = .30$) but *not* for the public in general ($r = .09$)
- NATIONAL ORIGIN means much more to the public ($r = .28$) than to college students ($r = .05$)

- The ATTITUDE-BEHAVIOR relationship is lower for college students ($r = .34$) than for the public ($r = .48$)

Indeed, I am sorely tempted to propose another, related, classification of the "CONSUMER" or "PUBLIC" rater entity—this time in the role of research participant—as LABORATORY STUDY versus FIELD STUDY. This is based not only on Peterson et al.'s (1985) and Peterson's (2001) meta-analyses, but also on another meta-analysis by Bell (2007). Although I am suspicious of meta-analyses in general, hers is pretty good. Bell analyzed the "Big Five" personality traits, adding a sixth, COLLECTIVISM, and also the abilities of GENERAL INTELLIGENCE and EMOTIONAL INTELLIGENCE, where the WORK TEAM—groups of individuals who worked together on various tasks—was the rater entity. Bell separated the studies into LAB STUDIES and FIELD STUDIES, which I said could probably qualify as another distinction between types of CONSUMER (or PUBLIC) rater entity. As in Peterson et al.'s (1985) meta-analysis, Bell (2007) found marked and significant differences in the correlations between the various predictors and the criterion variable according to whether the research was conducted in a *laboratory* setting—mainly with college students—or in a *field* setting, with a sample of the broader population. The correlations for lab and field studies from Bell's article are shown in Table 5.2.

Table 5.2 Predictors of work-team performance under laboratory and field conditions (from Bell 2007)

	Correlation with performance		
Predictor	Lab	Field	Difference
Conscientiousness	.04	.30	+.26
Agreeableness	.03	.31	+.28
Extraversion	.06	.15	+.09
Neuroticism	.03	.06	+.03
Openness to experience	.00	.20	+.20
Collectivism	.00	.35	+.35
Emotional intelligence	.20	.10	–.10
General mental ability	.31	.18	–.13

The point to be taken away from these studies is that theories—not just particular hypotheses—are going to be accepted or rejected based on the *rater entity*. This is why it is so important to specify the rater entity as part of the construct.

5.6 Individuals as Rater Entity (INDRAT)

The preceding four types of rater entity, for which I have used the abbreviations EXPRAT, CODRAT, MANRAT, and CONRAT, are each analyzed as a *group* rater-entity. The exception is EXPRAT when only one expert is making the ratings or assessment; this is the case for *qualitative research*, where one expert analyst is all that is needed, as explained in chapter 8 of this book (and it *should* be the case when designing *measures*—in which case you need to become expert in C-OAR-SE theory).

The final classification of rater entity is the *INDIVIDUAL* as rater entity (INDRAT). The individual is the rater entity for all *individual-difference* constructs—such as ABILITIES, PERSONALITY TRAITS, VALUES, and DEMOGRAPHICS, to use generic labels for what are really *sets* of constructs. These sets of constructs have the same rater entity (THE INDIVIDUAL), and the same object (THE SELF, or for those who can't face reality on *Facebook* or *Twitter*, one's AVATAR, or IDEAL SELF, though I take this snide remark back because I just came across a new study in *Psychological Science* that clearly disproves the "idealized virtual-identity" hypothesis for *Facebook* users—see Back, Stopfer, Vazire et al. 2010), and differ only in the *attribute* (the specific ability, personality trait, value, or demographic characteristic). Each individual produces a score on the particular construct and these scores should be analyzed at the *individual level*.

Although I have referred to measures for the "social" sciences throughout this book, the fact is that most theories in psychology and in consumer behavior are *individual-level* theories. However, there *are* some group-level theories in what is correctly named *social psychology* (here you should read G.W. Allport's excellent explanation and defense of "crowd" and "group mind" theories in his classic chapter on the history of social psychology in the 1985 edition of *The Handbook of Social Psychology*), and of course in the fields of organizational behavior and management. Group-level theories require the *group* to be used as the unit of analysis (e.g., WORK TEAMS in Bell's study summarized above).

When INDIVIDUALS are the rater entity, the scores should be analyzed not by using the conventional across-persons analysis (group means, group standard deviations, and group correlations) but *within-persons* analysis.

Within-persons analysis is the analysis of each individual's scores across *occasions*—that is, over time (but note that "time" is *not* a causal construct because in itself it cannot cause anything). This is the basis for calculating stability reliability, $R_{\text{stability}}$, which, after content validity has been established, is the second essential requirement of a measure (see chapter 2 on validity and reliability). For example, Coombs' (1964) "pick any" measures of BELIEFS (unipolar attribute-beliefs) and ATTITUDES (bipolar evaluative-beliefs), as explained in the next chapter on item types, are highly stable on a short-interval retest at the *aggregate or group* level, but are shockingly unstable at the *individual* level. As summarized by Rungie, Laurent, Dall'Olmo Riley, Morrison, and Roy (2005) beliefs and attitudes measured by "pick any" are, on average, only about 50% stable at the individual level, even when remeasured just 1 week later! What this means is that only 50% of the individuals who endorse the belief or attitude at time 1 will do so again a short time later at time 2 *and* that half of the endorsements of the belief or attitude at time 2, are contributed by individuals who did not endorse it at time 1! (See Rungie et al. 2005, Dolnicar and Rossiter 2008, and Rossiter et al. 2010, for these findings.)

Psychology and consumer behavior journals are replete with individual-level theories that are "proven" only by erroneous *across-persons* analysis, theories which very likely would not hold up under appropriate individual-level analysis. This work remains to be done and it is going to provide a valuable career path for many young psychologists and many young marketing academics—really worthwhile career opportunities that will be missed if they do not make the effort to understand and apply C-OAR-SE.

I am going to close this last section of the chapter with one "zinger" here and that is that all the major THEORIES OF LEARNING—my main coursework as an undergraduate in psychology and, in my opinion, the most important contribution of all to social science theory—were derived from the intensive study of *individual subjects* (i.e., INDRATs). Thorndike (1898) invented the original S–R "connectionist" theory of learning by studying the attempts of a few cats to escape from a box. Skinner (1935) studied two or three rats and later (1948) two or three pigeons. Freud studied one individual at a time—actually two, the patient and *himself*, since his theories were derived primarily by introspection (see Locke's 2009 article titled "It's time we brought introspection out of the closet" and also chapter 8 in the present book, on qualitative research). Piaget's theory—which is not a learning theory and in fact is precisely the opposite, a genetic maturation theory—was generated from observations of a few individuals (his own children).

5.7 End-of-Chapter Questions

(5.1) Based on what you've read in this chapter, write out arguments—in your own words—(a) in favor of including the rater entity in the definition of the construct and (b) against its inclusion as is the usual practice (see Diamantopoulos 2005). (7 points)

(5.2) I made the observation in the chapter that researchers' literature reviews often reveal that theory tests have received "mixed results." I also made the argument that a major reason for this is that the constructs in the theories involve different rater entities. Search through recent journal articles in your field for articles on a major topic and look at the end of the "introduction" section for the words "mixed results" or "mixed findings" (easy to do now online). Then look back at some of the original studies (which is always good practice, contrary to the common practice nowadays of citing studies in articles that the citer—you—has never read) and see if you can identify different rater entities as being responsible. Bonus question: If it wasn't different rater entities that produced the "mixed results," then what was it? There must be some logical explanation because a conclusion of mixed results or mixed findings is never acceptable. (7 points for the main question and up to 5 points additional for the bonus question)

(5.3) One classification of the rater entity is called the CONSUMER rater entity, a term that mainly applies to constructs in theories of marketing (MANAGERS are the other type of rater entity that is also relevant for many constructs in marketing, not just in management). What was the alternative label for this classification in public-opinion research? (1 point) What did I suggest as two other important distinctions for this rater entity, and why? (3 points for each)

(5.4) In constructs where INDIVIDUALS are the rater entity, how should their scores on the construct be analyzed, and why is this vitally important for testing theories that involve individual-difference constructs, that is, individual-difference "variables?" (7 points) Bonus question: Find and briefly describe and defend one example of a group-level theory in any social science field. (5 points)

Chapter 6
Selection of Item-Type and Answer Scale

Fraternité, Egalité, Liberté
—Nominally listed attributes of France's national motto

The sheep-like fraternity of social-science researchers is apparently under the impression that all item types are created equal. Social-science researchers seem to believe they are at complete liberty to choose what type of item to use to operationalize the measure. This *laissez-faire* state of affairs is evidenced by researchers' arbitrary and unjustified selection of Likert measures (most popular), Semantic Differential measures (second most popular), and Unipolar measures (occasionally) to represent almost any construct. They couldn't be more wrong about item types all being equal and interchangeable.

An *item* brings together a question part and an answer part. As we saw in Chapter 2, the question part of the item must be rationally argued to have high item-content validity (CV_{item}) and the answer part must be demonstrated to have high answer-scale validity ($\text{CV}_{\text{answer}}$). The purpose of the present chapter is to explain how both types of content validity can be achieved by selection of an *item type* and *answer scale* appropriate to the type of construct to be measured. There is quite a lot of innovative, and I hope useful, theorizing in this chapter.

After reading this chapter you should be able to:

- Appreciate the numerous variables hidden in an item
- Understand why a particular item-type is most valid for measuring, respectively, each of the eight major types of construct in the social sciences
- Then, afterwards, abstain from using this book simplistically as a "cookbook" for selecting items—because you yourself should now be able to make an intelligent selection

Before we begin the chapter, I will take the opportunity to review the criteria for evaluating measures that were outlined in Chapter 2 on validity and reliability. Unique to the C-OAR-SE theory of measurement, the criteria are *hierarchical*. The hierarchical criteria are summarized in Table 6.1.

J.R. Rossiter, *Measurement for the Social Sciences*,
DOI 10.1007/978-1-4419-7158-6_6, © Springer Science+Business Media, LLC 2011

Table 6.1 Hierarchical criteria for evaluating measures

Essential
1. High item-content validity
×
2. High answer-scale validity
↓
Derived essential
3. High (80% or higher) stability reliability on a short-interval retest
↓
Desirable (for predictor constructs)
4. Good (close to R_{pop}) predictive validity

There are two *essential* and multiplicative criteria, both of which concern *content validity*. The measure must be demonstrated to have both high item-content validity ($\mathrm{CV}_{\mathrm{item}}$) and high answer-scale validity ($\mathrm{CV}_{\mathrm{answer}}$). Accordingly, these are the two joint criteria that are the focus of item-type selection in this chapter. If the items and the answer scale are highly content-valid, then the next criterion, stability reliability ($R_{\mathrm{stability}}$), will automatically follow (hence it is *derived* essential). Good—close to R_{pop}—*predictive validity* is thereafter *desirable* for measures of the predictor variables. For the criterion or outcome variable—the scores of which are being predicted—content validity logically has to suffice. Nunnally (1978) but, I note for marketing readers, not Churchill (1979) makes this latter observation.

6.1 A Model of Item Variables

The model presented in Fig. 6.1 is not only a model of item variables, it is also a theoretical model of the item-based sources of measure distortion. (Measure distortion is the D_{m} term in Chapter 1's new true-score model, $O = T + D_{\mathrm{m}} + E_{\mathrm{r}}$.) There are at least six sources of (i.e., variables that cause) measure distortion. These variables will be illustrated in the measures discussed in the chapter.

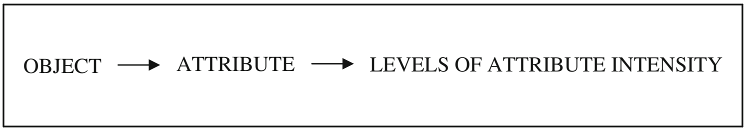

Fig. 6.1 Model of item-based sources of measure distortion of the true score.
Sources of distortion:
(1) Iconicity of object representation in the question part of the item
(2) Accuracy of the description of the attribute in the question part of the item
(3) Correct modality of the intensity levels in the answer part of the item
(4) Accuracy of the verbal or numerical descriptors of each level of intensity in the answer part of the item
(5) Match of the number of attribute intensity levels with the typical rater's discrimination of levels
(6) Clarity with which the answer levels indicate the polarity of the attribute

6.2 Attribute Beliefs (or Perceptions)

ATTRIBUTE BELIEFS—the rating of the focal object on an attribute—are by far the most common construct in the social sciences. This becomes obvious when it is realized that all *abstract* constructs (those made up of an abstract object, an abstract attribute, or both) are simply aggregations of ATTRIBUTE BELIEFS. This fact should be self-evident but if readers need further explanation of this obvious fact, see the discussion in Rossiter and Bergkvist (2009). Remarkably, when you are measuring an ATTRIBUTE BELIEF, no reviewer ever retorts that you should use a multiple-item measure! This ubiquitous construct rightly escapes the psychometricians' argument that multiple items are *always* necessary.

In my field of marketing, the really quantitative academics grandiosely label themselves as "marketing scientists"—whereas, as explained in Chapter 8, it is the *qualitative researchers* who are the true scientists in marketing. Marketing scientists quite confusingly call attribute beliefs "PERCEPTIONS." Quite by accident, this marketing science terminology points to an important distinction in the type of *rater entity* used to measure beliefs. You will remember from Chapter 4 that there are two types of concrete attribute. Concrete *perceptual* (CONCPERC) attributes—the marketing scientists' "perceptions"—can be self-rated, and so the rater entity is the INDIVIDUAL. The other type of concrete attribute is a concrete *psychological* (CONCPSY) attribute and, as explained in Chapter 4, such attributes cannot be self-rated; rather, the rater entity for concrete psychological attributes must be one or several EXPERTS.

At this advanced stage in the book, it is appropriate to identify two further types of belief, both of which are CONCPERCs and both of which have the INDIVIDUAL as the rater entity. They are worth distinguishing from normal CONCPERCs because their types of *answer scale* not only differ from that of a normal CONCPERC, but also differ from each other. I will call them a SENSORY CONCPERC and a SERVCOMP CONCPERC.

> It's huge!
>
> – A moderate compliment in Milleniumspeak but the ultimate compliment (and ultimate white lie) in Adultspeak . . .

What may be called SENSORY CONCPERCs are the type of attribute used in psychophysical research to measure ACUITY in any of the five senses: that is, acuity in visual, auditory, olfactory, gustatory, tactile, and kinaesthetic (sense of movement) attributes. To measure acuity on a SENSORY CONCPERC attribute, the researcher needs to use an answer scale that has "ratio" properties, meaning that there is a true psychological zero at the left or lower end of the answer scale and that the points along the answer scale are at equal psychological intervals from one another and in relation to the true zero (e.g., 2 is twice the intensity of 1, 3 is three times the intensity of 1, 4 is four times the intensity of 1, and so forth, and, if the points are verbally labeled, "maximum" intensity is twice as strong as "moderate" intensity). To give an everyday example of ratio measurement: current world record-holding sprinter, Usain Bolt, runs the 100-m dash averaging 46 km/h, or 29 mph for you Stone Age types, which is about 1.4 times faster than an elephant could, at 32 km/h; a cheetah, at 112 km/h, is more than 2.4 times faster than Bolt; and a snail, at 0.01 km/h gets down near enough to true zero (Salvado 2009).

A ratio answer scale is often called a *magnitude* answer scale and the most accurate and the fastest to administer—training or "calibration" time being a major inconvenience with magnitude-estimation methods—is Bartoshuk and colleagues' (2004) generalized labeled magnitude scale (abbreviated as gLMS) which was developed from Green and colleagues' (1993) LMS measure.

Psychologists and many general readers may recall that Stanley Milgram used a ratio or magnitude scale in his famous and nowadays-controversial "obedience to authority" experiments conducted in the U.S. city of New Haven, near Yale University, in the early 1960s (Milgram 1963). In Milgram's original study, eight in every ten adult participants recruited off the street in New Haven continued to administer what they fully believed was an electronic shock of at least 150 V to another person—merely on the experimenter's instruction! What is even more alarming, all but a tiny percentage of these individuals continued administering (fake but believed) shocks up to the maximum, which was clearly marked as 450 V. In a modern-day replication, the results were only slightly less shocking. Burger (first reported in Mills 2009, and later in Burger 2009) found that seven in every ten adults today were willing to continue beyond 150 V—despite, as in Milgram's 1961 experiment, hearing the "subject" crying out in apparent pain and saying "I have a heart condition. Let me out of here. I no longer wish to be in this experiment!" (though the experimenter, Milgram in a white lab coat, aided and abetted this by saying, "Go on, please, the shocks are painful but they are not harmful. There will be no permanent tissue damage . . ."). For Burger's replication study, a modern-day ethics committee stricture prevented continuation beyond the (fake) 150 V but the replication of this result, too, if it were allowed, would doubtless be very close to the original. The aftershocks suffered by Professor Milgram as a result of the *furore* over his methodology, fake though the shocks were, were all too real. According to Slater's (2004) floridly written biographical chapter, Milgram had five heart attacks and died at CCNY (Yale got rid of him during the scandal) at age 51.

In some cases, however, a relative rather than an absolute measure of a SENSORY CONCPERC is sufficient. When Cadbury unwisely substituted palm oil for cocoa butter in its chocolate, a decision made by the CFO no doubt, the company lost millions of consumers, perhaps forever, who made a *simple difference* discrimination (see "Good oil on Cadbury" 2010).

The other subtype of CONCPERC attribute is what I am going to call a SERVCOMP CONCPERC, where this label refers to a component attribute of the overall attribute of SERVICE QUALITY, and the object is a RETAILER or an E-RETAILER. The SERVCOMP attributes are most validly measured—especially for managerial diagnostic purposes—by using a "behavioral categories" answer scale in which the options or answer categories reflect meaningful (and concrete) levels of performance on that attribute. My article in *Service Science* (Rossiter 2009a) provides numerous examples of behavioral-categories answer scales (each item measuring a single attribute-belief) and the behavioral categories differ for each item, unlike with other answer scales.

Getting back to the usual measure of an ATTRIBUTE BELIEF, which is a "doubly concrete" construct (consisting of a CONCOB and a CONCPERC), I here introduce a radical recommendation, which is to measure beliefs using what I have called *DLF IIST Binary* items (see the working paper by Rossiter et al. 2010). The DLF IIST Binary measure is exemplified in Table 6.2 for measuring ATTRIBUTE BELIEFS about, or PERCEPTIONS of, brands of laundry detergent. DLF IIST Binary measures are "doubly level-free" (DLF), which means that there is no level of the attribute mentioned in the question part of the item, or in the answer part. The "IIST" in the label is explained shortly. DLF IIST Binary measures are also "forced choice," unlike Coombs' (1964) "pick any" (or, technically, Affirmative Binary) measure, which is by definition "free choice."

Table 6.2 DLF IIST Binary measures of laundry detergent attribute beliefs

For each of the laundry detergents pictured below, please click the "Yes" box if that description applies and click the "No" box if it does not apply.

TIDE [color picture of pack]		
Cleans	□ Yes	□ No
Whitens	□ Yes	□ No
Brightens	□ Yes	□ No
Cold-water washing	□ Yes	□ No
Environmentally safe	□ Yes	□ No
Value	□ Yes	□ No
OMO [color picture of pack]		
Cleans	□ Yes	□ No
Whitens	□ Yes	□ No
Brightens	□ Yes	□ No
Cold-water washing	□ Yes	□ No
Environmentally safe	□ Yes	□ No
Value	□ Yes	□ No

DLF IIST Binary measures of attribute beliefs, remarkably, turn out to be the *most stable* of all belief measures—and also the most *predictively valid* in terms of producing the most stable regression coefficients in multivariate predictive regressions.

DLF IIST Binary measures probably work so well because, when answering "Yes" or "No" about a particular object in terms of possessing a particular attribute, the rater plugs in his or her *individually inferred satisfaction threshold*—hence "IIST"—of intensity before answering. If the brand's perceived attribute performance is at or above the individual's threshold the individual answers "Yes" and if below, "No." Previous theorists, following Coombs' (1964) revelation that *all* rating scales consist of a series of binary choices, have conceptualized binary (one choice between two alternatives) answer scales as involving an individual-level threshold (see Lord 1980, Ferrando and Anguiano-Carrasco 2009). In Lord's theory, however, the threshold applies only to *ability-test* items; the threshold is the level of *item difficulty* at which the individual can make a correct answer. This "item-response theory" approach—really the "item-response difficulty theory" approach—has been applied by later theorists to BELIEFS (actually to bipolar beliefs, or ATTITUDES) and BELIEFS, of course, do *not* vary in difficulty. Also, the IRT or rather IRDT theorists did *not* postulate an information-processing explanation as I do here. Rather, they "explained" the threshold in a circular manner as simply "the transition from the tendency to deny the item to the tendency to endorse it" (this from Ferrando and Anguiano-Carrasco 2009, p. 502). As previously mentioned, I theorize—and I'm the first to do so as far as I am aware—that there is an *individually inferred satisfaction threshold* operating for this measure. The individually inferred threshold represents the perceived level of the attribute in terms of which the rater says (subvocally and likely subconsciously) in relation to the first item in the table, for example, "Does TIDE detergent clean well enough for *me*?" If so, the rater clicks the "Yes" box and if not, the "No" box.

Watch for impressive results for DLF IIST Binary measures of BELIEFS in forthcoming articles with colleagues Sara Dolnicar and Bettina Grün.

6.3 Evaluative Beliefs (or Attitudes, Plural)

ATTRIBUTE BELIEFS, or PERCEPTIONS, are *unipolar* and can extend in levels of intensity of the attribute only from zero to some *positive* amount. In contrast, EVALUATIVE BELIEFS—or ATTITUDES (plural) in the traditional meaning of the term—are *bipolar* in that the ratings differ in direction (negative, positive) and then vary two ways in *intensity* (a negative amount through to zero, and zero through to a positive amount).

The most popular item-type for measuring EVALUATIVE BELIEFS or ATTITUDES in the social sciences is the so-called "Likert item" in which the polytomous (multipoint) answer scale ranges from "Strongly Disagree" to "Strongly Agree." (I say "so-called" because Likert in his classic 1932 monograph never actually used the attribute of DISAGREEMENT–AGREEMENT in the several answer scales in his studies. His closest wording was "Strongly Disapprove" to "Strongly Approve," using the attribute of DISAPPROVAL–APPROVAL, on p. 14 of his article. But most researchers have never bothered to find and read Likert's original article—despite citing it.) Many researchers in marketing and psychology naively refer to *any* multipoint answer scale—bipolar or unipolar, for Heaven's sake, where Rensis Likert surely is now—as a "Likert" answer scale.

The second most popular item-type for measuring EVALUATIVE BELIEFS or ATTITUDES is the so-called Semantic Differential item, in which a negative adjective appears at one end of the answer scale and a positive adjective appears at the other, so that the attribute descriptors are "polar opposites" (see especially the definitive article on semantic differential items by Heise (1969)). The term "semantic differential" is a misnomer because this term refers to a *factor-analytic structure*—the Evaluation, Potency, and Activity orthogonal (independent) factors—discovered by Charles Osgood and reported in the book by Osgood, Suci, and Tannenbaum (1957) which at one time was the most-cited book in all of social psychology. Nevertheless, I have to yield to the weight of common usage of "semantic differential" as referring to a type of *item*. Semantic Differential items, like Likert items, are bipolar in direction and have accordingly two *continua* of intensity, with zero intensity in the *middle* of the answer scale.

However, appallingly ignored—with devastating consequences for the measurement of true scores—is Peabody's (1962) article in *Psychological Review* (along with *Psychological Bulletin*, this is the most authoritative journal in psychology, so any self-respecting psychometrician can hardly claim to have missed the article). For four diverse constructs, Peabody demonstrated that most of the variation in scores obtained from bipolar belief items is accounted for by simple *direction*, with very little of the variation accounted for by *intensity*. In Table 3 in his article (p. 71) the correlations between the total scores and the direction scores (which were of course binary, just negative or positive) across eight data sets, for four different but very typical multiple-item constructs, each measured on a US sample and a U.K. sample, ranged from $r = .87$ to .97, for an overall average of $r = .94$! For r^2 fans, this average translates to $r^2 = .88$, which signifies that *only 12% of the variance on average* in scores from bipolar polytomous answer scales is due to bipolar intensity judgments! Yet, you would be hard-pressed to find a psychometrician today who does not believe that 100% of the variation in belief or attitude scores is due to intensity judgments, which is the fundamental rationale for using multipoint, "polytomous" answer scales as in Likert and Semantic Differential items.

Also, likely to be ignored is the finding by Swain, Weathers, and Niedrich (2008) that negatively worded Likert items—which often make up approximately half of a multiple-item battery—are so confusing to raters that they usually produce a second (and entirely artifactual) factor in measures of an allegedly unidimensional construct.

In terms of the new true-score model (of Chapter 1), Likert items *and* Semantic Differential items are vulnerable to a substantial and serious extent to the five major forms of measure-induced distortion (D_{m}). These are: *midpoint evasion* (for Likert and Semantic Differential answer scales that employ an uneven number of categories, as is typical); *extreme(s) responding*; *response-order* distortion; *acquiescence* or "yea-saying;" and distortion due to *overdiscrimination* (too many levels of intensity provided in the answer alternatives). These measure-induced distortions produce unacceptably *unstable* ratings on a short-interval retest—found in an extensive study by Rossiter et al. (2010) and note that we only have to find that these measures *sometimes* fail in order to recommend against using them in general.

I allege that all the findings in the social sciences based on Likert items and Semantic Differential items are *suspect*—and this means the *majority* of findings!

This leaves the DLF IIST Binary measure as the *only content-valid measure* of bipolar EVALUATIVE BELIEFS, just as it is the most content-valid measure of unipolar ATTRIBUTE BELIEFS. With DLF IIST Binary items, the rater *can't* evade; there are *no* extreme responses possible; response order and acquiescence *don't* affect the scores (as Rossiter et al. 2010, found empirically); and the rater obviously *can't* overdiscriminate.

For measuring bipolar EVALUATIVE BELIEFS, the binary answer categories in the DLF IIST Binary measures can be either "Yes" or "No" (exactly as for unipolar ATTRIBUTE-BELIEF measures) or they can be in "condensed Likert" format, that is, "Disagree" or "Agree" (wherein the bipolar evaluative beliefs are traditionally called ATTITUDES). Both answer wordings are illustrated in Table 6.3. These items are from two studies that I have conducted with my colleagues Sara Dolnicar and Bettina Grün, one for consumers' evaluative beliefs about brands of fast-food restaurants, and the other for the public's perceptions of, or attitudes toward, the candidates in the last election in Australia for the office of Prime Minister.

Table 6.3 DLF IIST Binary measures of bipolar evaluative beliefs or attitudes

Fast-food restaurants		
For each of the fast-food restaurants named below, please click the "Yes" box if that description applies and the "No" box if it does not apply		
McDonald's is		
Yummy	□ Yes	□ No
Convenient	□ Yes	□ No
Value	□ Yes	□ No
Unhealthy	□ Yes	□ No
Fattening	□ Yes	□ No
Australian candidates for the office of Prime Minister in the 2007 national election		
For each of the politicians named below, please click the "Agree" box if that description applies and the "Disagree" box if it does not apply		
Kevin Rudd is		
Likeable	□ Agree	□ Disagree
In touch with voters	□ Agree	□ Disagree
Decisive	□ Agree	□ Disagree
Has a vision for Australia	□ Agree	□ Disagree
Understands the major issues	□ Agree	□ Disagree
Trustworthy	□ Agree	□ Disagree

For these bipolar evaluative or "attitude" attributes, note that the attributes are all *worded unipolar* in the DLF IIST Binary items. The actual attribute in the fast-food restaurant item is, in some cases, truly bipolar (e.g., YUMMY–YUK, CONVENIENT–INCONVENIENT, GOOD VALUE–BAD VALUE, and UNHEALTHY–HEALTHY) but is represented only by its *managerially meaningful pole* (i.e., YUMMY, CONVENIENT, VALUE, and UNHEALTHY). In these cases, the CONSUMER, as rater entity, is possibly but not necessarily endorsing the implicit antonym—or opposite pole—in answering "No." But if the opposite pole is relevant to consumer choice, it should be included as a separate item (e.g., a separate item might be "Healthy □ Yes □ No"). Another example is that the last attribute in the fast-food restaurant list, FATTENING, in the past *was* functionally *unipolar* but with the introduction of "healthy" food lines such as sandwiches and salads, this attribute may now have become bipolar like the others (i.e., FATTENING–SLIMMING, or FATTENING–DIETARY, or whatever descriptor consumers would typically use to describe the opposite of FATTENING). If so, the two separate attributes should be included. But note that in the FATTENING and SLIMMING cases the attributes would now have to apply to two separate attitude *objects*, the "regular" food line and the "healthy" one, so these two sub-objects would have to be incorporated in the questionnaire (for each restaurant that offers these alternatives). The same analysis could be invoked for the attributes in the items in the survey of attitudes toward Australian politicians in the foregoing table. Some of the attributes are obviously bipolar (e.g., LIKEABLE–DISLIKEABLE, in British spelling) whereas others are clearly unipolar (e.g., HAS A VISION FOR AUSTRALIA).

The use of unipolar attribute wording, even for bipolar attributes, solves a problem that is unique to *Semantic Differential* items, where bipolar and unipolar attributes are often *mixed* within the set of items (e.g., in an experiment by Rossiter and Smidts, 2010). The mixing of bipolar and unipolar attributes leads to measure distortion in *midpoint* ratings where, with the typically applied 1–7 enumeration, a score of 4 means "neutral" for a bipolar attribute and "moderate" for a unipolar attribute. Heise (1969) makes the anecdotal but plausible case that *all* Semantic Differential scales are *interpreted* bipolar, but I am yet to be convinced.

In several studies soon to be published, we have found that these DLF IIST Binary measures of what are usually bipolar EVALUATIVE BELIEFS—or ATTITUDES—are highly stable ($R_{\text{stability}} = 85\%$ for the fast-food restaurant attributes and $R_{\text{stability}} = 82\%$ for the politicians' attributes). High stability (80% plus) is a *derived essential* property of rating scales (see Table 6.1 earlier). Also, unlike Likert and Semantic Differential items, DLF IIST Binary items avoid the severe *multicollinearity* induced by the Likert and Semantic Differential items' polytomous answer scales (the multicollinearity is mainly due to *measure distortion* by spurious intensity in the answer scales, which produces what is known as common-method bias) when the BELIEFS are used as multiple predictors. Accordingly, DLF IIST Binary is the only type of measure that can produce *stable regression weights* in a multivariate prediction (of a relevant criterion variable such as PREFERENCE or BEHAVIORAL INTENTION).

In case you haven't realized the importance of these findings, I will state categorically (in the Kantian sense) here that DLF IIST Binary is the most significant breakthrough in belief and attitude measurement since Likert's classic contribution. Sadly but importantly, the breakthrough is a *requiescat* for Likert measures, and also for Semantic Differential measures.

The essential properties of DLF IIST Binary item-type for measuring both ATTRIBUTE BELIEFS and EVALUATIVE BELIEFS or ATTITUDES are worth emphasizing. (Why? Because BELIEFS are easily the most-measured construct in the social sciences.) The first essential property is that the items are *doubly level-free*: the attribute in the question part of the item must be worded level-free, with no qualifying adjectives or adverbs, and the answer alternatives must also be level-free, not going beyond the personally subjective binary division in the *intensity* of a unipolar attribute (recognizing again that bipolar attributes are *worded* unipolar and represented by one pole or by both poles as separate items, depending on the managerial relevance of the direction of the attribute). Go back and study both sets of items in Table 6.2 and you will see what I mean.

The second essential property of DLF IIST Binary items is that they have *forced-choice* answers. Regarding this second property, we have tested an alternative answer scale that "unforces" the choice by adding a "Don't Know" answer to the binary answer alternatives—as Rensis Likert did almost 80 years ago (1932, p. 14)—and we have found that this third option *worsens* stability. The "Don't Know" option is sometimes included in the answer scale because of the suspicion that raters may be unfamiliar with, or have little knowledge about, some of the *objects* they are being asked to rate. However, it is likely that the presence of the "DK" option prevents or disrupts the rater's automatic mental process of *inferring a threshold*.

To get around the problem of low-familiarity objects, raters should be instructed to make the ratings that they *expect* the object to have, based on its picture or name. This is a common procedure in commercial marketing research surveys because this expectation is what a consumer must act on in deciding for or against trying a new product. This procedure should also be used for *public opinion surveys*. It is of course true that some raters may *not* know, in the literal sense of "Don't Know," but it is highly unlikely that they have *no* attitude or a *neutral* attitude toward a major political figure or a major social or economic issue! This "attitude polarization" tendency in public opinion was noted as far back as 1930 by the philosopher Robert Thoulless in his classic book, *Straight and Crooked Thinking*, one of my favorite books because it is based on rationality, not empirics!

6.4 Emotions (Type 1 and Type 2)

EMOTIONS are enjoying a *renaissance* in the social sciences, replacing the previous fad known as the "cognitive revolution" but really representing a return to very old—and very great—theorizing in psychology by James (1884) in his article (in a philosophy—which at that time incorporated psychology—journal) titled "What is an Emotion?" This is another classic article that you should read, although James' misguided idea that emotions have an "organic" cause led indirectly to pointless physiological and now predominantly brain research on emotions, an avenue that was long ago disproved (in the famous Schacter and Singer 1962 *Gedanken* experiment.)

There are two very different kinds of emotions and they require very different types of items to measure them. Both kinds of emotion involve *concrete perceptual* attributes (CONCPERCs) and can only validly be measured by *self-report* (see especially Barrett 2004, but not her later publications in which she back-pedaled on this correct claim by also endorsing the use of physiological measures of emotions—caught up in the rush to do "brain research," the latest fad in the social sciences).

'Big-Time Sensuality'
(one of younger son and club-deejay
Stewart Rossiter's favorites)

– Composition by Björk

First there are TYPE 1 EMOTIONS (so-named by Rossiter and Bellman 2005), which are also called "core" emotions (Russell and Barrett 1999). There are probably just two type 1 emotions: AROUSAL, which is unipolar (though is often measured bipolar), and AFFECT, which is bipolar (NEGATIVE AFFECT and POSITIVE AFFECT). A highly valid and very efficient measure of both AROUSAL and BIPOLAR AFFECT is the 9 × 9 grid in which both emotions are rated simultaneously as the coordinates of a single point in the grid. This grid measure was invented by Russell, Weiss, and Mendelsohn (1989) and although these researchers wrongly conceptualized AROUSAL as bipolar—in which the logical negative pole would be "Dead" or perhaps "Cremated"—this simultaneous rating measure appears to work very well in practice. It delivers highly content-valid measures of the dual TYPE 1 EMOTION predictor variables. An adapted version is given in Fig. 6.2.

For measurement of the TYPE 1 EMOTION of AFFECT, emotion researchers should *not* use the widely used Positive Affect Negative Affect Schedule, or PANAS, measure (Watson, Clark, and Tellegen 1988), for which others have claimed to show that NEGATIVE AFFECT and POSITIVE AFFECT are *not* bipolar, but rather are completely independent, or "orthogonal" (an illogical notion). For example, the PANAS-style study by Tuccitto, Giacobbi, and Leite (published in *Educational and Psychological Measurement*, 2010) is badly flawed by its use of measures of TYPE 2 EMOTIONS of which five of the 10 "positive affect" emotions are not "affect" at all (INTEREST, ALERTNESS, ATTENTIVENESS, ACTIVITY, and STRENGTH). Instead, these descriptors obviously represent Osgood et al.'s (1957) *two* orthogonally and allegedly *nonevaluative* connotative meaning dimensions labeled Activity and Potency. You cannot validly measure TYPE 1 emotions with PANAS!

State of emergency is where I want to be

—From the composition "*Jöga*," by Björk
(older son B.J. Rossiter's co-favorite,
together with her composition "*Violently Happy*")

TYPE 2 EMOTIONS (Rossiter and Bellman 2005), or what Russell and Barrett (1999) called "prototypical" emotions, are completely different from TYPE 1 EMOTIONS. TYPE 2 EMOTIONS are specific *emotion-states* that are experienced all-or-none and must be measured binary. Following a 2,000-year-old idea proposed by the Stoic philosophers, specific emotions require "cognitive labeling" consciously by the rater and, therefore, *cannot* be validly measured by physiological recordings. This includes "neuroscience" or "brain-imaging"—a superficially objective but conceptually blunt methodology which is unnecessary, and indeed misleading, for measuring psychological *and* perceptual constructs.

The most common mistake in research on TYPE 2 EMOTIONS is to measure them with continuous or polytomous answer scales (see Bellman 2007, Rossiter

Mark an “X” in the square to show how you feel *right now*:

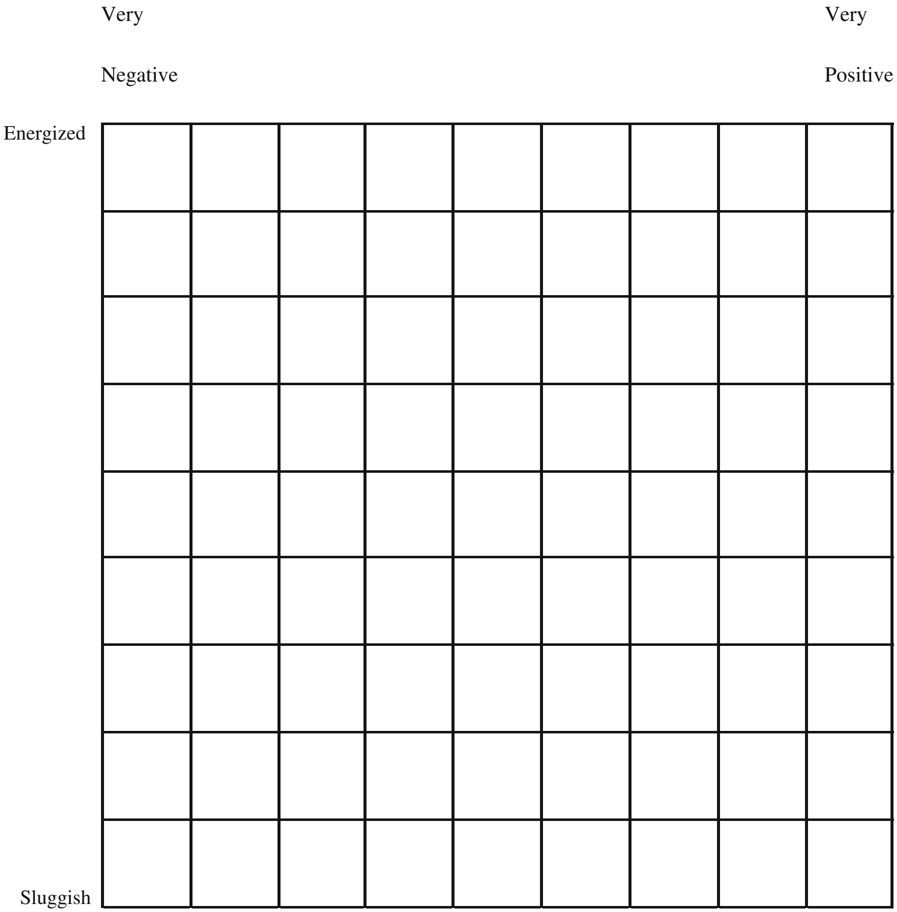

Fig. 6.2 Arousal-affect grid measure of type 1 emotions

and Bellman 2010). Type 2 emotions are *not* continuous in intensity. Below and sometimes above their inherent and specific level of intensity, it’s a *different* type 2 emotion. ROMANTIC LOVE at lower intensity is “I LIKE YOU VERY MUCH BUT . . .”and at higher intensity is OBSESSION. ANGER at lower intensity is just ANNOYANCE and at higher intensity becomes RAGE. See also, Sauter’s (2010) excellent article demonstrating that the Type 2 emotions have *discrete* (0, 1), rather than continuous, bodily signals—HAPPINESS (facial), PRIDE (posture), and GRATITUDE, LOVE, and SYMPATHY (various forms of touch)—although I don’t recommend trying to measure them this way! Famous writers have provided us with a multitude of examples of discrete TYPE 2 EMOTIONS; read Herman Hesse’s novel, *Der Steppenwolf*, for some of the finest literary emotion-state discriminations ever.

The second common mistake in research on TYPE 2 EMOTIONS is failure to measure *emotion-shift*. A *shift* in emotions is the necessary mechanism for the operation of MOTIVES, measures of which are discussed in Section 6.6 of this chapter, but I will add a quick illustration of this problem here. In a social policy-relevant study of anti-drug PSA (public service announcement) TV commercials, Fishbein, Hall-Jamieson, Zimmer, von Haeften, and Nabi (2002), in a study published in the *American Journal of Public Health*, reported that the commercial's "rated effectiveness was highly related to ... negative emotion ($r = .88$) ... and [negatively related to] positive emotion ($r = .35$)" and recommended that anti-drug PSAs "should point out the negative consequences of drug use" (p. 238). Whereas this is partly true, what anti-drug messages should *actually* do is incorporate a *shift* from a neutral to a negative emotion (e.g., from COMPLACENCY to DREAD) paired with pre- to postdrug-taking. Technically this is called *positive* punishment (see Skinner 1959) and, to be effective, it must be perceived to follow the drug-taking behavior specifically. What Fishbein and colleagues measured (see p. 241), however, were respondents' reports of experiencing various negative emotions (SADNESS, ANGER, FEAR, and DISGUST) and positive emotions (HAPPINESS, EXCITEMENT) *singly* and at *any* time during the commercial rather than before and after the drug-taking was shown (or implied). This all too typical *static* measurement of single TYPE 2 EMOTIONS regardless of the sequence and timing of their occurrence renders all "emotion testing" of all TV commercials, not only PSAs, misleading, with consequently ineffective recommendations for advertisers.

Many of the TYPE 2 EMOTIONS are abstract attributes (the SELF is the object); however, they don't have "multiple meanings" in the usual sense but are *multicomponential*. As explained by Rossiter and Bellman (2010), multicomponential attributes require a *single-item* measure, but one that must not be confused with a "double-barreled" and hence ambiguous measure. A multicomponential TYPE 2 EMOTION relevant to marketing (and advertising, e.g., for McDonald's) is LOVE, in the sense of "romantic" or "passionate" love (see Hatfield and Rapson 2000). ROMANTIC LOVE has *two* component attributes that must be present *simultaneously* and must not be measured as separate items. Here is the item wording we used in our study to measure the type 2 emotion of ROMANTIC LOVE: "I would say that I feel deep affection, like 'love,' for this brand, and would be really upset if I couldn't have it. □ Yes or □ No." Note that there are two component attributes that the rater is *simultaneously* endorsing if he or she answers "Yes," one being an intense positive feeling beyond mere liking (see Langner, Rossiter, Fischer, and Kürten 2010), and the other being a painful sense of loss when the rater (the ACTUAL SELF) is separated from the love object.

After writing a multicomponential single item, you can test and edit the item by asking the excellent methodological question devised by my Rotterdam School of Management mates van Rekom, Jacobs, and Verlegh (2006): "Would it still be [concept] if it *didn't* have this [component]?" For example, it wouldn't be an AUSSIE BURGER if it didn't have a big slice of Beetroot! Or as one of the *Desperate*

Housewives said, “She’s just one dead dog away from a country song . . .” (the other two essential components, as everybody knows, are a lost job and a departed love).

In our recent study (Rossiter and Bellman 2010) we developed measures of five TYPE 2 EMOTIONS representing the typical focii of “EMOTIONAL BRANDING,” which is very popular in advertising worldwide at present. These items were written by the researchers after intensive open-ended pretesting with consumers, in which we asked them to explain *in their own words* what the single-word emotions of TRUST, BONDING, RESONANCE, COMPANIONSHIP, and LOVE mean (these are the five most prevalent “attachment” emotions used in TV commercials by creatives in advertising today). We—the researchers—then decided on the final wording. The items are given in Table 6.4. The last item—measuring LOVE—is *multicomponential*. The other attachment-like emotions have only a one-component attribute. The “softer wording” tactic—putting sensitive terms in quotation marks—was adopted in order to get adults, and especially middle-aged and older men, to disclose some of these rather personal admissions about how they feel toward the inanimate commercial objects that are brands.

Table 6.4 Measures of some attachment-like type 2 emotions for brands

• I trust this brand	□ Yes	□ No
• This brand fits my “self-image”	□ Yes	□ No
• I regard it as “my” brand	□ Yes	□ No
• It is like a “companion” to me	□ Yes	□ No
• I would say that I feel deep affection, like ‘love,’ for this brand and would be really upset if I couldn’t have it	□ Yes	□ No

Note also our avoidance of uncommon noncolloquial words such as “empathy” or “regret,” which too often appear in erroneous *multiple-item* measures of emotions. If you are interviewing REAL CONSUMERS as opposed to COLLEGE STUDENTS (see Chapter 5), you need to use everyday language in the items—and I wouldn’t assume that the average college student has much more than average semantic knowledge these days, either! It is *always* better to err on the “dumber” side in choosing item content.

6.5 Overall Attitude (Singular)

> *Clockwork Orange*
>
> – Famous Anthony Burgess novel, and excellent Stanley Kubrick film, about *human evaluative conditioning*

OVERALL ATTITUDE (singular) is actually a *bipolar belief* expressing evaluation of an object (as conceptually defined by Fishbein 1963). However, whereas BELIEFS are stored verbally in memory (Collins and Quillian 1972) and, therefore, should be measured with a single-item *verbal* question *and* answer scale, OVERALL ATTITUDE is a *quantitative conditioned response* and, therefore, should be measured with a *numerical* answer scale (see Hull 1952; see the neo-Hullian theorist Weiss, 1968, for the OVERALL ATTITUDE construct in social psychology; and see Rossiter and Foxall, 2008, for this construct in consumer or buyer behavior). The most valid item-type for measuring OVERALL ATTITUDE is one that in the question part names or depicts the attitude object *exactly* as the rater recalls or encounters it (see Chapter 3) and then provides a fully *numerical* answer scale, end-anchored by appropriate bipolar evaluative adjectives or adjectival clauses.

For a relatively *simple* attitude object, raters are unlikely to be able to discriminate beyond five categories and so the following answer scale is appropriate:

Dislike very much	− 2	−1	0	+1	+ 2	Like very much

For a *complex* attitude object, an 11-point bipolar numerical scale is appropriate:

Bad	−5	−4	−3	−2	−1	0	+1	+2	+3	+4	+5	Good

OVERALL ATTITUDE should always and only be measured by *one good single item* (see Rossiter and Bergkvist 2009) and this item will *differ* according to the precise nature of the conditioned evaluative response (see Chapter 4). Other items necessarily will be to some degree "off attribute" and, therefore, will always distort the observed score away from the true score. This realization was discussed in Chapter 4 and illustrated in Fig. 4.1 in that chapter. Also, the OVERALL ATTITUDE answer scale should be *numbered bipolar*, with a midpoint of *zero* (as above). It should not be numbered 1–5, 1–7, 1–9, or 1–11, because unipolar numbering obscures the bipolar attribute for the rater. Fishbein himself makes this mistake—or made, since he is now, sadly, no longer with us—see Fishbein and Ajzen's (1975 and 2010) books.

6.6 Motives (Conscious and Subconscious)

Human MOTIVES are another construct that has largely dropped out of favor among psychologists and academic market researchers—but certainly not among clinical and market research *practitioners* in either field. Unlike with emotions, there has been no important academic resurgence of interest in this fundamental psychological construct (and here I discount the shallow reinterest in Kurt Lewin's "goal" notion, which is far too cognitive for what are essentially mechanistic processes).

MOTIVES, or more generally MOTIVATION, is one of the three constructs in the universal "performance equation," which postulates that PERFORMANCE = MOTIVATION × ABILITY. Two-predictor equations of this form originated from Hull's (1952) *systematic behavior theory*, the core of which is the equation BEHAVIOR = DRIVE × HABIT STRENGTH, where the HABIT, if not learned, may be an innate one, that is, a DISPOSITION. (HABIT is an example of a *concrete psychological* attribute—a CONCPSY attribute—and one that is measurable with a single item, unlike ABDISP attributes; see Chapter 4.) For my money, Hull's and particularly the Hull-Spence version of Hull's theory (see Weiss 1968, Rossiter and Foxall 2008) is the best theory ever invented for explaining human behavior because it incorporates clear constructs—concretely measured—and it specifies the exact *causal* relationships between the constructs.

The Vroom-Yetton (1973) model of PERFORMANCE in organizational behavior is another example of the MOTIVATION × ABILITY equation.

Even Fishbein's famous ATTITUDE model was later extended to include motivation—by Fishbein and Ajzen (1975)—in the MOTIVATION TO COMPLY WITH THE REFERENT construct, where the dependent variable in the equation is BEHAVIOR. Fishbein and Ajzen realized, like most of the early psychologists, that some sort of *motivation* is required to translate an attitude into action. The omission of motivation is the principal inadequacy of *cognitive* theories of performance—notably Miller, Galanter, and Pribram's (1960) widely heralded (in academia) TOTE theory.

People *cannot validly self-report* the MOTIVE that energizes a particular behavior. Freud long ago understood this as, later, did McClelland, Maslow, and other motivational psychologists. The operative motive must be inferred by the researcher based on *qualitative interviews* (see Chapter 8).

The only comprehensive yet parsimonious classification of MOTIVES is the typology devised by Rossiter and Percy (1987, 1997). Their typology specifies eight distinctly different motivational processes, as listed in Table 6.5. Five of the motives are based on negative reinforcement and three on positive reinforcement. This indicates just how much of our behavior—I would say at least two-thirds—is prompted aversively by a "stick" rather than stimulated by the hope of getting a "carrot."

Based on my recent interest in LEADERSHIP in management (research in preparation) and also having successfully "climbed" the academic and corporate-consulting "ladders" in life, I am seriously considering adding another motive class which would be POWER (positive ego-reinforcement). The POWER motive seems to involve an amalgam of all three types of positive reinforcement listed

Table 6.5 Rossiter and Percy's eight basic motives

Negative reinforcement motives

1. Problem removal
2. Problem avoidance
3. Incomplete satisfaction
4. Mixed approach-avoidance
5. Normal depletion

Positive reinforcement motives

6. Sensory gratification
7. Intellectual stimulation or mastery
8. Social approval

above. The theorizing of McClelland (1975) and Allport (1985) supports this addition (although Maslow 1943, surprisingly omits it). I also sense that POWER has become a motive for the purchase of *ego-reinforcing* consumer products in a society in which people are feeling increasingly powerless to influence the world they live in.

Correct identification of the individual's main motive for performing a particular behavior can only be validly made by an *expert qualitative researcher* (again see Chapter 8). Each of the Rossiter-Percy motives, with the sole exception of NORMAL DEPLETION, is manifest in various forms (but adheres to a single mechanistic process). For example, for a substantial segment of car buyers, there is no doubt that a largely subconscious *sexual* form of SENSORY GRATIFICATION is involved. (Ernst Dichter, his name perhaps Freudian, introduced this idea into marketing research, which was originally called *motivation research*; see Dichter 1964.) Large, long vehicles are, for many people, phallic symbolic; motorbikes are intercourse symbolic; and highly protective cars are womb symbolic.

Two further examples of the idea of a single motivational mechanism manifest in different forms would be the PROBLEM-REMOVAL class of motive (where the common energizing mechanism is the emotion-state of Annoyance) and the PROBLEM-AVOIDANCE class of motive (where the energizing emotion is not Annoyance but Fear). The now-classic "problem-solution" advertising format pioneered by the Procter & Gamble company should be deconstructed to fit one of these two classes of motives. For example, Tide laundry detergent "gets out tough stains" (PROBLEM REMOVAL) while Crest toothpaste "helps prevent cavities" (PROBLEM AVOIDANCE).

Each of the eight motives, mechanistically, involves an *emotion-shift* process (Rossiter and Percy 1987, 1997, Rossiter and Bellman 2005). The idea of a *shift* between *different* emotions was suggested to me from the comprehensive learning theory devised by Mowrer (1960)—an excellent *clinical* as well as experimental psychologist. In the case of negative reinforcement motives, the shift is from a negative emotion (such as Annoyance or Fear) to a neutral or mildly positive emotion (Annoyed → Relieved, or Fearful → Reassured). In the case of the positive reinforcement motives, the emotion-shift is from a neutral state or from a mildly negative emotion (such as Bored or Apprehensive) to an intense positive emotion (Bored → Excited, or Apprehensive → Flattered). It is the self-report of these specific emotion-states *and* the transition or shift from one to another that is picked up by the qualitative researcher, enabling the valid inference of a particular motive being operative to be made.

I'll end this section on MOTIVES with a personal example or two. If I am maddened by the fact that very few colleagues have picked up on my C-OAR-SE theory of measurement and will be placated only if the present book is widely understood and used, which motive energized me to write it? Answer: the first one. I am also annoyed, and dismayed, that no one in marketing has picked up on my emotion-shift theory, which is undoubtedly valid. Instead, my colleagues—and quantitative practitioners—piddle round studying emotion-states in *isolation* when trying to measure motives.

6.7 Personality Traits and States (and Types)

PERSONALITY TRAITS are abstract dispositional (ABDISP) attributes of an individual that predispose—and cause—the individual to think and behave in a certain way almost regardless of the external situation. PERSONALITY *STATES* are merely the individual's short-term, real-time manifestations of a *high level* of one of the traits. An example of a personality-state manifestation that most of us can relate to among our acquaintances, or even within ourselves, especially if we are academics, is the CHRONIC INTROVERT who displays HIGHLY EXTRAVERTED behavior in a circumscribed *situation*. My dear friend and colleague, Larry Percy, co-author of my advertising management textbooks (1987, 1997) is a self-confessed chronic introvert, but when he has to go "on stage" to give a lecture or make a presentation to business clients, he becomes a total extravert! In fact, when we have to present together, as we often have done at conferences and also in consulting roles, we jokingly look at each other and say "Showtime!" as we morph into extraversion. Another example is that even the CALMEST individuals will become HIGHLY ANXIOUS in at least one circumscribed situation. For instance, some people are acrophobic, some are agrophobic, and many otherwise placid women go nuts if they spot a mouse or a cockroach or see their male partner brush his teeth in the kitchen sink (at least before the advent of pausable TV, which allows the euphemistic bathroom break)!

PERSONALITY *TRAITS*, on the other hand, are apparent in the individual from about age three and are "wired in" throughout life with, for the vast majority of people, only marginal fluctuations over their lifespan (see Harrington, Block, and Block 1983). However, I predict that in studies of the increasingly older adult population we are going to see meaningful changes in individuals' EMOTIONAL STABILITY and INTROVERSION late in life as a result of major *health scares* (more unstable, more introverted) and as a result of *surviving* a major health scare (more stable, more extraverted). A recent study by Dutch researchers (Klimstra, Hale, Raaijmakers, Branje, and Meeus 2009) claimed to find "maturation" of the "Big Five" traits in late adolescence, but inspection of their age plots (on p. 904) reveals that EXTRAVERSION is the only trait among the largely taciturn Dutch—I know they are because I lived in Holland for a year—that increases from childhood to late adolescence by more than half a scale-point on the 1–7 unipolar answer scale (which the researchers wrongly referred to as a "Likert" scale). The increase is mainly evident among males, a finding easily accounted for by the tendency for most males to get more confident and outgoing by the *end* of adolescence.

An interesting illustration of a possible abstract dispositional (ABDISP) TRAIT manifestation at the time of writing is the scandal surrounding world champion golfer, Tiger Woods, who was recently "outed" for multiple instances of cheating on his wife, following the long-held public perception of him as introverted and just plain boring (though not at golf!). David Buss, a psychologist at the University of Texas at Austin, has postulated that men's "mating strategies" (which are obviously not evident at age three!) may be genetically predisposed and indeed evolutionary. Buss (2009, p. 360) points out that "Some pursue a long-term mating strategy marked by lifelong monogamy. Others pursue a short-term mating strategy marked by frequent partner switching. And still others pursue a mixed mating strategy, with one long-term mate combined with some short-term sex on the side." Tiger might justifiably blame evolution, or his designer genes, for his indiscretions! Buss also believes the mating strategies apply to women but that social norms have led to repression of these genetic predispositions more than among men. As I hinted earlier in this book, *Libido* is a mightily ignored drive, which is manifest in behavior differently according to the individual's personality type.

I have given frequent examples in this book of the important construct of PERSONALITY TRAITS and I made two observations that should be repeated here. One is that PERSONALITY-*TYPE* theory is, in my view, a much more sensible—and useful—theory than PERSONALITY-*TRAITS* theory (the popular example of the latter being "Big Five" theory). What matters is the *constellation* of *trait-levels* in the individual, not the specific traits themselves. To give an example relevant to the research world in academia, if you wanted to hire a young researcher who does not yet have an established publication record, then a well above-average bet would be to hire an *emotionally stable*, *introverted*, but *egotistical* personality type (read the riveting—for academics—review of this research by Rushton 1997). A surprising proportion of top researchers in the social sciences—like me—are, to put it politely in German, *Egozentriker Arschlochen* personalitywise. Skinner, Eysenck—both of whom I had the pleasure to meet—and even the late, great sociocognitive psychologist Bill McGuire come to mind, though McGuire was reportedly a delightful extravert at the frequent cocktail parties he and his wife hosted.

PERSONALITY-TYPE theory is starting to take hold in health psychology. The Dutch medical psychologist Johan Denollet, for instance, found that "Type D" individuals are much more at risk of heart disease than the average person (see Wark 2005, and Denollet 2005). It was previously believed—and still is by many medical researchers and practitioners—that heart trouble is most prevalent in "Type A" individuals ("driven workaholics" who report chronically feeling stressed and angry), but this turns out not to be the case. The better predictor is "Type D," which is a combination of EMOTIONAL INSTABILITY (NEUROTICISM), which Denollet labeled not too well as Negative Affectivity, and SOCIAL INHIBITION (a component of INTROVERSION). These two traits—combining to form a *type*—are measured by Denollet's 14-item questionnaire reproduced in Table 6.6 (from Wark 2005).

However, like most predictive validity relationships in medical and health research, the predictive accuracy of this TYPE D PERSONALITY measure is well above chance, but not absolutely that high, suggesting, of course, that there are other causes of heart disease, including no doubt, lack of exercise and a fatty diet. In Denollet's study with 300 Belgian cardiac patients, 27% of the Type Ds (using his questionnaire and scoring method) had died of heart disease within 10 years, versus 7% of the others—a numerically modest but pretty serious difference.

Also, the longest-used and most widely used personality measure in *industry*, for personal selection, is the Myers-Briggs Type Indicator (Myers and McCaulley 1985). This test is based on PERSONALITY-*TYPE* theory.

Table 6.6 Are you a type D?

Read each statement and circle the appropriate number. There are no right or wrong answers: your own impression is all that matters.

1. *Take the test*

		False	Less false	Neutral	Less true	True
1	I make contact easily when I meet people	4	3	2	1	0
2	I often make a fuss about unimportant things	0	1	2	3	4
3	I often talk to strangers	4	3	2	1	0
4	I often feel unhappy	0	1	2	3	4
5	I am often irritated	0	1	2	3	4
6	I often feel inhibited in social interactions	0	1	2	3	4
7	I take a gloomy view of things	0	1	2	3	4
8	I find it hard to start a conversation	0	1	2	3	4
9	I am often in a bad mood	0	1	2	3	4
10	I am a closed kind of person	0	1	2	3	4
11	I would rather keep people at a distance	0	1	2	3	4
12	I often find myself worrying about something	0	1	2	3	4
13	I am often down in the dumps	0	1	2	3	4
14	When socializing, I don’t find the right things to talk about	0	1	2	3	4

2. *Add your answers*

Negative Affectivity: add scores for questions 2, 4, 5, 7, 9, 12 and 13
Social Inhibition: add scores for questions 1, 3, 6, 8, 10, 11 and 14

3. *Score the results*

You qualify as a Type D personality if your Negative Affectivity score is 10 or higher, *and* your Social Inhibition score is 10 or higher

The second point I want to reemphasize about personality theory is that all PERSONALITY-TRAIT theories are based on aggregated (i.e., group-statistical) observations. But it has been found recently—in a journal article that deserves wide reading—that the well-known “Big Five” traits do *not* apply within individuals (see Molenaar and Campbell 2009). This is another logical reason for favoring a PERSONALITY-*TYPE* theory, because personality types are based on stereotyped (or prototyped) *individuals*.

I will, therefore, close this section with a few important (and new) comments on the measurement of personality traits. I am not going to be inconsistent with Molenaar’s finding here because I am of the view that Eysenck’s four fundamental traits of GENERAL INTELLIGENCE, EXTRAVERSION-INTROVERSION, EMOTIONAL INSTABILITY-STABILITY, and PSYCHOTICISM-IMPULSE CONTROL, though *derived* from group observations, are sufficiently well-supported by individual-level biological theory (Eysenck’s) to qualify as genuine *within-individual* personality traits.

Getting back to item types for measuring PERSONALITY *TRAITS*, the items should have three properties as spelled out below:

- The question part of the item should describe a *mental or behavioral activity* predefined as a component of the trait. For an INTROVERT, for example, his or her “mind is always racing” (a mental activity). An EXTRAVERT “loves to chat” (a behavioral activity). Do not be lured by the purely empirical (atheoretical) “narrow traits” idea surfacing in applied psychology—see for instance the article by Dudley, Orvis, Lebietke, and Cortina (2006) in the *Journal of Applied Psychology*. In a recent article in *Psychological Bulletin*, researchers Su, Rounds, and Armstrong (2009) echoed this nonsensical view, stating that “... compared with broad personality factors [read *traits*], personality facets [read trait *components*] are more useful for understanding subgroup differences and for predicting certain behaviors” (p. 877). A trait is a trait, and you can’t pick out “narrow bits” of it that happen to be good predictors of bits of behavior!
- The *adverb* in the question part of the personality trait-component item should be chosen such that approximately 50% of the *general* population—not the college-student population or the clinically treatable population—would respond one way and the other 50% would respond opposite. A Ph.D. thesis of genuine utility would be to locate the 50:50 adverbs for the many available items in the major existing personality batteries. Tentative examples might be “I *seldom* go to parties;” “I *often* do things at the last minute;” “I find it *fairly* easy to overcome setbacks in my life;” and “I *don’t always* respect others’ feelings” (I took the *stems* of these items—but not the adverbs—from a widely used “Big Five” personality inventory).
- The answer scale should be *binary* (“Yes” or “No”). Eysenck used a *trinary* “Yes” “?” “No” answer scale in his brief 12-item EXTRAVERSION and NEUROTICISM questionnaire (1958); however, the question-mark “Don’t Know” option in the middle is likely to encourage evasion of answers, and individuals cannot possibly *not* know how they stand on these self-report items, provided they are worded plainly and colloquially and with an appropriate “cut-point” adverb.
- Use at least 20 total items *per trait*, and 4–6 per *sub-trait*, if you seek to classify individuals accurately (see Emons, Sijtsma, and Meijer 2007) and note that these researchers found *binary* answer-scales to do just as well as polytomous ones (e.g., as used in the Type D questionnaire on the previous page).
- Don’t bother to include a Lie scale. Faking is *not* a problem (see Hogan, Barrett, and Hogan 2007).

6.8 Abilities (General Intelligence, Creative Ability)

ABILITIES are *abstract dispositional* attributes (ABDISPs). This type of attribute is one of the two cases where item-response *difficulty* theory makes sense. (I said IRT—item-response theory—should properly be abbreviated IR*D*T.) The other is *abstract-achieved* attributes, ABACHDs, such as types of KNOWLEDGE, discussed later.

Readers may note that I hold to the theory that ABILITIES—such as GENERAL INTELLIGENCE (which Eysenck, incorrectly I think, considered to be a personality trait), EMOTIONAL INTELLIGENCE, LEADERSHIP, and CREATIVE ABILITY—are largely innate. These abilities are only minimally increased by training and are suppressed only slightly by a *very* unfavorable childhood environment. (I realize I have just lost most of the politically correct "egalitarian" readers by making these statements—but they should seriously consult the research literature, particularly Lykken's important work at the University of Minnesota; e.g., Lykken 1982.) I will discuss measurement of the two most important abilities—GENERAL INTELLIGENCE and CREATIVE ABILITY—below.

General intelligence. GENERAL INTELLIGENCE, GENERAL MENTAL ABILITY—or I.Q. in plain language—is pretty much fixed by the time the child enters primary school. Attempts to "train" or be "coached" for an I.Q. test are pointless. For example, the Educational Testing Service has published a number of studies demonstrating that both coaching and re-taking of the Scholastic Aptitude Test used for university entrance in the U.S. makes a negligible difference to individuals' SAT scores—both MATH and VERBAL. A perhaps more objective review by Anastasi (1981) reached the conclusion that "[t]he usual short-term high school coaching programs yielded average gains of approximately 10 points in SAT-Verbal scores and approximately 15 points in SAT-Mathematical scores" (p. 1089). These gains translate to 3 and 5% of the respective maximum possible scores of 300. Not to be sneezed at—until you realize that most U.S. high schools coach for the SAT anyway!

The most important individual ability in life is undoubtedly GENERAL INTELLIGENCE and, most unfortunately, it is mostly inherited. This is the "great tragedy" of humankind. The best available evidence (studies reviewed in Lynn 1997), and swept under the mat by "environmentalist" psychologists, especially in the U.S.A.) demonstrates that Negroids—the technical biometric label—in Africa start life, *on average*, with an I.Q. of only 75; the now Caucasoid–Negroids in the U.K. and U.S.A., 84; Caucasoids in Europe, 100; Mongoloids (Asians) in East Asia, 107; and Asians reared in the U.S., 104. Quite disturbing to many "White Americans" will be the news that the *average* Mainland Japanese *Verbal* I.Q. is now 122 and Mainland Chinese, 115, versus the U.S. Whites' average of 100. In terms of overall I.Q., as reported by Richard Lynn in the prestigious journal, *Nature*, 77% of Japanese, and I estimate from the data 74% of Chinese, have a higher I.Q. than the average White American or White Western European. And barring a breakthrough in pharmacology—or an unlikely removal of the ban on human cloning—we can't do much about this. Nor does it do any good to ignore natural human differences. They can't be ignored anyway, because they will show up at school and in the workplace. The best that can be done is to tailor education and jobs to fit each person's mental ability.

GENERAL INTELLIGENCE (a.k.a. GENERAL MENTAL ABILITY) has now been shown "in literally hundreds of studies" to be the best predictor of JOB PERFORMANCE—in all occupations—"whether measured by supervisor ratings, training success, job knowledge … or ongoing job learning" (Rushton 1997, p. 416). The predictive validity coefficients (which are examples of $R_{\text{pop,}}$ see

Chapter 2) *increase* as you go up the occupational scale, being about $r = .42$ for blue-collar jobs and about $r = .65$ for managers and professionals (Hunter 1986). see Appendix C (on p. 153) for the binomial effect sizes on these correlations.

However, INCOME is *not* well-predicted by I.Q., the correlation between the head-of-household's income and I.Q. being only about $r = .10$ (Blau and Duncan 1967). As a result, in a very telling econometric analysis, Dickens, Kane, and Schultze (1995) demonstrated that even if all Americans had *equal* I.Q.s of 100, the income-distribution *disparity* in the U.S.A. would be about the same as it is now—shockingly wide. Looking at the data another way, Ceci and Williams (1997) concluded that a far more equal income distribution would be observed if the U.S. government were to distribute income solely according to each adult's I.Q. score! This policy might seem to be the exact opposite of the meritocratic "American way." However, given the high correlation between I.Q. and job performance, this method of paying people actually would reward *worthwhile* achievement and be *truly* meritocratic.

Critics of I.Q. testing have generally dismissed all I.Q. tests as being "culture biased" because they think it somehow unfair that only high-I.Q. individuals can score highly. Raven's Progressive Matrices I.Q. Test is, however, a highly valid test of GENERAL INTELLIGENCE that completely avoids this criticism by being nonverbal and nonmathematical (it measures concept learning and abstract problem-solving; see Jensen 1970). "Matrices," as it is called for short, is equally doable by individuals belonging to any cultural, subcultural, or socioeconomic group, but try not to tell the test-takers that it's an "I.Q. test," otherwise low-intelligence individuals are likely to underperform and high-intelligence individuals are likely to overperform (see Brown and Day 2006). The Matrices test measures *general* intelligence, or Spearman's *g*, a factor found in all factor analyses of mental-ability test scores, and it has always been assumed that the items' scores represent the effects of a *unidimensional* dispositional attribute (i.e., *g*) as required by the "reflective" psychometric model. You may recall that in the new version of C-OAR-SE theory—in this book—I hold that unidimensionality (and also factor analysis) is unnecessary for an abstract attribute because *all* abstract attributes are *formed* (or "formative," not "reflective," in psychometrics terminology). I recently came across a factor analysis of item-scores on the Standard Progressive Matrices Test that strongly suggests that I am correct. Educational psychologists Arce-Ferrer and Gruzmán (2009) computed coefficient alphas for high-school students' scores on the five series A through E of the Matrices test, where the A series of items are the easiest and the E series the most difficult (hence the description "Progressive" in the test's name). The researchers also readministered the test a short time afterward to the same students in a counterbalanced design (a paper-and-pencil test then a computer-delivered test or *vice versa*). I reproduce here the mean scores and coefficient alpha correlations for both versions in Table 6.7 (from their Table 2, p. 862).

The low alphas for each of the series A through E, which were 12 items each and *should*, therefore, produce a high alpha, were all below the value of .8 recommended by Nunnally (1978) for final use of a test, and, therefore, were *not* sufficiently unidimensional. The *total* scores had high alphas, but this is because they were composed of a very large number of items (60) and coefficient alpha is mainly a

Table 6.7 Means and internal consistency (coefficient α) of item series' scores on Raven's Standard Progressive Matrices test (from Arce-Ferrer and Gruzmán 2009)

	Mean		Coefficient α	
Item series	Paper	Computer	Paper	Computer
A	11.3	11.2	.36	.28
B	10.7	10.6	.58	.56
C	9.1	8.9	.49	.54
D	9.2	8.9	.65	.71
E	4.9	5.2	.68	.75
Total score	44.2	44.8	.81	.86

function of the number of items. The other result in the table, though unfortunately based only on aggregate rather than within-person means, suggests that the Matrices test of GENERAL INTELLIGENCE does have high stability-of-scores reliability, $R_{\text{stability}}$, which is the "derived essential" property of a measure in the hierarchical measure-evaluation theory in C-OAR-SE (see Chapters 2, 6, and also the summary Chapter 9).

The best *publicly available* tests of INTELLIGENCE SUBABILITIES—Verbal Ability, Numerical Ability, and Visuo-Spatial Ability—are in Eysenck's paperback book, *Check Your Own I.Q.* (Pelican Books, 1990). I recommend that these tests not be self-administered privately, but rather administered under supervised testing conditions because the two jointly essential characteristics of I.Q. are accuracy and speed (number correct within the time limit) and it is too tempting for most individuals to cheat or allow themselves extra time if self-administered.

Creative ability. Probably the second most important individual ability in the modern world—although it always *was* important if you look back in history, as far back, at least, as Leonardo da Vinci—is CREATIVE ABILITY. Identifying individuals with high creative ability is obviously vital in industry—for continued inventions. The industries widely include engineering, production, the arts, and my specialized applied field, advertising.

It is very important to make researchers aware that CREATIVE ABILITY and GENERAL INTELLIGENCE are *independent* abilities for people with an I.Q. of 120 or higher (which is the I.Q.-base score needed by great majority of graduate students). I.Q. tests, such as Matrices, measure convergent thinking, or abstract problem-solving ability where there is only one correct solution, whereas CREATIVITY requires *divergent* thinking—the ability to generate many different solutions, which can then be evaluated for their originality and usefulness. In a very revealing article in the *Psychological Bulletin*, the New Zealand psychologist James Flynn (1987) demonstrated that there were "massive" gains in I.Q. in most developed nations over the 30-year (a generational) period 1952–1982, but that these gains have not been accompanied by a higher rate of inventions. The most striking case is the Netherlands; the average Dutchman's I.Q. went up by 21 points on the Matrices test over that period (from a mean I.Q. of 100 to a mean of 121), yet the rate of patents granted for inventions in the Netherlands *fell* by at least 35%. The obvious conclusion is that CREATIVE ABILITY, on average, is not increasing and may well be declining. It is, therefore, vital that we can validly identify individuals with well-above-the-norm creative talent.

The problem with most tests of CREATIVE ABILITY is that they are based on *past creative behavior* in the form of previous "creative outputs." This is a particular problem in the advertising industry where "creatives" (art people and copywriters) have to present a portfolio of their past work. Potential creative employees who have just graduated from high school or university usually do not have such a portfolio (I encourage them to submit to potential employers advertising class projects as long as they can attest in a cover letter that the creative part was mainly their own contribution). What is needed, therefore, is a test of CREATIVE ABILITY that has high predictive validity—the "desirable" criterion for assessing predictor measures as summarized at the beginning of this chapter.

Several such tests of CREATIVE ABILITY (misnamed "creative potential") have been devised and these were evaluated in a book chapter about 15 or so years ago by Kabanoff and Rossiter (1994). Since then, my previous doctoral student at the Rotterdam School of Management, Niek Althuizen, now an assistant professor at a leading business school in France, ESSEC, found a brief version—which takes only about 15 minutes to administer—of the Torrance Tests of Creative Thinking which, in two preliminary studies, has shown very good predictive validity in identifying individuals with high creative ability (see Althuizen, Wierenga, and Rossiter 2010). Goff and Torrance's (2002) set of short tests is definitely highly content-valid, being based on J.P. Guilford's three essential creativity components of Fluency, Flexibility, and Originality (see Taft and Rossiter 1966—my first journal article publication—in which we showed that Mednick's well-known Remote Associates Test, the RAT, is *not* a test of CREATIVITY, but of VERBAL INTELLIGENCE, and in my psychology honors thesis I also showed that Eysenck's theory that CREATIVITY and I.Q. are positively correlated up to I.Q. 120 but independent thereafter is supported—unless you measure CREATIVITY with the RAT test!). Torrance also postulated a fourth factor, called Elaboration, which adds to the testing time (and scoring time) but which, we found, does not add to predictive validity (see Althuizen et al. 2010).

Yesterday I came across an obscure, even shorter, and very thoughtfully scored test, which I will check for predictive validity in identifying highly creative individuals.

I would also argue that *social scientists* be rewarded for creative contributions and not for behaving like sheep, which is how the journals reward them now. Look at the "cookie cutter" articles that predominate in any journal. Experience the "onion tears" from the rectally constricted reviews you'll receive if you try to submit anything really new. Read Paul Rozin's (2009) autobiographical article about the difficulty of getting unconventional ideas—like C-OAR-SE—published. I would exclude only the innovative journals published by the Association for Psychological Science from the "cookie cutter" criticism. Not surprisingly, Rozin's article appeared in one of them.

6.9 Knowledge Tests

KNOWLEDGE—in whatever domain—is an *abstract-achieved* (ABACHD) attribute. This means that the scores on the knowledge-testing items do *not* need to be "unidimensional." (Psychometricians, listen up.) Here it would be instructive to look back at the lengthy anecdote I recounted in Chapter 4 about my experience in developing my relatively brief (30-item) test of MARKETING KNOWLEDGE, K-Test, which, when the scores were factor-analyzed, split into numerous fragmented orthogonal "factors." The items should be in the test because of their *high content validity* alone. They should be selected according to a carefully worked out prior definition of what constitutes "knowledge" in that field, ideally agreed upon by several domain EXPERTS (who would, therefore, serve as the initial *rater entity* for the test, i.e., EXPRATs, see Chapter 5). Factor analysis should *not* be used and no item or items should be dropped.

Unlike with ABILITY tests, KNOWLEDGE tests should be scored for *accuracy only*, not speed, meaning that the test should be *untimed*. A relevant and interesting choice for the knowledge questions would be items that reflect POP PSYCHOLOGY beliefs. (Here is such an item: "The 'werewolf phenomenon' has some basis in fact in that violent mental patients are approximately twice as likely to become violent during a full moon. True or False." It's true—see Calver, Stokes, and Isbister 2009.) Several tests of pop psychology beliefs—many of which are true, by the way—have appeared in the journal *Teaching of Psychology* over the years (see the references in the article by Burkley and Burkley 2009). Indeed, in writing this book, I hope that the pop psychology belief is true that "Fortune favors the brave"!

A test of beliefs versus facts in consumer behavior was devised by Wharton professor Steve Hoch (1988) and published in the *Journal of Consumer Research*. See my article (Rossiter 2004) in the same journal for how to devise a more comprehensive test of students' CONSUMER BEHAVIOR COURSE KNOWLEDGE for consumer behavior instructors.

A test of COURSE KNOWLEDGE—in any course in any field—administered at the beginning of a course in the first lecture and then again in the last lecture or in the final exam would be the most valid test of students' *knowledge gain*. A pre–post *knowledge-gain* analysis would also be the most valid gauge of *TEACHING EFFECTIVENESS*—replacing the meaningless "personality contests" that currently characterize student evaluations of their instructors (see especially Armstrong's, 1998, letter in the *American Psychologist*).

Here are the C-OAR-SE-based recommendations for constructing KNOWLEDGE tests:

1. *Recruit a panel of experts* in the substantive field of the knowledge. Three acknowledged experts should be sufficient. (Knowledge-test construction is one rare type of measurement where substantive-field experts are needed.)
2. Ask each expert to independently submit a list of up to 20 main *facts*—items of knowledge—that each considers essential to know for mastery of the

substantive field. Ask each expert to also submit a list of up to 10 popularly believed *nonfacts*—pop falsehoods. Then you—the researcher—should compile nonoverlapping lists of facts and nonfacts. Discuss with panel members if you need to.

3. Now give the two lists to each expert panel member and ask him or her to write (a) one good question per knowledge component—factual as well as nonfactual—*plus* (b) five alternative (multiple-choice) answers that include the correct answer, two close distractor answers, and two unarguably wrong answers, with "don't know" not among them.
4. From these questions and answers, you—the researcher—select and write the final set of items and answers. There must be no words in the *questions* or the *answers* that the least competent test-takers do not understand. Also, there must be no items that merely ask the testee to select the correct *definition* of a technical term in the substantive field—these are trivial displays of knowledge. Aim for at least 20 facts and 10 nonfacts, that is, at least 30 items.
5. You—the researcher—must then conduct an open-ended pretest with at least ten of the *least* competent individuals for whom the knowledge test is intended. There's no special interviewing skill required for this—just ask the person to say out loud what each question *and* answer alternative means and to point out anything that's ambiguous (use the word "unclear"). Tape-record or digitally record the interviews—always a sound practice when doing "qualitative" research.
6. Finalize the wording of the questions and answers. Do not bother the expert panelists again. This is a measure construction task, not a task requiring substantive expertise. It's *your* test and you, as the researcher, are in the "hot seat" to justify its content validity.
7. Make up clear *instructions* for the test. Next, make up a *scoring key*. Weight (enumerate) the scoring key as follows: correct answer (the true fact or the true alternative for a nonfact) = 2, distractor = 1, wrong answer = 0. Although this is most unlikely, this weighting means that an individual could just pass the test (50%) by always choosing one of the two distractor answers for every question. Very high scorers must have a high proportion of correct answers, but a "clearly knowledgeable" score—say 75% – could be composed of half 2's and half 1's.
8. Eschew item-response analysis, factor analysis, and coefficient alpha. (As emphasized in C-OAR-SE, statistical finagling after the fact cannot fix problems of low item-content validity or low answer-content validity.) The items in the test are all *defining* items (representing the experts' definitions of KNOWLEDGE in the field). Items cannot be dropped.

A final note on KNOWLEDGE tests is that at least 30 items are needed to precisely classify individuals ($R_{\text{precision}}$—with 30 the number where the distribution of test scores will approximate the Normal distribution). It doesn't matter if critics later claim that there are essential facts missing, because high scorers on the test would very likely know these also.

6.10 End-of-Chapter Questions

(6.1) The following items are taken from measures of attribute beliefs published in the social-sciences literature by, respectively, Sen and Block (2009), Melnyk et al. (2009), Qin et al. (2008), and Ajzen (1988) (you won't find these articles or books in the References and there's no need to look them up). For each, identify the *major* thing wrong (1 point each) and then correct the item, explaining your changes. (up to an extra 2 points each, so possible 12 points total)

(a) Yoplait yoghurt:
Never buy 1 2 3 4 5 6 7 Always buy

(b) I would definitely 1 2 3 4 5 6 7 I would definitely
buy the cake at the buy the cake at
closest bakery my classmate's bakery

(c) Memory records and stores all of our experiences since birth:
□ Strongly Disagree □ Disagree □ Neither □ Agree □ Strongly Agree

(d) Breast-feeding protects a baby against infection:
Likely : _____ : _____ : _____ : _____ : _____ : _____ : _____ : Unlikely

(6.2) Read the article by Rossiter and Bergkvist (2009) and then criticize the following measure (from Chaiken and Yates 1985, no need to look up). First, define the apparent construct and then write and justify a more content-valid measure. (7 points)

Capital punishment is
good _____ _____ _____ _____ _____ _____ _____ bad
foolish _____ _____ _____ _____ _____ _____ _____ wise
sick _____ _____ _____ _____ _____ _____ _____ healthy
harmful _____ _____ _____ _____ _____ _____ _____ beneficial

(6.3) In a study by Brakus, Schmitt, and Zarantonello (2009) published in the premier journal in my field, the *Journal of Marketing*, the researchers first identified, through qualitative research, many literal and relevant Type 2 emotion descriptions given by consumers for major brands (examples: "Crest toothpaste—I feel refreshed;" Nike sports gear—"Makes me feel powerful;" BMW—"I feel young"). The researchers then backed off these specific emotions in their quantitative study and instead measured reports of consumers' *general* "brand experience" (sample items: "This brand makes a strong impression on my visual sense or other senses;" "This brand is an emotional brand;" "This brand results in bodily experiences;" "I engage in physical actions and behaviors when I use this brand"). Why does their measure have *zero* content validity? What type of measure would you recommend in its place and why? (7 points maximum for the first answer, and 7 points maximum for the second, equally difficult answer)

(6.4) In the chapter, I said that motives can be validly measured only by an expert qualitative researcher's inference. Based on your own experience and introspection, which *one* of the eight Rossiter-Percy motives, listed in this chapter, do you estimate is the major buying motive for these brands in the following product categories—and why?

(a) Donations to the animal-protection charity, PETA
(b) Sara Lee Lite frozen-baked dessert pies
(c) Skype Internet phone service
(d) The university you chose or would choose for your graduate studies (if not an academic) or for your first academic job
(3 points each for quality of arguments, not necessarily for what I think is the correctly identified motive, so possible 12 points total)

(6.5) Explain how you would construct a test of how much students learned in their introductory social-science course—"Psych 101," "Sociology 101," "Management 101," and so forth. Then explain how you would validate the test. (10 points maximum for your description of the test construction and 5 points maximum for the validation procedure, thus 15 points total.)

Chapter 7
Enumeration and Scoring Rule

The King is in his counting house, counting…
—"Sing a Song of Sixpence," *Tom Thumb's Pretty Song Book,* ca. 1744.

The E in C-OAR-SE stands for the final aspect of the theory, which is a double selection procedure called *enumeration and scoring rule*. "Enumeration" means how the answer scale is scored quantitatively. "Scoring rule" had two applications: it is the rule adopted by the researcher for deriving a total score for an individual *within one item* if the construct is doubly concrete, or *across multiple items*, if the construct measured is abstract in either the object or the attribute; and it is also the rule that the researcher adopts when combining scores from individuals to compute a *group* statistic such as a mean or median.

After reading this chapter you should:

- Decide whether to use objects or raters as the units of analysis
- Know how to enumerate scores for a unipolar versus a bipolar attribute
- Distinguish the four main scoring rules for within-persons analysis
- Understand and know when to use alternative group-scoring rules

7.1 Objects or Raters as the Units of Analysis

The first big decision is whether to make the *objects* or the *raters* the unit of analysis.

In applied research it is often the case that the alternative *objects* being rated are the relevant units of analysis. Examples are COMPANIES as objects rated on the attribute of CORPORATE REPUTATION; POLITICAL CANDIDATES as objects rated on PERSONAL CHARACTERISTICS or on the attribute of INTENTION TO VOTE for them; BRANDS as objects rated on ATTRIBUTES; and TV COMMERCIALS rated in terms of degrees of LIKING.

In fundamental contrast, *raters* are the appropriate units of analysis for testing theories or hypotheses about individual-level constructs and individual-level causal relationships between constructs. It is all too easy to slip into the "aggregation error"

J.R. Rossiter, *Measurement for the Social Sciences*,
DOI 10.1007/978-1-4419-7158-6_7, © Springer Science+Business Media, LLC 2011

here. For example, many people believe that so-called LEARNING CURVES are "smooth" and either concave (psychologists believe this) or S-shaped (marketers believe this), or J-shaped (economists believe this). As Skinner demonstrated long ago (1938) these curves are smooth only because they are *averages* of individuals' learning trajectories, most of which go in the same "direction." Individually, however, they are very bumpy, consisting of small jumps, large jumps, and plateaus of varying durations.

Most interesting for psychologists (and also for parents and teachers) is that Piaget's MATURATIONAL STAGES theory can only be tested at the disaggregated individual level. This is because individual children transition into a particular stage at quite a wide age interval. Any parent who has not delegated their child's upbringing to almost total daycare would know that it is quite sudden that a child grasps the idea, for example, that objects don't go out of existence when they become obscured by another object, such as when the cat walks behind the sofa. But one's own children, if you have more than one child, may differ by as much as a year to 18 months in grasping this Piagetian concept. The ability to engage in abstract reasoning or "formal operations," Piaget's final stage of cognitive development, shows an even wider variation across individuals. The tragic fact is that about 22% of US adults *never* reach this stage, testing "below basic" in quantitative skills according to the U.S. National Center for Education Statistics. This poses a very serious and potentially fatal problem for HEALTH NUMERACY (see Reyna, Nelson, Han, and Dieckman 2009). When individual children's or teenagers' progression paths in cognitive achievement are averaged, the average path shows *no* evidence of Piaget's stagewise jumps. This is an example of an innovative and sound theory (Piaget's—see, of all people, Eysenck's, 1979, favorable review, and who said he wasn't open-minded) being overlooked or rejected because the wrong analysis was applied in the test of it. Any parent or teacher could verify Piaget's theory of "sudden grasping" but most psychologists would tell you that it has been disproved. In truth, Piaget's thoughtful stage tests correlate about $r = .6$—which is a "large" effect size—with I.Q. scores as measured by the culture-fair Matrices test (see Eysenck 1979).

Another (completely unrelated) reason why psychologists reject Piaget's theory is that it is a *genetic theory* (see Rossiter and Robertson 1976). Genetic theories are pooh-poohed by most U.S. psychologists, whom I've said before have an overwhelming egalitarian bias. Parents of multi-child families, and teachers who teach siblings, give much more credence to the influence of genetics on cognitive skills, on personality traits including sensation-seeking, and even on job and lifestyle choices (see especially Lykken's analyses of the Minnesota Twins ongoing database—e.g., Lykken 1982—and don't get confused with the baseball team of the same name). Scientifically speaking, the only good thing about the recent spate of DNA research is that genetic theories will make a comeback. It is fascinating, for example, that even a construct as evidently learned as TRUST IN PEOPLE IN GENERAL apparently has a significant genetic basis (see Uslaner 2008). As a neo-behaviorist, I can't resist pointing out that Hull's theory includes an ${}_{S}U_{R}$ construct ("unlearned," i.e., innate, stimulus-response connections) that accommodates this.

7.2 Enumeration of Ratings on Unipolar and Bipolar Attributes

Failure to distinguish a unipolar from a bipolar attribute has resulted in many stupid misinterpretations of empirical findings. One common error is to measure unipolar attributes on a bipolar answer scale and then score the answers as if they were unipolar. As noted in Chapter 6, BELIEFS, the most widely measured attribute in all social sciences, are very often measured on "Likert" answer scales, which range from "Strongly Disagree" to "Strongly Agree" and are thus bipolar (negative to positive). Researchers very often score these answers as unipolar (i.e., 1–5 if a 5-point Likert scale, or 1–7 if a 7-point Likert scale, or 1–9 if a 9-point Likert scale). The consequence is that scores of 3, 4, and 5 (respectively) make it look like the rater's attribute belief was *moderately positive* whereas on the rating scale the rater was trying to indicate a *neutral or zero* degree of belief. Unipolar attributes should be measured on a unipolar-worded answer scale (e.g., "Not at all" to "Extremely") and then scored with a series of positive numbers *beginning with 0*.

What is roughly the converse mistake is to use a unipolar answer scale when measuring a bipolar attribute. The most pertinent example in the social sciences is the attribute called ATTITUDE in the traditional sense of an EVALUATIVE BELIEF (see previous chapter). Evaluations *by definition* range from negative to positive. A notable example in the organizational behavior, management, and marketing literatures, is SATISFACTION. However, in the very widely used AMERICAN CUSTOMER SATISFACTION INDEX, the ACSI (see Fornell et al. 1996), SATISFACTION is measured with a unipolar numerical answer scale that ranges from 1 to 10, enumeration that is patently ambiguous at the lower end.

It may be noticed in many journal articles that a bipolar answer scale is often used to measure the *unipolar* attribute inherent in the construct of an INDIVIDUAL'S BEHAVIORAL INTENTION. Even the great attitude theorists Martin Fishbein and Icek Ajzen always make this mistake, and they make it not only for the unipolar attribute of BEHAVIORAL INTENTION, but also for the unipolar attribute, in their "extended model" which has had various reincarnations as "the theory of reasoned action" (TRA) and "the theory of planned behavior" (TPB), in their construct labeled MOTIVATION TO COMPLY (WITH THE REFERENT). Check the rating scales in the classic books by Fishbein and Ajzen (1975) or Ajzen (1988) or most recently Fishbein and Ajzen (2010, pp. 449–463). Logically, the least INTENTION is zero and the least MOTIVATION TO COMPLY is zero; neither can be negative as implied by the *bipolar* answer scales that these two much-imitated psychologists use.

This wrong enumeration produces *measure-distorted ratings* (D_m in the new true-score model, $T = O + D_\mathrm{m} + E_\mathrm{r}$, of Chapter 2). Many studies in social psychology, health promotion, and marketing rely on the TRA model or the TPB model and, with their low content-valid enumeration, their findings must now be questioned. The studies should be redone with properly enumerated rating scales.

7.3 Sum Scoring Rule

The final enumeration consideration is the *scoring rule*. The first three scoring rules, discussed in this and the next two sections, apply to the scoring of *multiple-item* measures. These are the Sum Scoring Rule, the Average Scoring Rule, and the Cutoffs or Profile Scoring Rule.

The Sum Scoring Rule is used when the meaningful result sought by the researcher is the individual's cumulative score over the multiple items. Obviously this applies when the measure is a *test* of some sort, such as for measuring KNOWLEDGE attributes or ABILITY attributes.

Sum-scoring is appropriate for tests and also for most other *abstract achieved* (ABACHD) attributes. One notable example is SOCIAL CLASS (see Coleman 1983), which is wrongly measured by social researchers and market researchers who use Occupation Prestige only (see Chapter 4 for a discussion of SOCIAL CLASS *versus* SES). Another example is the well-known Holmes-Rahe LIFE-STRESSORS index (Holmes and Rahe 1967) in which Death of one's spouse scores the maximum of 100, and second is Divorce with a stress-score of 73, with a Vacation scoring only 13 (despite the Chevy Chase movie, although that was a vacation *avec enfants*!).

Total scores on these ABACHD attributes are always *formed* (i.e., summed) from the scores on the component attributes. The component attributes are always *unipolar* attributes. The researcher should always specify the *theoretical minimum* sum-score and the *theoretical maximum* sum-score.

The researcher should also specify the *actual midpoint* score as well, instead of dumbly splitting scores at the empirical sample median as is common practice—the so-called "median split." As a real-world example of the importance of correct midpoint-score interpretation, a parent in the United States or Australia told that his or her child has an I.Q. of 100 can rest assured that the child is "right on average" whereas a Japanese parent told the same news would have reason for concern because the average among 5-year-olds in Japan is well above 100 (more like 120; it was 117 in the early 1980s, see Reid 1983). Actually, both of these hypothetical parents would be even more satisfied if they knew the fact that average scores on I.Q. tests have been moving steadily upwards over the past few decades in major Western and Eastern countries—the "Flynn effect" (see Flynn 1987). Thus, their child is in reality more intelligent than those of several decades ago. Each successive generation of kids and teenagers may have worse manners—this is an "oldie" talking—but they are definitely getting smarter! Apparently, however, as mentioned in the previous chapter, they are *not* becoming more creative.

I wish to warn again here about a type of measure that is *not* additive—*not* sum-scored—and this is the *multicomponential single-item* measure illustrated in the previous chapter for (some) TYPE 2 EMOTIONS. I recently saw the movie *Old Dogs* in which genius comedian Robin Williams's character opined, "STRUGGLING is TRYING WITHOUT SUCCESS." This is a two-component single-item definition of STRUGGLE. You cannot measure STRUGGLE, as multiple-item theorists would, by subtracting the individual's score on SUCCESS from his or her score on TRYING! The correct measure is: "I am trying very hard but have been unsuccessful. □ Yes □ No." This measure concretely represents the construct of STRUGGLING in a single item, which is also correctly answered *binary* (see Chapter 6) and correctly enumerated (scored) 1, 0.

7.4 Average Scoring Rule

The Average Scoring Rule at first seems to be a harmless variation of the Sum Scoring Rule because it is the *sum* of the items' scores (on a multiple-item measure) divided by the number of items. An individual's scores on the multiple items, each measured on an answer scale of (e.g.) 1–7, are sometimes "averaged back" to the numbers on the answer scale. For example, on the 1–7 scale, one person might achieve an average score of, say, 6.1 and another person 6.3. The Average Scoring (or Average Score) Rule is regarded as appropriate by psychometricians because of their belief in "domain sampling" theory, whereby the multiple items are a random sample of possible items, and in any random sample it makes sense to take an average score.

However, as pointed out in Chapter 2 on validity and again in Chapter 4 on attribute classification, domain sampling has no place in C-OAR-SE theory. This is because all abstract attributes are measured by items that are not an arbitrary-sized random sample, but rather by items that are there *by definition* (of components) and are *fixed* in number. This means, if you think about it, that the *Sum Scoring Rule* is the right one.

In other words, the Average Scoring Rule should not in any circumstances be used for reporting INDIVIDUAL rater entity (INDRAT) scores. Provided the answer scale is properly enumerated (meaningfully, content-wise), *sum-scores* are correct, with the theoretical minimum score, maximum score, and middle score explicitly indicated in the report.

An *average*—the Average Scoring Rule—is of course appropriate in *group* scoring, though we will see at the end of this chapter that it is the *median*, not the mean, that is sought.

7.5 Cutoffs or Profile Scoring Rule

What I am calling the Cutoffs or Profile Scoring Rule is a very widely applicable—and almost always overlooked—scoring rule. Measurement buffs might recognize it as the *conjunctive* scoring rule (applied to multiple items' scores). However, it can be applied to scores in two situations: group scores on a single item, and individual scores on component items of abstract constructs—constructs with either an abstract object, an abstract attribute, or both—which of course require multiple items (at least one item per *component*).

Single-Item Cutoff(s). Cutoffs or Profile-Scoring of group scores on a *single item* is essentially what happens when researchers dichotomize the total sample of *respondents*—that is, *raters*—when making a *median split* of individuals' scores. On bipolar rating scales the "split" should be made at the *theoretical numerical midpoint* of the scale if there is an even number of points in the answer scale, or at the *actual numerical midpoint* if it is an odd-numbered answer scale. Otherwise, the split doesn't make psychological sense and the dichotomy is neither rational nor real. Researchers should cease using *empirical* median splits!

Another example of Cutoffs or Profile-Scoring application is a bit more sophisticated and this is a *threshold split* of scores on a predictor variable. An example of a single-threshold split is for a bipolar SATISFACTION answer scale where the verbal answer category "Delighted" is added to the positive end of the scale (as recommended by expert service researcher Roland Rust; see Rust, Zahorik, and Keiningham (1995)). Answers (scores) of "Delighted" apparently predict REPEAT PATRONAGE differentially well and so the *enumeration* of this scale point should be given a nonlinear and larger weight. The size of this weight depends on the degree of upturn in REPEAT PATRONAGE empirically observed. *Hypothetical* weights might be: "Very Dissatisfied" (−3), "Somewhat Dissatisfied" (−2), "Slightly Dissatisfied" (−1), "So-So" (0), "Slightly Satisfied" (+1), "Satisfied" (+2), "Very Satisfied" (+3), and "Delighted" (+6). Don't argue, you "quant jocks"—this is only a hypothetical example! And it's "interval," not "ordinal."

Sometimes, there are *two* thresholds. A fiducially and fiscally important example is Reichheld's (2003) Net Promoter Score, widely used by marketers to measure the percentages of word-of-mouth "Promoters" and word-of-mouth "Detractors" of the brand (the difference is the "net" of Promoters). Net Promoter is now being applied to evaluate word-of-mouth (WOM) comments by the twits who use online social networks such as Twitter (Klaassen 2009). Net Promoter relies on a 0–10 numerical answer scale of RECOMMENDATION INTENTION that has *two* thresholds, a lower one at 6 for Detractors (scores of 0–6) and an upper one at 9 for Promoters (scores of 9 and 10). Presumably (but probably *just* a presumption, knowing practitioners) the two thresholds have been predictively validated against the twits' actual WOM behavior.

Australian wines—which many (with the obvious exceptions of Francophiles, Germanophiles, and Kiwis) regard as the best in the world—provide an influential practical example of *multiple* Cutoffs scoring on a single-item verbal rating scale. The authoritative *Langton's Classification of Australian Wine* is "Distinguished" (the lowest category!), "Excellent," "Outstanding," or "Exceptional" (e.g., Penfolds Grange Shiraz, Cullen Diana Madeline Cabernet-Merlot, and Giaconda Chardonnay from rainy Victoria!). All these adjectives translate to scores of 90+ on a 0–100 numerical scale. The recently released 2004 Penfold's Grange Shiraz is rated 99 and the 1976 vintage of Grange scored 100, according to U.S. wine critic Robert Parker (Speedy 2009).

Multiple-Item Profile. The second application of the Cutoffs or Profile Scoring Rule is when scoring and *adding* component scores for abstract and thus *multiple-item* constructs. The Profile Scoring Rule (I'll just call it this for convenience), therefore, must be sharply distinguished from the Sum Scoring Rule. The Profile Scoring Rule has most often been exemplified historically in clinical psychology, psychiatry, and health and medical research as "patient profiles," of which the Minnesota Multiphasic Personality Index is the best-known exemplar (there's another plug for my lovely friend Vicki's beautiful U.S. "twin cities" Minneapolis-St. Paul, of which St. Paul is the usually neglected twin).

Profile scoring can also be applied to *verbal* answer scales. Fascinating clinical profiles of the Nazi war criminals Joachim von Ribbentrop and Hjalmar Schect, as rated by one psychiatrist, reported in an article by Harrower (1976, p. 343), are reproduced in Fig. 7.1. You may notice, as the psychiatrist did, that Schect appears to be disturbingly normal! In a story reminiscent of *Shindler's Ark* (the book) or *Shindler's List* (the movie) it has recently been discovered that it was an intellectually adoring Nazi, Anton Sauerwald, who arranged the exit visas for Sigmund Freud family's flight to Britain in 1938 when Hitler invaded Austria (Cohen 2010).

I wish to point out that the clinical type of profile scoring is based on personality *type* theory (a so-called "idiographic" theory) and not on the far more widely

	(1)	(2)	(3)	(4)	(5)
PRODUCTIVITY	**Impoverished**	Reduced output	Adequate	Better than average	Rich and well-ordered
RELATION TO REALITY	Loose	Lapses – together with good focus	**Not noticeably disturbed**	Essentially firm	Firm and good
THOUGHT CONTENT	**Stereotyped**	Tendency toward stereotypy	Adequate	Original trends	Original
CONSTRUCTIVE FANTASY	Absent	**Barely accessible**	Accessible	Readily accessible	Active but not hampering
DRIVE	**Hampering passivity**	Insufficient drive	Adequate	Clearly sufficient	Exceptionally well directed
EMOTIONAL TONE	**Lacking**	Indicated but repressed emotions	Trend toward emotional expression	Warmth available	Warm, readily available
ANXIETY	**Disintegrating**	Marked	Moderate	Not marked	Lack of evidence of anxiety
OVERALL EVALUATION	**Markedly disturbed personality**	Less than adequate personality with some psychological problems	Adequate personality	Better than verge functioning personality	Exceptionally well-integrated personality with excellent potential

J. VON RIBBENTROP H. SCHACT

Fig. 7.1 One psychiatrist's clinical profile scoring of the personality characteristics of the Nazi WWII criminals Joachim von Ribbentrop (*bolded*) and Hjalmar Schact (*underlined*). Source: Adapted from Harrower (1976)

assumed personality *trait* theory (a so-called “nomothetic” theory—see Allport 1935). You will remember this distinction if I ask you to visualize, for example, a “George W. Bush type” of politician versus a “Barack Obama type” of politician—as opposed to my listing each politician’s scores on the “Big Five” personality traits!

Profile Scoring may be applied to the component objects of a construct that have an *abstract* (multiple-meanings) *object*, such as PERSONAL VALUES. In the usual procedure adopted by VALUES-researchers, continuous scores on measures of each of the values are simply added together (Rokeach, for instance, does this). But *cutoffs* should first be applied. For example, I’m working on a new measure of AUSTRALIAN VALUES, which I am dubbing “AUSSIE VALUES” because many are unique to Australian or Australian-acculturated individuals (see especially the Australian clinical psychologist Ronald Conway’s book, *Land of the Long Weekend*, 1978). More precisely, Australians’ *cutoffs*—their *profile*—is what is unique, the uniqueness being that the combination of *above-threshold scores* on the values—forming a TYPE—is what really defines an “Aussie.” This new measure will hopefully sink the “culturally universal values” idea, which anyone who has traveled outside their own country (or even their own subculture) can see to be so obviously untrue. For instance, how many CROCODILE DUNDEE types have *you* met if you don’t know any Australian blokes or haven’t been Down Under? Regrettably, there aren’t too many left here, either, but nearly all Australian males aspire to this personality type, consciously or not.

The other case is where the *attribute* of the construct is abstract and, therefore, has *multiple components* (an ABACHD or an ABDISP attribute, see Chapter 4). In my new version of C-OAR-SE theory (in this book), *all* abstract attributes are *formed* attributes and, therefore, require an *adding* of the component scores (which themselves may be single-item scores or summed or averaged multiple-item scores). It is the *above-cutoff* scores that should be added.

An example of Profile Scoring by adding above-cutoff ratings is when judging an ADVERTISING IDEA, or a fully produced NEW AD, as CREATIVELY EFFECTIVE. The attribute of CREATIVE EFFECTIVENESS when applied to an ADVERTISEMENT as the object of the construct is not a CONCPERC attribute, but rather is an abstract achieved (ABACHD) attribute, consisting of two components—actually two concrete perceptual components (CONCPERCs)—namely, Creative and Effective (see Rossiter and Percy 1997, or Rossiter and Bellman 2005, or Rossiter 2009b). These two-component attributes are usually judged—by EXPERTS as the rater entity (EXPRAT)—on single-item unipolar rating scales (e.g., 0–10) for each component. Academic advertising researchers sometimes use these two correct measures but, incorrectly, they *average* the scores. What should be done is to apply a theoretically and practically safe *cutoff* (say 7+) to each and then *add* the expert judges’ scores above the cutoff (i.e., 7, 8, 9, or 10) and at the same time recode all scores below the cutoff as *zero*. This means that an ADVERTISING IDEA or a NEW AD can only get a positive score for CREATIVE

EFFECTIVENESS if it exceeds the cutoff on *both* components. Moreover, since the more creative the better and the more effective the better, the IDEA or AD as the *object* (a CONCOB) is additionally rewarded if it achieves more of either.

The second example of single-item component scores comes from Amazon.com, the world's most successful e-retailer (as at the time of writing). Amazon, although its managers may not realize it, uses Profile Scoring in its so-called "mission statement" or "value proposition." Amazon.com's stated value proposition (reported in a newspaper article by Stahl and Fowler 2009) is: "Lowest Prices + Best Selection + Great Customer Service." I hope you can see that this is an application, inadvertent or not, of the Cutoffs or Profile Scoring Rule.

An application of this rule to components that are *themselves* abstract, and, therefore, require multiple items (as *second-order* components), which may vitally affect POPULATION MIGRATION is the United Nations Human Development Index (Callick 2009). In this index, countries are ranked on the basis of an overall HDI score made up of four-component scores on the attributes of Gross Domestic Product, Education, Literacy, and Life Expectancy. These components are themselves abstract because they must have more than one concrete first-order component (e.g., Life Expectancy would be computed for men and women separately since in every known country, as I recall, women's life expectancy is longer than men's and what is more, these forecasts are based on historical life durations and projected for children born in the current year and medical conditions in the country may have improved greatly in that country in recent times that would throw these forecasts off). I'm not sure how the total HDI score (the index) is computed but I'm willing to bet it is done by simple addition of the four-components' scores. If so, this is silly. It means, for instance, that many middle-ranked countries may be outstanding on two of the components but poor on the other two. Actually, this could even be the case for top-ranked countries—which in 2007, the latest year for which the HDI is available, are Norway, Australia, and Iceland—because the index provides *rank* scores, not *absolute* scores. I know personally that this is *not* the case for Australia because my country rates absolutely highly on all four components, so I suspect it is not the case either for the country that "beats" us, Norway, or for the first 10 or so countries below us. (In case you're wondering, the U.S.A. didn't make the top 10 in 2007, ranking 13th.) The point is that this index should be scored using the Cutoffs or Profile Scoring Rule. In other words, assuming that the four components, GDP, Education, Literacy, and Life Expectancy really do signify "human development," then one would want to live in, and raise children in, a country that has *high* GDP, and a *very high minimum* level of Education and *universally functionally adequate* Literacy. So that you fully understand the importance of using a valid scoring rule, in this case the Cutoffs or Profile Rule, imagine the hypothetical case of a middle-ranked country—realizing that most people *live* in the middle-ranked countries—that achieved its middle ranking by scoring just *average* and thus *below the cutoffs* on all four factors.

7.6 Multiplicative Scoring Rule

Perhaps the best-known application of the Multiplicative Scoring Rule in the social sciences is Fishbein's (1963) COMPONENTIAL WEIGHTED BELIEFS ATTITUDE model (and formula) which is $A_{\mathrm{o}} = \Sigma i(B_{\mathrm{oi}} \times E_{\mathrm{i}})$, where A_{o} = overall attitude toward object o; B_{oi} (wrongly singly subscripted by Fishbein and followers as B_{i}) = belief that object o has attribute I; and E_{i} = evaluation of attribute i. The *B* and *E* terms are multiplied for each belief before adding, which is denoted by the Σi symbol.

Another well-known example, especially in health psychology, is Rogers' (1983) model of FEAR-APPEAL EFFECTIVENESS, where ACCEPTANCE of the fear appeal is estimated by multiplying, within person, the PERCEIVED THREAT score by the PERCEIVED SELF-EFFICACY score for achieving the suggested remedial action. Most smokers, for instance, do not accept "quit" messages (cancer researcher Ron Borland even allowed that anti-smoking ads probably encourage relapse by increasing cravings among attempting quitters when they see, hear, or watch these ads—see Dart 2009). Although smokers perceive the THREAT as very high (because of its severity, not its probability; the probability that a U.K. male lifetime smoker will contract lung cancer is about 16%, not "nearly 100%" as anti-smoking ads imply, and it only reduces to 9% if a White male smoker quits late in life; see Peto, Darby, Deo, et al. 2000), their felt self-efficacy to quit smoking (and the actual success rate, which is never publicized but I have heard it admitted at medical health conferences) is quite low when relapse is taken into account and when nicotine ingestion is objectively measured. None of this would be apparent from regression models, which mistakenly *add* THREAT and EFFICACY (e.g. Witte and Allen 2000).

A recent study by Kusev, van Schaik, et al. (2009) showed that for *real-life decisions*, namely what things to take out personal insurance against, as opposed to *monetary gambles*, objects on which much of the PERCEIVED RISK literature is based, people greatly exaggerate low-to-moderate probabilities. This finding says that Kahneman and Tversky's Nobel Prize-winning PROSPECT THEORY doesn't hold up for major life-threatening events! This is an example of the importance of defining the *object* in constructs—as emphasized in my C-OAR-SE theory.

Another example of the Multiplicative Scoring Rule occurs in all PERFORMANCE EQUATIONS (of the form PERFORMANCE = MOTIVATION × ABILITY, see Chapter 4) such as in Hull's (1952) or Vroom and Yetton's (1973) theories.

At the more micro-level in *marketing*, Urban and Hauser (1993) describe the ASSESSOR model in which BRAND PURCHASE INTENTION = BRAND AWARENESS × BRAND PREFERENCE (see also Rossiter and Percy 1987, 1997, or Rossiter and Bellman 2005, for more detail). This *multiplicative* model of these two communication effects is routinely ignored by advertising and consumer behavior academics who (a) wrongly treat BA and BPREF as *additive* independent variables and (b) wrongly analyze the scores at the *group* level rather than the individual-consumer level. These mistakes have produced many, many (that's many × many) nonsensical empirical findings in both academic and commercial advertising research.

A "pop psychology" but I think real-world plausible application of the Multiplicative Scoring Rule is Hollywood talent-agent Larry Thompson's RAGE TO SUCCEED model (reported in Lawson 2009). Thompson says of actors and actresses, "The more rage you have, the less talent you need." (He actually said "may need," suggesting some lack of confidence in his own model.) This may be recognized as a variation of the PERFORMANCE = MOTIVATION × ABILITY model where, here, SUCCESS = RAGE × TALENT.

7.7 Alternative Group-Scoring Rules

A *group rater-entity* is many times involved in constructs in the fields of management (e.g., TOP MANAGEMENT TEAM), organizational behavior (e.g., STRATEGIC BUSINESS UNIT, or SBU), and social psychology (e.g., THE FAMILY or, these days, CO-DWELLERS) and it matters crucially how group-level measures are scored.

Steiner long ago (1972) rationally worked out the alternatives for group scoring, as summarized in Table 7.1.

Table 7.1 Group scoring rules (Steiner 1972)

Rule	Situation
Sum	Performance equal to "sum of the parts"
Average (median)	Individuals' contributions likely to be compensatory
Maximum	Problem solvable by one individual
Minimum	"Weakest link" group tasks

I'm now going to return (see Chapter 5) to Susan Bell's "totally excellent, dude" (2007) study, with acknowledgments to Keanu Reeves' dialogue writer in the movie *Bill and Ted's Excellent Adventure*. She claimed in general not to find much support for Steiner's typology—that is, for a match of the group scoring rule to the type of team task. However, closer inspection of her findings shows considerable support for matching, for most of the predictor variables: Conscientiousness ($r = .13$ for matched versus $r = .02$ not-matched), Agreeableness (.18 versus .11), Openness (.11 versus .03), Collectivism (.23 versus –.05), Emotional Intelligence (.26 though too few studies for not-matched), but no difference for Extraversion (.05, .06) or General Intelligence (.28, .29). In any event, I believe the correct group scoring rule should be used for the reason of rationality—regardless of any empirical results.

Also, an *objective* measure of TEAM PERFORMANCE should be used. I would, for instance, discount the findings reported in the *Journal of Personality and Social Psychology* recently by De Rue and Morgeson (2009), in which the researchers used a subjective measure (supervisors' ratings). The researchers also measured (and aggregated) individuals' performances, instead of measuring *team* performance and using a task-appropriate *group* scoring rule.

7.8 End-of-Chapter Questions

(7.1) Find a study in your field of social science where raters were wrongly made the units of analysis instead of objects as the units. Explain why raters are the wrong units in that study. (7 points)

(7.2) Are "lucky charms" effective? As reported in Bialik (2010), researcher Lysann Damisch and two unnamed colleagues at the University of Cologne in Germany recently told half the golfers on a putting green that they were playing with a lucky ball and told the other half nothing. Those with the "lucky" ball sank 6.4 putts out of 10, an average of almost two more sunk putts than those using the "normal" ball—a 35% relative improvement. Participants were 28 college students (Bialik 2010). How would you analyze and interpret these data? (5 points)

(7.3) From what you have learned in the last chapter and this chapter, write a thorough (at least 2,000 words) evaluation of Young & Rubicam's new five-factor Brand Asset Valuator measure (details publicly available in the *Journal of Marketing Research*, 2008, 45(1), 15–32, though relevant to management, organizational behavior, and finance). In the process, construct better items and scoring for the BAV. (13 points)

(7.4) Researchers Liao et al. —see the *Journal of Applied Psychology*, 2009, 94(2), 371–391—set out to measure the effects of "high-performance work systems," in the banking industry, on customers' perceptions of service quality. In their measures and analysis, what did they do right and what did they do wrong? (9 points)

(7.5) Explain why "median splits" should not be used. If you want to dichotomize raters into two groups, where should you make the split for (a) a unipolar attribute—give an example, and (b) a bipolar attribute—give an example. (3 maximum points for the general answer, and 4 each for (a) and (b), thus 11 points total)

(7.6) Evaluate the measures *and* the scoring rules in Mandeep Dhami's very important experimental investigation of what "reasonable doubt" means to jurors in criminal trials —see the *Journal of Experimental Psychology: Applied*, 2008, 14(4), 353–363—. (11 points)

(7.7) Find and read Richard Coleman's (1983) excellent review of social class measurement and scrutinize his recommended four-component measure. (a) What scoring rule does it use and why? (3 points) (b) Which of the weights would you modify—in today's Western society—and why would you not change the other ones? (7 points)

Chapter 8
Qualitative Research from a C-OAR-SE Perspective

'Simple' causal models are logically wrong, and the empirical estimation by LISREL or some other software is not going to 'confirm' them.

– Gilles Laurent, Champion of the original C-OAR-SE article

We believe the data; we don't care about the truth.

– Jeffrey Deaver, *Broken Window* (2008, p. 95)

Qualitative research is always the output of the fruitfulness of a human mind, both in generating hypotheses and in being insightful enough to select measures that will test the hypotheses.

– Muckler and Seven (1992, p. 442)

After reading this chapter you will learn that:

- Qualitative research—the most important type of research by far—is necessary to formulate a theory (which consists of constructs and causal relationships between constructs)
- Analytic qualitative research, AQR—which requires the researcher to be able to conduct skilled open-ended interviews with diverse rater entities and then apply deep-level introspection to formulate a mini-theory of action—is the only valid method of qualitative research measurement
- Other methods appearing in the qualitative research literature—notably "grounded theory," "exploratory qualitative research," and "interpretive qualitative research"—should be rejected
- You should commission AQR from an expert practitioner and according to the C-OAR-SE-based guidelines in this chapter—because very few academic researchers are expert enough to conduct this vital type of measurement themselves

The opening quotation, from Laurent (2000), epitomizes the "statistical tail wagging the conceptual dog" problem, which is almost totally due to the neglect of, and ignorance about, *qualitative research* in the social sciences. Gilles Laurent was the "champion" of my original C-OAR-SE paper for *IJRM* (after it was rejected

J.R. Rossiter, *Measurement for the Social Sciences*,
DOI 10.1007/978-1-4419-7158-6_8, © Springer Science+Business Media, LLC 2011

by the leading marketing research journal, *JMR*, as too radical) and I am eternally indebted to this friend, scholar, and delightful French gentleman. He and Bobby Calder (1977), a focus-group practitioner before he became a full professor of both psychology and marketing at Northwestern University, are the only two academics who have realized and spoken out about the vital role of qualitative research for *theory-building* in the social sciences. Qualitative research is naively bypassed in social-science research textbooks. The most likely and widely used book, Kerlinger and Lee's (2000) *Foundations of Behavioral Research*, devotes a paltry three of its more than 900 pages to it, and then dismissively.

Qualitative research—and note that *introspection by the researcher* is essential in qualitative research—is absolutely necessary for theorizing about the *causal relationships* between constructs, and for *identifying constructs* in the first place. An example I gave earlier is the definition of the JOB SATISFACTION construct as either COMPONENTIAL JOB SATISFACTION or OVERALL JOB SATISFACTION, a difference which matters depending on whether the theory is about the causal relationship between *specific aspects* of the job (Work hours, Salary, Colleagues, etc.) and, say, INTENTION TO STAY, or is the theory that a *global feeling* of OVERALL JOB SATISFACTION (a global feeling that may be persistently salient every time the employee gets up to go to work, as in the commonplace thought extremes of "I love my job" or "I hate my job") is what mainly motivates INTENTION TO STAY. The effective predictor construct, and the causal process by which it operates, can only be validly discovered through *qualitative* interviews with EMPLOYEES in which the RESEARCHER also serves as a respondent by engaging in *introspection*. The researcher's own experiences will tend to tip the balance toward specifying one theory or another.

In *quantitative* research, the ogre-like nemesis of qualitative researchers, especially nowadays with the trendy travesty of structural equation modeling (which I earlier called "silly empirical meddling"), causal relationships are simply *assumed* and *drawn in* between constructs in *box-and-arrow diagrams*. The direction of the arrows (of the inferred causality) is added *afterward* depending on empirical statistical tests (e.g., "model fit" tests, of which there are alarmingly many to choose from, and you're bound to find a test statistic that shows "fit"). This *quantitative* approach is devoid of *theory*. Also, learning theorists have argued convincingly—see Hilgard's *Theories of Learning*, especially the 1956 edition that I learned from—that there cannot be response → response (R → R) causal relationships, which is what box-and-arrow diagrams imply. There can only be stimulus → response (S → R) causality, as in paired-associates learning and classical conditioning, or response-produced stimulus (R → S) causality, as is operant learning. A simplified ("cleaned up") example of a "conceptual model" is shown in Fig. 8.1. This box-and-arrow model is supposed to be a mini-theory of the emotional state of "FLOW" EXPERIENCED BY COMPUTER USERS (it is loosely based on an actual published model, whose authors I won't name, which had 24 constructs in the boxes!). This is really no more than an empirical model. In box-and-arrow models such as this one, theory is either absent or is supplied *post factum*, whereas it should be supplied from qualitative research.

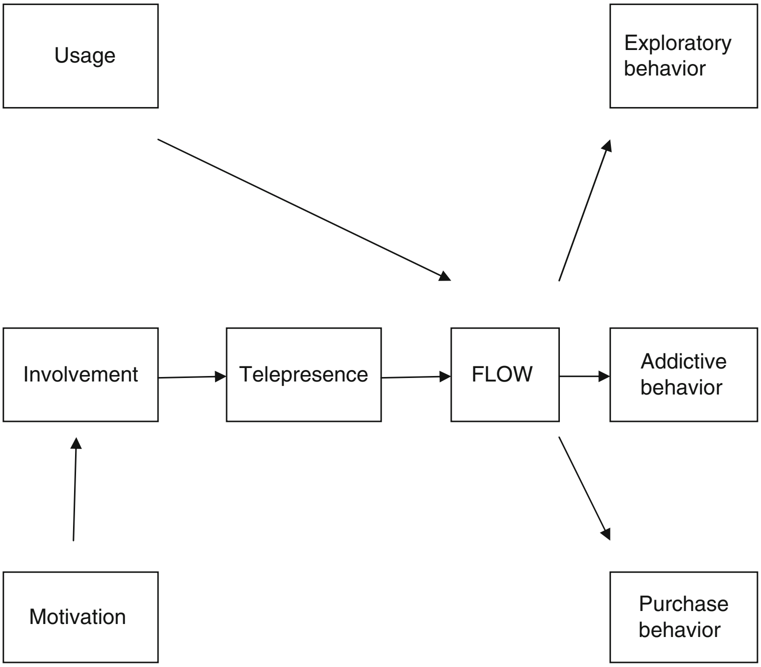

Fig. 8.1 A "cleaned up" example of a box-and-arrow conceptual model (the original had boxes for 24 constructs)

In *all* social-science fields there has been a scandalous neglect of theory in recent times. In psychology, where I did my undergraduate training, I believe this neglect has come about, first, because social-science students don't get taught the "battle of the learning theories," as was taught to psychology students in the 1960s and 1970s from Hilgard's (1956) famous textbook and, second, because social-science students are neither taught nor encouraged to *introspect*, that is, to think up theories themselves. This incredibly thorough textbook, with its sophisticated theoretical debates, was the best way to learn *theory*, not just learning theory. Later texts on comparative theories of attitude formation-and-change helped (especially Greenwald, Brock, and Ostrom's (1968) edited book on attitude theories, though that book did not much comparatively *evaluate* the theories, unlike Hilgard's book). No psychology text was as instructive as Hilgard's book on the great theories of learning—and apart from broad innate dispositions, all human behavior is *learned*. Classical (now "human evaluative" as contrasted with Pavlovian) conditioning, paired-associates learning, and operant learning (with an initiating "discriminant stimulus" and a following "reinforcing or punishing stimulus") continue to be the main—and as far as I know the *only*—causal mechanisms by which causal *relations* between constructs can arise.

How did the great early psychologists such as Thorndike, Guthrie, Hull, Tolman, Lewin, and Freud (the psychologists whose theories are discussed in Hilgard 1956) arrive at their descriptions of these causal processes? Well, for the first four of these theorists, it was pure *introspection*, since those theorists mainly studied lower animals who can't be interviewed. Lewin and Freud, on the other hand, talked to *people*. However, they *inferred* their theories, just as the animal research psychologists did.

Qualitative researchers have to use *aided introspection* to formulate their theories. (Those in the more social of the social sciences tend to use what they call "grounded theory," an oxymoronic term which is intended to absolve the researcher of much thinking and of any personal responsibility for the final theory.) *Aided introspection* allows (a) identification of the relevant constructs and (b) theorizing of causal relationships between them, including estimation of the causal importance of each construct that precedes another. Aided introspection to achieve these two components of a *theory* is a much more complex endeavor than any solely *quantitative* researcher, a label that covers psychometricians, ever had to undertake. As John Locke theorized long ago (1690) in his famous treatise "Of the Association of Ideas," only simple ideas come from external "sense-data" whereas complex ideas come from the "inner experience" of thought (I am indebted to Mautner's, 2000, *Dictionary of Philosophy* for this interpretation of Locke's insights). Four hundred years later, another Locke (2009) makes much the same point about the value—and under-use of—introspection.

Qualitative research necessarily involves *measurement*. This chapter spells out the nature of the measurement properties of validity and reliability as these terms apply to qualitative research and also explains how *valid and reliable* qualitative research can be done. The "bad news" punchline is that in about 98% of cases only *professionals* should do it. There are very few academic researchers who have enough *natural* "people reading" ability and enough knowledge of what a *theory* involves to qualify, and that's a sad fact.

I had to exclude qualitative research in the original C-OAR-SE article (see Rossiter 2002a, p. 308, note 3) because coverage could not be fitted in to its already long, for a journal article, 30 pages. Well, here is that coverage. Thanks to my long-time friend and colleague, Les Johnson, much of the coverage was later published in Rossiter (2009b), but unfortunately in a somewhat inaccessible Australian research journal. This chapter really should have been the first or second chapter in this book. It isn't, because I didn't want to turn off the "quant jocks" for whom this book is most necessary.

8.1 Types of Qualitative Research

The only type of qualitative research useful for theory-building is *analytic qualitative research* (which I will frequently abbreviate in this chapter as *AQR*). I will explain what analytic qualitative research entails in the next section, but first I want to dismiss two other types of qualitative research, namely, "exploratory" qualitative research and "interpretive" qualitative research.

The meaning of qualitative research as "exploratory" is appropriate if the researchers intend it only to be an input for subsequent quantitative research, such as to generate items for a questionnaire survey, or to construct stimuli for an experiment. This type of qualitative research should really be called *pre-quantitative research*. In this unfinished "exploratory" capacity, pre-quantitative research does not qualify as qualitative research.

In recent years, the term "qualitative research" has assumed a more philosophical and I would say *sociopolitical* meaning among those academics who equate qualitative research with "interpretive" research (see Tadajewsky 2008). So-called *interpretive research* is a form of research that is based on qualitative interviewing methods, including self-interviewing (i.e., introspection). However, it has as its purpose the understanding of behavior as an "end in itself," rather than the philosophically "positivist" research purpose of the understanding and then prediction of behavior (see Hudson and Ozanne 1988), which is what social science is all about. Even worse is that the "interpretist" school rejects quantitative research—which I certainly do not, as evidenced in this book.

All qualitative research involves interpretation, indeed "heavy interpretation," as will be explained in this chapter. The problem is that the purported episodes of "understanding" reported by most exponents of the interpretist school fall far short of ending up as explanatory, testable theories. As Mick (1997, p. 259) has commented with regard to interpretive research reports: "Reaching the end of the their articles the reader is often hard pressed to know exactly what new knowledge has been contributed through the exercise of semiotics' esoteric terminology or techniques." One of my favorite realists, academic, and practitioner Bill Wells (1993, p. 498), asked the following terse question about much academic research: "So what?" This question needs to be answered by the interpretist school.

8.2 Analytic Qualitative Research (AQR)

The label of *analytic qualitative research* (Moran 1986) describes the widest and most valid use of qualitative research by *practitioners*. Analytic qualitative research (AQR) is a procedure consisting of a data collection methodology and a mode of analysis for deriving insights into human behavior. The results are reported in the form of a proposed causal explanation of behavior—a *mini-theory of action*—with the theory's action recommendations to be tested in the real world. The action recommendations depend ultimately, for validity and reliability, on the ability of the individual qualitative research analyst.

The professional purpose of AQR is not to produce theory or even results that could be generalized beyond the specific behavioral situation. In fact, those who commission professional qualitative research—analytic qualitative research—seek a *unique* mini-theory that competitors cannot easily imitate (see Rossiter 1994). This places analytic qualitative research in marked contrast with the purpose of quantitative academic research, which is to produce either *broad* contingency theories or to identify empirical generalizations. The professional analytic qualitative researcher must come up with a theory that will work now, for the issue or product at hand.

Because "analytic insights" specifically and "mini-theories" more generally are major outputs of analytic qualitative research, it is useful to point to some examples of these. The specific insights or more comprehensive mini-theories could not have been obtained from quantitative research, that is, by measuring what people say or do, counting it, and interpreting the results literally.

Various examples of qualitative research *insights*—from the field of marketing, to which I have best access—are:

- Toilet Duck's package innovation (U.K.) where the research input was focus groups.
- Benetton's "shock tactics" advertising campaign (which ran in Europe for many years but was banned in the U.S. as too controversial) where the research input was the creative director's introspective selections from his personal collection of professional photographs.
- U.S. Post Office's "We deliver for you" advertising campaign, which profiles letter carriers personally, based on the inference from participant observation by social anthropologists that many people symbolize the mail deliverer as a means of contact with broader society and even as an antidote to loneliness (Levin 1992).
- California Milk Board's (and later nationally adopted) "Got milk?" advertising campaign, based on the insight from focus groups that people really only ever think about milk when they need it and it isn't there, and then the identification, from a post-group experiment with the consumers in which they were asked not to drink milk for a week and then report back, of which foods they most missed milk with, such as cereals, cookies and coffee, foods which were then featured in the ads (Morrison, Haley, Sheehan, and Taylor 2002).
- McDonald's "I'm lovin' it" advertising campaign criticized by the experts and by consumers in a quantitative ad test, but shrewdly and successfully introduced to overcome consumer resistance to McDonald's "nutritious menu" items (Rossiter and Bellman 2005).
- Numerous instances of re-weighting of attributes rated low in survey research, but inferred from focus groups or individual depth interviews to be *highly* important, such as peer acceptability in beer brand choice (Rossiter and Percy 1997), or taste in toothpaste as a surrogate indication that the toothpaste is cleaning and freshening the mouth (Langer 1984).

Other compelling examples of analytic qualitative insights are given in Durgee (1985), Calder (1994), Rossiter and Bellman (2005), and Zaltman and Zaltman (2008).

There are, of course, thousands of one-off *mini-theories* in professional qualitative research reports that could be cited, if we had access to them. The author's own contributions include proprietary reports such as a beer brand choice model developed for Anheuser-Busch; an advertising launch plan for Stouffer's Lean Cuisine in the U.S.A. and later used almost identically in Australia (branded Findus); a consumer-market brand choice model developed for Rockwell's power tools; and a persuasion model commissioned by the (U.S.) National Potato Promotion Board (cited in Rossiter and Percy 1987, 1997). I put myself through grad school (Ph.D.) at Penn and then supported my young family with a nice lifestyle by doing professional qualitative research.

These mini-theories that I came up with were product- and time-specific, and tested in real-world campaigns with successful outcomes. They are examples of the science of *analytic* qualitative research.

Turning to broader *theories*, in which a large number of insights are combined and interrelated, it is evident that most of the influential theoretical models in marketing, such as Howard's (1977) EPS/LPS/RRB product lifecycle-based series of models, or Bass's (1969) diffusion model, and in advertising, the FCB grid (Vaughan 1980) and the Rossiter-Percy grid (Rossiter and Percy 1987, 1997; Rossiter, Percy, and Donovan 1991), were essentially qualitatively derived. (And I wish my theoretical model of MARKETING KNOWLEDGE—see Rossiter 2001, 2002b—would become more influential!) No quantitative survey or experiment "produced" these models. Rather, they were the result of analysts' inferences from various sources, including other studies, everyday "anthropological" observation, and introspection. They represent a necessary continuation of the early tradition of introspection in psychology (see Locke 2009).

8.3 Methodologies of Qualitative Research

Four principal types of data collection methodologies are employed in qualitative research to collect consumer data (which I will call *data 1*). The first four are well-recognized in the social sciences as qualitative techniques (Walker 1985a). In market research, and especially in advertising research, group-depth interviews and individual-depth interviews are by far the most prevalent methodologies, although the others are used occasionally. The four data collection methodologies are:

1. Group depth interviews (commonly called focus groups)
2. Individual depth interviews (including individual interviews with company personnel in organizational case studies)
3. Participant observation (including "anthropological" or "ethnographic" studies, the latter a politically incorrect label if ever there was one!)
4. Projective techniques

A comparison of the four types of qualitative research in terms of their data collection methodologies is shown in Table 8.1. Four attributes of the methodologies are identified: the number of consumers per interview, total consumer sample size, the consumer's (respondent's) role during the interview, and the analyst's role as question-asker during the interview. These attributes are used to compare and evaluate the methodologies.

Table 8.1 Comparison of the interview methodologies of qualitative research

Type of qualitative research interview	Respondents per interview	Total respondents	Respondent's role	Analyst's role as question-asker
Group depth interviews	2–12	Any number	Active	Active
Individual depth interviews	1	Any number	Active	Active
Participant observation	Any number	Any number	Passive	Active
Projective techniques	1	Any number	Active	Active

Group depth interviews, also known as GDIs or "focus groups," are the most widely practiced type of qualitative research. Calder (1994) has estimated that about 700 focus groups are conducted each day in the U.S.A. alone! Focus groups probably account for about 50% of social and commercial market research projects, although their low cost means that they constitute, as very conservatively estimated for the U.S. market by Baldinger (1992), only about 20% of market research expenditure (the percentage is much higher in smaller countries and less-developed countries). Group depth interviews typically employ 2–12 consumers per interview. The smallest number, two consumers, to form a group, known as a "dyadic group," is employed quite often with husband–wife or cohabiting-partner interviews. Another commonly used number of interviewees is four, in what is known as "mini-groups," which are used when the researcher wants to obtain more information per individual than with a larger group, although often the total time for mini-group interviews is reduced from the usual 2-hour group interview to about 1 hour. Focus groups usually employ about 8–10 consumers per interview, occasionally going as high as 12 if the question-asker (group moderator) feels capable of handling a larger group; larger groups are sometimes requested by clients who confuse the number of respondents per group with the quantitative projectability (reliability) of the results. Any number of interviews, totaling any number of consumers, can be conducted, although there is a theoretical upper limit as explained below. The interview "unit" in group depth interviews is the *group*, so the number of groups to be conducted, rather the total number of consumers interviewed, is the relevant methodological decision variable (see Lunt and Livingstone 1996).

The following comments about the number of GDIs that should be conducted apply also to the number of individual interviews in the individual depth interview (IDI) method, other aspects of which are discussed shortly. In theory, the number of groups or interviews to conduct is governed by the judgment of the analyst, who assesses the point of which the "marginal insights" from successive interviews seen to be leveling off (Walker 1985b). This is aptly called "theoretical saturation" by the "grounded theory" school (Glaser and Strauss 1967). In the question-asker's role, the point of theoretical saturation is fairly evidently reached during later interviews when the question-asker can almost exactly predict what the interviewee's answer to the question will be (Lunt and Livingstone 1996). If the question-asker is also the analyst, which is ideal theoretically although professionals differ on this procedure in practice (McQuarrie 1989), then "hearing the same information" will be reasonably equivalent to the likelihood that the marginal *insights*, which are *analytic inferences*, have approached their asymptote. In practice, it is only the rare client who is willing to let the number of groups or number of interviewees be "open-ended," with the total to be decided by the analyst. Usually, a fixed number of groups of interviews is decided with the client in advance. A rule of thumb that works well in practice (Rossiter and Percy 1987, 1997) is to conduct three groups, or about 12 individual interviews, per segment, if the target market is known *a priori* to be segmented, or just this number in total in an unsegmented market. Should a new, heretofore unrealized segment or segments emerge during the course of the initial interviews, then another set of three groups—or 12 more individual interviews—should be added. The reason for three *groups* is that, because of group dynamics, it is quite likely that the first two groups will drift in contradictory directions and thus a third group is needed as a "tiebreaker" to resolve conflicting data for the analyst to make inferences from. In terms of the final two attributes in the table, the respondent's role is "active" in group depth interviews, as is the role of the interviewer.

Individual depth interviews, or IDIs, quite obviously consist of a single question-asker and a single question–answerer on each question-asking occasion. The main methodological difference between group depth interviews and individual depth interviews is that, in the group setting, the question–answerers interact with each other, rather than with just the question-asker, in providing answers to questions. That is, the participants' answers are presumably not the same as if they had been interviewed individually. In IDIs, the answers are in no way peer-dependent and usually are assumed to be more forthcomingly personal than in the group setting. The questions in IDIs are exactly the same as in GDIs.

Individual depth interviews are employed, rather than group interviews, in several well-known circumstances (McQuarrie and McIntyre 1990, Rossiter and Percy 1987, 1997). One is where the analyst's prior knowledge of the category indicates that the purchase (or other) decision of interest is made largely by the individual acting alone. This includes *personal-product* decisions, such as in the personal health and hygiene area, where consumers would not be comfortable talking about their personal needs in a group setting. An extreme application is the use of hypnosis of individuals to try to elicit "deep" product experiences (Cuneo 1999). A second well-known use of individual interviews is when the qualitative researcher has to interview *high-level professionals* such as medical specialists or very senior executives—the sorts of individuals who would not readily come to a group discussion. For this purpose, "executive interviewers" are often employed as the question-askers—these are the research firm's most skilled individuals in obtaining and successfully completing these hard-to-get interviews. A third use is when a *developmental history* is sought to find out how consumers arrived at their current state of knowledge and attitudes in the category (e.g., Fournier 1998). I did developmental history IDIs every 5 years or so for Stouffer's in the United States

(Findus elsewhere) to map the decline of women's cooking skills—and later the isolated rise in men's—which of course favored purchase of "gourmet" frozen-food products. Because of the longitudinal questioning, developmental history gathering is much better suited to IDIs than to group interviews. A final use of individual depth interviews is in *pre-quantitative* qualitative research, which, as discussed previously, is not really qualitative research as its purpose is only to formulate items for a quantitative survey. Griffin and Hauser (1993) have shown that it is much more cost efficient to use individual, rather than group, interviews to identify important attributes for survey items.

Participant observation is a type of qualitative research, long established in social anthropology and more recently adopted in market research (see Belk 1991, Levin 1992), in which the question-asker immerses himself or herself in the question–answerer's natural social and cultural milieu. The idea, of course, is for the question-asker, as both asker and analyst, to walk a fine line between participation (outwardly) and detachment (inwardly) so as not to be an overly reactive part of the measurement process. Participation is necessary to obtain measurement in the first place, given that *unobtrusive* observation (Webb et al. 1966) would provide an inadequate understanding. But detachment is necessary, also, to "factor out" the analyst's participation.

In participant observation, any number of respondents can be "interviewed" on any single occasion, and in total, but with a theoretical (and a practical) upper limit as explained above for GDIs and IDIs. Respondents per interview may range from one respondent per interview, such as in "shill" shopping whereby the question-asker poses as a shopper in order to observe and in a sense "interview" the salesperson, to many respondents per (unobtrusive) interview, used in the anthropological type of setting as in the Consumer Behavior Odyssey (Belk 1991). The total number of interviews is arbitrarily decided. The respondent's role in participant observation is different from the other types of qualitative research interviews in that the respondent should not be aware that an interview is taking place. In the shill-shopper situation, for example, the respondent (a salesperson rather than a consumer) is unaware that he or she is being interviewed. In the more anthropological situation, it is assumed that the question-asker has previously established rapport and become part of the social group to the point where its members don't feel they are being interviewed or observed, although this perception may not be removable entirely. The respondent's role in participant observation in Table 8.1 is, therefore, classified as "passive" rather than the respondent as an active answerer of questions as in the other qualitative interview methodologies. However, the question-asker's role is "active," because rarely is participant observation simply "observation." Questions have to be asked to clarify what is going on, and also to test analytic inferences as the research proceeds; all this is done in a fairly unobtrusive manner but can hardly be described as passive.

Projective techniques are included as a method of qualitative research because they use evolving questions, open-ended answering, and heavy interpretation, and the results are formulated as a mini-theory. In the Rorschach Inkblots test, for instance, the interviewer first asks the respondent, "What do you see?", and then follows up with neutral probe questions such as “What else?” In the Thematic Apperception Test (TAT), the interviewer asks the respondent to “Tell a story” about each TAT picture and probes each initial answer with “Tell me more.” However, in Speaker Balloons, a commonly used projective technique in market research, the respondent merely receives an initiating question from the analyst or interviewer, along the lines of “What might this person be saying?” but even this could be regarded as an evolving question because the respondent is likely to mentally ask further questions such as “Why is the interviewer asking me this? What does the interviewer really want to know?” Another commonly used projective technique in market research is Sentence Completion, such as “People drink Coca-Cola because ________________________.” With Sentence Completion, the first part of the sentence forms the question and so this particular projective technique less clearly meets the criterion of evolving questions, and indeed the question can be administered via a self-completion survey, but it does produce open-ended answers and require interpretation. All projective techniques require the analyst’s interpretation of the extent to which the consumer’s answer reflects a *personal* explanation, and thus is “projective.”

Projective techniques employ one respondent per interview and any total number of interviews. In projective techniques, the respondent’s role is active, as the interviewee responds projectively to the ambiguous initiating stimulus. The question-asker’s role is also active because he or she not only provides the initiating stimulus, but also formulates the evolving questions when projective techniques are used in individual depth interviews.

8.4 First-Order and Higher-Order Data

What emerges from qualitative research interviews, from each of the four interview methodologies, are *first-order data*. This is Schutz’s (1967) term to describe the immediate “surface” information obtained from the interview, that is, the open-ended answers, or what can more simply be called *data 1*. *Higher-order data* are the analyst’s *interpretations* of the first-order data, which can be called *data 2*. The nature of higher-order data will be discussed in the next section.

First-order data (data 1) produced by qualitative interview methodologies are extremely messy. The open-ended answers are mainly “what is” descriptions and “why” reasons (the latter as detected by the respondents themselves). The label “messy” is justified for two reasons. In the first place, the open-ended answers may, of course, be organized by question areas, but there is not much more structure to the data than this. Eventually, the data have to be organized into a model that will little resemble the question order. The second reason for messiness is that in all but some forms of content analysis, the first-order data are *multi-modality data*. Group depth interviews, the most prevalent methodology, provide first-order data in all of five modalities: verbal (what is said), paraverbal (how it is said), facial-visual (nonverbal expression), somatic (“body language”), and intersubjective (group dynamics). The extra modalities' contents, not just verbal content, are part of the data 1 that the analyst receives.

The question-asker is not just a question-asker, but is also unavoidably an analyst. The question-asker is *simultaneously* adopting the role of answer-interpreter. In order to know which further questions to ask and what to probe, the question-asker must simultaneously be assessing the incoming first-order data. This is true even for the most "neutral" probes, such as "Tell me more about that;" numerous studies by Wilson and his colleagues (for a review, see Wilson, Dunn, Kraft, and Lisle 1989) have demonstrated that asking people to reflect on or explain their attitudes, such as their liking for a particular product, gives *lower* prediction of their behavior than had they not been probed. In other words, the question-asker has to know "when to stop," or what to later "discount." This is another reason why, for highly valid results in qualitative research, the analyst should also be the question-asker. The questions and answers are open, reactive, often interactive, and typically are provided in multiple modalities. Even if the sample of respondents could be duplicated, their answers couldn't. And this is *prior* to the first-order data being interpreted!

The next phase of qualitative research is the analytic procedure, in which first-order data are interpreted in higher-order terms and become data 2.

8.5 Modes of Analysis in Qualitative Research

No matter which type of qualitative interviewing methodology is employed, there is always a choice of analytic modes. The alternative analytic procedures in qualitative research are *in general* compared in Table 8.2 (only the *first* and *last* are used in AQR). Four analytic modes are identified:

Table 8.2 Modes of analysis employed to derive inferences (data 2) from qualitative research

Qualitative research analytic procedure	Total analysts	Background knowledge needed	Upper limit of interpretive skills	Range of results possible	Typical predictive validity
Analyst's content analysis*	1	Social-science clinical theory plus psychology plus knowledge of the subculture	Very high	Extreme	Moderate to very high
Coders' content analysis	2—10	Subject matter only	Very low	Small	Low
Computerized content analysis	1	None except lexical	Virtually zero	None	Very low
User observation*	Total users	Commonsense marketing experience	Moderate	Quite large	Moderate

*Analytic procedures used in AQR

1. Analyst’s content analysis
2. Coders’ content analysis
3. Computerized content analysis
4. User observation

These analytic procedures are compared in terms of five attributes, namely, the number of analysts, the background knowledge involved in the analysis, the estimated upper limits of interpretive skills in each type of analysis, the estimated range of results possible with each type, and the estimated predictive validity of each.

As I pointed out previously, qualitative research analysis is increasingly equated in academic writings with “grounded theory,” a description proposed by Glaser and Strauss (1967) in their widely cited book. This is usually taken to mean that qualitative research is a “theory of social action” (such as the action in a particular area of civic, political, organizational, or consumer behavior) based on, or “grounded in,” the *experiences of the participants*. The description is incomplete and misleading. The theory—the result—that emerges is much more “grounded” in the *experiences of the analyst*. It is wrong to believe, as many commentators apparently do, such as Calder (1977, 1994) and McQuarrie and McIntyre (1990), that qualitative research results can consist of completely atheoretical “participant phenomenology.” There is *always* an analyst involved as part of the *measure* whenever phenomenological reports (which are data 1) are received and recorded. This point is far too little appreciated, especially by academic qualitative researchers who refer to the objective-sounding term “grounded theory”—a similar clichéd term is “evidence-based”—without realizing that the analyst is a very large part of that theory and that the “grounding” is, therefore, not what they think it is. The importance of this point will become clearer in the following discussion of the four qualitative research analytic modes and also in the final section of the chapter, which examines qualitative research in the light of psychometric quantitative measurement theory.

Analyst’s content analysis. In analytic qualitative research, the interviews are analyzed by an independent, professionally trained researcher who interprets the first-order data (data 1) in terms of *higher-order concepts* consisting of an overall theoretical framework, major variables, and inferred causal relationships between the variables (data 2). Occasionally, in very difficult research situations, some clients employ two or three professional analysts independently, who often also conduct their own qualitative research interviews separately. In the overwhelming majority of studies, however, there is only one analyst. For marketing applications of analytic qualitative research, the analyst ideally should have a strong background in *psychology*, so as to be able to interpret the first-order data and convert them to higher-order concepts. Also, the analyst should have a strong background in *social-science theory*. The state of qualitative research in practice is that most analysts have a good background in psychology with a clinical emphasis—or are intuitively “natural” psychologists—but too few are up to date in terms of theory (Gordon 1997). The analyst who can combine psychological knowledge with an extensive and current knowledge of social science theorizing is much more likely to achieve high predictive validity—provided the analyst also has very good knowledge of the subculture(s) of the respondents. When I was a qualitative researcher in the

U.S., for instance, I never conducted focus groups with all Blacks, all Hispanics, or all teenagers, and neither did I resume qualitative research consulting when I returned to Australia because after 15 years as a U.S. resident, I had become far too "Americanized" (or at least "White Americanized"). I am firmly back in mainstream Aussie culture now, and could now qualify as an analyst in my own country.

With analytic qualitative research, AQR, the most important form of qualitative research, the client is essentially "buying the analyst," not simply buying the interview data. The best analysts, as evidenced by their historical high success rate, charge very high fees, and these fees are justified for all but the most mundane research topics. It follows that the upper limit of interpretive skills of analysts is "very high." However, it also follows that analysts differ widely in abilities and thus the range of results possible from qualitative research is "extreme." (Here's a hint for instructors. The author has demonstrated the variability of analysts' results numerous times by assigning students in advertising research classes into teams to jointly conduct a focus group; then, using duplicates of the audiotape or videotape of the group discussion [the first-order data], the team members as individuals have to interpret the results and write reports. The range of results—findings and recommendations—is always extreme, with grades ranging from A+ to F, yet the results are based on identical data 1! This many-times replicated result led to the proposition, discussed later, that analytic qualitative research results are characterizable as about 50% respondents' data and 50% analyst's interpretation.) The author has also read many professional qualitative research reports on behalf of clients and arrived at a similar conclusion: that the range of insightfulness and usability of the results shows extreme variation from poor to very high.

However, because the client is obtaining an independent viewpoint, even a poor qualitative research report will generally be quite helpful, mainly because the manager—that is, the *user*—steps in as the *analyst*. The manager interprets and converts the findings into action recommendations. For this reason, the predictive validity of analytic qualitative research is assessed as ranging from "moderate to very high," rather than very low to very high. Nearly always, clients will state that "some value" has been obtained from the qualitative research even when an experienced qualitative researcher reading the report would class it as poor. And the client may be right—if the client is a good *analyst*.

Coders' content analysis. Sometimes, multiple coders are employed as low-level analysts of qualitatively derived interview data. The use of multiple coders is typical in academic studies where the content-analytic convention of "inter-coder reliability" is well established. To establish inter-coder reliability estimates, between two and ten coders in total may be employed (Perrault and Leigh 1989, Rust and Cooil 1994) but usually there are just two or three. The background knowledge brought to the analytic situation is "subject matter only," in that coders are temporarily trained "on the data itself" and no prior psychological or marketing knowledge is assumed or typically present. The upper limit of interpretive skills when using multiple-coder content analysis is therefore "very low." The range of results possible is "small"—which is a paradoxical but logically necessary outcome of achieving high inter-coder reliability. If the coders' findings are taken at face value with little further analysis, as they most often are in academic qualitative studies, the typical predictive validity is "low."

Computerized content analysis. An increasingly favored mode of content analysis, especially in academic qualitative studies, is computerized content analysis, using computer software programs such as NUD*IST$^{\mathrm{TM}}$, N-VIVO$^{\mathrm{TM}}$, or LEXIMANCER$^{\mathrm{TM}}$. The researcher or research team has to set up the content codes initially, as in coders' content analysis, but thereafter the computer merely does a blind and mechanical numerical count according to the pre-established codes. The computer needs no background knowledge other than lexical (word form or parts of speech) discrimination and there are no marketing or psychological skills in the computer's interpretation (contrary to the *deus in machina* hope of computerized content analysts). An example is the computerized content analysis of corporate advertising slogans by Dowling and Kabanoff (1996). The analyst has to select the content categories to be coded by the computer in the first place and computerized content analysis requires heavy subjective interpretation—contrary to the belief that it is the most objective of all types of content analysis. For instance, the largest category in Dowling and Kabanoff's study, into which 50% of the slogans were placed, was "Equivocal," that is, uninterpretable by the computer (1996, p. 71)! Because of its use of the computer, computerized content analysis gives the semblance of being more "scientific" than other forms of content analysis but it is actually the least scientific and the least valid.

Computerized content analysis is little more than a quantitative summary of first-order data, a simple frequency count of data 1. The upper limit of interpretive skills is "virtually zero" though not entirely zero because there has been some analyst's input initially. It should go without saying that the typical predictive validity of computerized content analysis used in qualitative research is "very low."

User content analysis. By "user" is meant the marketing manager or government policy maker or, in applications of qualitative research in advertising, its most frequent applied use, the creative team or copywriter at the advertising agency. User observation is the method of analysis employed by definition with the so-called *phenomenological* type of qualitative research (Calder 1977) in which the interview data from the participants (respondents) are taken at face value. The data 1 are the interviews themselves (such as focus groups) and are observed directly by the user in person, on videotape, on audiotape, or indirectly from a summary report that is quite literal (with minimal higher-order interpretation). With phenomenological qualitative research, the *user* must act as the analyst and the data 2, the inferences, are never made explicit but remain implicit in the plan that is implemented.

User observation is not, however, confined to the phenomenological type of qualitative research. User observation almost always enters as an additional mode of analysis prior to application of qualitative research findings when the *analytic* type of qualitative research, AQR, is conducted in which higher-order findings from an independent analyst—the qualitative researcher—are available to the user. It is, therefore, very important to examine what the user brings to the analytic process.

For user observation, there may be several users physically observing an interview at one time, as when a number of managers, or creative people from the advertising agency, observe focus groups or individual depth interviews directly, typically from behind a one-way mirror or on videotape or DVD. The total number of analysts is equal to the total number of users. The background knowledge brought to the analytic situation by the user is described as "commonsense marketing experience." This description fits in most cases, in that very few managers have extensive training in social-science *theory*. It is also rare for users as analysts to have extensive psychological training; hence, the upper limit of interpretive skills for user observation would be "moderate." Because different managers will bring to the analysis different levels of marketing experience and also different levels of "natural" interpretive ability, the range of results possible via user observation as an analytic procedure is assessed as "quite large."

When user observation is the sole analytic procedure, as with phenomenological qualitative research, the typical predictive validity of the findings is assessed as "moderate." The degree of predictive validity with the user-as-analyst is constrained by a presumed limited ability to develop higher-order conceptual insights (if the typical user had this ability, there would be no need to employ an independent professional analyst). On the other hand, the user's ability to *apply* the findings, drawing on marketing experience and marketing skills, should be quite high. Users might, therefore, be typified as being able to apply lower-level, first-order data very well. If the research "problem" that led to the research in the first place does not require a "deep" solution, then user observation alone, with the user's subsequent analysis being translated into marketing action, will be sufficient for high predictive validity. Most situations in which qualitative research is called for, however, have a far more complex causal structure which users are usually not able to detect in full and thus the overall assessment of the predictive ability of user observation as "moderate" seems justified.

A special form of user observation *not* shown in the table but well worth discussing occurs when people read mini-theoretical accounts, or "interpretations," based on qualitative research that are published in the literature. The reader is cast as a *user* and thus an analyst. Due to the unfortunate schism between the academic and practitioner literatures in the social sciences (e.g., Wells 1993), there are usually two situations here. One is where managers read practitioners', or sometimes academics' simplified-for-practitioners, theoretical proposals in the *practitioner* literature (in the British *Admap* publication, for instance, and in the U.S. *Journal of Advertising Research*, both of which have large practitioner readerships, with articles submitted by practitioners and academics—as do the *APA Monitor*, the *Australian Psychologist*, and other magazine-like journals). The manager-as-reader situation is exactly akin to user observation in the table and the table's classification of interpretive skills and predictive validity applies.

The other situation is where *academics* read (usually other academics') qualitative mini-theories. Well-known (to academics) examples of such mini-theories would be Gould's precedent-breaking article in the *Journal of Consumer Research* (1991) based entirely on his own introspection, and Fournier's more recent article in the same journal (1998) on brand relationships based on individual depth interviews. Who knows what analytic skills various readers bring to judging the worth of these mini-theories? It would seem that these accounts pass only the test of their internal logic (or else, presumably, they wouldn't be published in respected journals) and perhaps some sort of test of "empathy" with the reader's own experiences as an observer and, therefore, analyst. But these qualitative mini-theories are not yet science until they have passed an empirical test, something which only practitioners are usually in a position to provide via a marketing or advertising campaign. By this argument, it may be concluded that Ph.D. theses consisting of untested qualitative mini-theories are precarious as far as the normal Ph.D. criterion of "contribution to knowledge" is concerned.

The variables that can affect the vital contribution of the *analyst* in qualitative research may be identified as follows:

- The analyst's *comprehension ability*—including verbal ability and the probably intuitive nonverbal ability to "read people" (which Gardner 1983, calls "interpersonal intelligence" and others popularly call "emotional intelligence"—cf. the TV series *Lie to Me*, based on the psychologist Paul Ekman's research).
- The analyst's knowledge of *psychological concepts and processes*.
- The analyst's knowledge of *causal mechanisms* in social science.
- The analyst's *personal values*.
- And what can only be described as a *stochastic* element, in that different items of first-order data will be focused upon depending on the analyst's attention span, state of mind and perhaps physical fatigue, and other environmental quirks in the situational context occurring while the first-order data are being analyzed.

This list of variables should make it clear that the analytic process itself is highly variable across analysts except, of course, when low-validity forms of analysis are employed such as coders' content analysis or computerized content analysis. This means that the results will be highly variable, depending on the *analyst*. This essential realization is reinforced in the final section, where *analytic* qualitative research is compared with quantitative research in terms of the conventional criteria that are used to evaluate the worth of research.

8.6 Analytic Qualitative Research Compared with Quantitative Research

"Quantitative" research refers to structured-questionnaire surveys or laboratory-administered experiments providing *numerical answers*, which, after statistical analysis, are interpreted as the results. (In academic research, but not practitioner research, quantitative research measurement is synonymous with "psychometrics," the approach that I have criticized in this book).

Quantitative research can be compared with the most important form of qualitative research—*analytic* qualitative research, AQR—in terms of a simple model: C (consumer or other respondent data) + A (analyst's interpretations) = R (results).

Table 8.3 Comparison of analytic qualitative research and quantitative research in terms of sources of variance (%), which lead to the results (author's estimates)

	Consumers or other respondents (*C* factor)		Analyst (*A* factor)		Results (*R* factor)
Analytic qualitative research	50	+	50	=	100%
Quantitative research	90	+	10	=	100%

I will make the comparison in terms of my social-science field, marketing, where analytic qualitative research is most used. Table 8.3 offers a comparison of the "sources of variance" leading to the results in the two types of research. The reasoning underlying the estimated weights is discussed under the *C*, *A*, and *R* factor headings, next.

The C factor. In analytic qualitative research, the analyst's purpose is to build a mini-theory of the behavioral domain of interest (Walker 1985a, Moran 1986). In marketing, this theory might range from quite elaborate models of buyer or consumer behavior (see Howard 1977) to a much narrower set of recommendations for positioning and advertising a particular branded item. In constructing this theory, the "sample" that the analyst is drawing from in the first-order interview data is really a sample from the population of *ideas* rather than the population of consumers or other respondents. As Walker (1985a, pp. 5–6) puts it: "The units of analysis generally consist of ideas, experiences and viewpoints and the reported and logical relationships between them." This is stated directly by Lunt and Livingstone (1996, p. 92) in the case of focus groups: "The unit of analysis in focus groups is the thematic content of discourse used in the groups, not properties of the individuals composing the groups." Additionally, the analyst is sampling from his or her *own* ideas and experiences, via *introspection*. This second phenomenon is examined in the "A factor" later. Meanwhile, the contribution of the C factor in analytic qualitative research is estimated to be no more than 50%.

Realization that the relevant "C factor" in AQR is the population of ideas rather than the population of consumers or other respondents explains why analytic qualitative research, unlike quantitative research, should not be concerned with random sampling. In fact, to maximize the range and variety of ideas, *purposive sampling* should be employed. The researcher should deliberately recruit not only some average respondents, but also extremes such as very heavy users of the category, averse nonusers, users of "niche" brands, and so forth. This means that even so-called "professional respondents," or "groupies" as they are disparagingly called, are suitable as subjects in qualitative research; they are proven talkers and, having experienced many qualitative interviews (for which they must meet the respective product category screening criteria, of course), they are likely to be able to contribute more *ideas* than the typical "naive" respondent. More ideas means more *unique* and thus valuable ideas (Langer 1984, Rossiter and Lilien 1994).

The consumer (*C*) or other respondent data collected in an analytic qualitative study are ideas and, because of the people-sampling method and the varying ability of the analyst to elicit ideas from consumers, the data are almost impossible to replicate. The main reasons for this very low "reliability" of data 1 are:

1. The consumer or respondent sample employed in the research. (Research companies have very different "lists" from which they recruit qualitative research respondents.)
2. The question-asker's personal characteristics and also the perceived social relationship between the question-asker and the respondent, as perceived by the respondent, and to some extent as perceived by the question-asker. (Only in content analysis are personal characteristics not relevant.)
3. The actual questions asked (and not asked).
4. The degree and quality of probing of respondents' answers. (The probing in content analysis is seen in the addition of unanticipated content categories.)

It can be seen from the last two factors on the list that the *analyst*, not just the consumer (or other respondent), contributes to the quality of data 1, not just of data 2. The consumers' contribution (*C*) to the AQR results in the overall research equation $C + A = R$ is, therefore, justifiably estimated at no more than 50%. In fact, a great many marketing (including social marketing) plans are formulated by a single manager's qualitative introspection, *without* any consumer input at all, just as a great many advertising campaigns are formulated by one introspecting copywriter in an advertising agency (Kover 1995). Two classic examples of this were portrayed in *Mad Men*—Lucky Strike's "It's toasted," which came from an incidental comment by the client, and Kodak's "carousel of your life's memories" idea for the Carousel revolving slide-projector tray, suggested by the creative director. From a cynical perspective, consumers or other respondents—the *C* factor in analytic qualitative research—may be seen as just an input instrument for the analyst's inferences.

In quantitative research, by comparison, the consumers' or respondents' contribution to the results as providers of data 1 is more like 90%. The analyst converts these data 1 to a very simple form of data 2 (inferences) via statistical analysis. As argued below, the analyst's contribution to the results of quantitative research is only about 10%.

The A factor. The fundamental and most important difference between analytic qualitative research and quantitative research is the analyst's contribution (the A factor), which is major in analytic qualitative research and relatively minor in quantitative research. In AQR, the analyst's role is analogous to that of a clinician (Calder 1977), who observes and then infers to reach a diagnosis and recommendation. The biggest myth about qualitative research, including the analytic type, perpetuated in textbooks, academic journal articles, and increasingly by doctoral committees, which now accept qualitative theses, is that "anyone can do it." Practitioners know better (Gordon 1997). Just as clinicians exhibit differing abilities to correctly diagnose patients' problems, the analysts in qualitative research have differing abilities and, therefore, *differing predictive validities* (see also Westen and Weinberger 2004). This was demonstrated in the student studies referred to earlier, where each student, acting as a separate *A*, analyzed the same *C*-data, with widely differing results (*R*). It is certainly evident in the field of professional qualitative market research where some analysts are highly paid and highly sought after, based on their predictive track record, whereas other low-success analysts leave the profession after few attempts.

It follows that neither the content validity nor subsequent predictive *validity* of analytic qualitative research (AQR) results can be improved by *averaging* the interpretations of a highly expert analyst with those of one or two less-expert analysts. This is also the problem with trying to invoke "trustworthiness" as a sort of validity claim by dragging in other analysts to review the data and "confirm" the inferences (Erlandson, Harris, Skipper, and Allen 1993). Imagine, in a not too far-fetched analogy, trying to estimate a person's height by averaging the results from three judges, one of whom uses a tape measure, another just looks at the person and guesses, and the third plugs in the average height of people in the population. Only one judge would be right. Averaging (or seeking "consensus") is a false analogy with *internal-consistency reliability* (coefficient alpha) in quantitative research. In the qualitative case, the different analysts would be regarded as multiple items on a test, and the fallacy would be to look for a high "alpha." This fallacy is demonstrated in coders' content analysis. Inter-analyst "reliability" is a nonsensical concept in analytic qualitative research and does not provide the equivalent of the quantitative researcher's comforting coefficient alpha to indicate the internal-consistency reliability of measurement results (well, comforting to those uninitiated in C-OAR-SE theory!). Analytic qualitative research is highly unreliable in the inter-coder or inter-analyst sense but *can be* highly valid: it is just that *analysts*' predictive validities differ.

The second point is that internal-consistency reliability *is* important in qualitative research but in a reconceptualized way. The correct equivalent to internal-consistency reliability in qualitative research is the *leveling off of a marginal insight* referred to earlier—that is, the reliability of the *idea*. From this perspective, the "items" in the "test" are *A's successive feelings of confidence* about each inference he or she makes (data 2). The analyst must interview enough consumers or respondents—thereby sampling enough of the "confidence episodes" pertaining to the idea—until the analyst experiences a "cumulative confidence of inference" that the idea is approaching "100% correct," in which case it goes into the report (or 100% wrong, in which case the hypothesized idea is dropped from the report). This is equivalent to a coefficient alpha approaching 1.0. (I am merely drawing an analogy here; as we saw in Chapters 2 and 4, I don't endorse alpha for quantitative measures.) Of course, there will be multiple insights or ideas, and hence multiple "confidence alphas," but usually these will begin to simultaneously reach their *maxima* as more consumers are interviewed and the overall explanatory theory begins to be fitted together in the analyst's mind.

Analytical qualitative research, as stated before, is largely unreplicable. However, the analyst's performance over successive projects provides a practical indication of *test–retest reliability*—test–retest reliability in the sense of the analyst's cumulative "track record" over successive jobs. This is what distinguishes successful analysts from the unsuccessful.

In *quantitative* research, by contrast, there is no direct analogy to the individual analyst's test–retest reliability because the analyses (and the results) are supposed to be completely objective and replicable, that is, achievable by anyone. In truth, this is hardly ever the case because exact replications of *data 1*—the 90%-weighted C factor in $C + A = R$—are hardly ever achieved. In hindsight—see also the excellent book by the experienced researcher Robert (Bob) Abelson (1995)—I have probably underestimated the differences between quantitative statistical analysts by weighting the A factor at only 10%.

The R factor. Little has been written in the literature of qualitative research about the research *outcome*: the results, or R factor. Wells (1986) is one of the few to address this topic. He makes the important observation from his long experience that "words" reports, as are typical in qualitative research, are much more likely to be taken as a basis for managerial action than are "numbers" reports, as in quantitative research. The words and numbers division is deceptive, as Overholser (1986) and Scipione (1995), among others, have demonstrated (see Table 8.4). But the conceptualization of results involves much more than this distinction.

Table 8.4 Quantitative interpretation of qualitative reporting: base = 160 executive users of qualitative research (adapted from Scipione 1995)

Degree descriptors	Mean (%)	S.E.* (%)	Change descriptors	Mean (%)	S.E.* (%)
Virtually all	85	(1.1)	A significant change	47	(2.4)
Most	69	(1.7)	A substantial change	34	(1.8)
A large majority	61	(1.7)	Much more than	32	(1.7)
More than half	59	(0.8)	Somewhat more than	31	(2.0)
A majority	56	(1.4)	Somewhat less than	29	(2.0)
A large majority	41	(1.7)	Much less than	26	(1.0)
Less than half	40	(1.0)	A slight change	20	(2.2)
A minority	24	(1.7)			
Hardly anyone	12	(1.6)			

S.E. = one standard error (plus or minus) from the Mean, by my estimates from Scipione's data.

As mentioned several times throughout this chapter, the results of AQR should be presented in the form of *mini-theory of action*. This means that a *theoretical framework* for the results is required. Such theoretical frameworks are usually nonexistent in qualitative research reports. Not surprisingly, buyers of qualitative research studies complain that there is a lot of "free-form," unsatisfactory reporting (Moran 1986). Lack of a theoretical framework for the results is also a serious problem for academic "interpretive" qualitative research reports. One such framework, applicable especially to developing action recommendations for *advertising campaigns*, is available in Rossiter and Percy (1987, 1997). Without getting too specific here, the framework, to be "filled in" by the qualitative research analyst, consists of a behavioral sequence model, a listing of communication objectives, a positioning statement, and recommended persuasion tactics. For a theoretical framework for *media planning*, see Rossiter and Danaher (1998). Another framework, applicable to qualitative research for *new-product positioning*, could be easily adapted from the market research book by Urban and Hauser (1993). The qualitative researcher must *think*. The researcher must formulate an appropriate *R*-framework for presenting qualitative research results.

Finally, there is another difference between quantitative research and analytic qualitative research that resides in the *R* factor. Quantitative research can be fairly distinctly divided into theory-testing research and applied research (Fern and Monroe 1996). For example, *theory-testing research* is usually conducted with an experimental design in a laboratory setting, and statistical significance takes precedence over effect sizes. On the other hand, *applied research* is usually undertaken by using a nonexperimental survey or quasi-experimental field study (Campbell and Stanley 1973) where effect sizes are paramount regardless of statistical significance (Fern and Monroe 1996). But analytic qualitative research, AQR, *combines* theory-testing and applied research. The qualitative research analyst is developing and, in the interviews, tentatively testing, a *mini-theory*. The mini-theory requires the analyst to discover and define the relevant constructs, decide how they are causally related, infer consumers' or other respondents' scores on the measures of these constructs (thereby engaging in *measurement*), and then test the theory "on the run" before writing it up in the report. Statistical significance does not apply, but predictive effect sizes definitely do, though only in a loose, ordinal, small–medium–big quantitative metric rather than as precise numbers.

Along with its methodological and analytic difficulties, the lack of an "effects" test makes AQR unsuitable for most Ph.D. dissertations. With a qualitative dissertation, the examiners (or other readers) can say: "Well, that's your view—but who's to say you're right?" The Ph.D. student's theory may well be right (valid), but no one can know without a field test of it. Untested qualitative research cannot contribute to knowledge in the field, the usual defining requirement of Ph.D. dissertations.

In practice, qualitative research results (action recommendations) are usually tested for predictive validity in some form of *field experiment*. In marketing—and this includes political marketing and health promotion, two of the fast-growing subfields—this is most often an advertising pretest, a product or service test-market, or simply launching a product or running an advertising campaign and then "tracking" its results over time. Rarely is an ideal, fully controlled experimental design, such as the Solomon 4-group design (Campbell and Stanley 1973; and see Rossiter and Percy 1997, Chapter 19), affordable. Advertising pretests such as ARS^{TM} and ADVANTAGE*ACT^{TM} typically employ the one-group pretest–posttest design (observation–treatment–observation) in which the possible effect of pretesting on the posttest outcome is uncontrolled, although this effect appears not to be a problem in practice due to well-disguised pretest measures. Some advertising pretests, such as ADTEST^{TM} and RPM Test^{TM}, employ the true-experiment, posttest-only control-group design, which is well-controlled but requires more than double the sample size of the pretest–posttest design to obtain statistically reliable results.

In-market tracking studies employ the quasi-experimental design of either a single-group time series (panel sample) or equivalent-group time series (so-called "continuous" tracking, such as Millward Brown^{TM} or MarketMind^{TM} tracking research). Health promotion campaigns, for instance, are often monitored with continuous tracking (or should be). The main threat to quasi-experiments comes from potential rival causal factors operating in the marketplace but these are usually well-measured and assessed by these tracking research suppliers so that managers can be highly confident in the truth of the findings (Rossiter and Percy 1997, Chapter 20). However, this is not to condone what is too often a *non*experiment, the "one-shot case study" (Campbell and Stanley 1973) in which the manager merely observes sales (or other behavioral) results of the new marketing campaign and deems it a success or failure, with no measurement and control of potential alternative causal factors. If prior sales are observed as a pre-measure, this becomes a one-group pretest–posttest design, which is safe enough for advertising pretests, or "theater tests," but not for in-market tests without measurement and control of all likely other real-world causes of sales change.

In sum, validated advertising pretests (see Rossiter and Percy 1997, Chapter 19), responsible test-market experiments or quasi-experiments, or multiple-measurement market tracking is necessary to establish the *predictive validity* of a mini-theory proposed by analytic qualitative research.

8.7 Summary

This C-OAR-SE-theory perspective on qualitative research can be summarized in terms of four major points:

1. Qualitative research is not simply the use of qualitative interview methodologies. The *analyst* is a crucial and inseparable part of qualitative measurement, and the results are the analyst's causal mini-theory of the behavior that is the topic of investigation. This is properly called *analytic* qualitative research (AQR). The interpretive ability of the analyst contributes about 50% to the predictive validity of the results, which is enough range from analyst to analyst to produce qualitative research that is of very low to very high predictive validity. Interpretive ability requires interpersonal intelligence for collecting data and making inferences about causality, and knowledge of social sciences theory for making recommendations for an action plan.
2. Evaluation of qualitative research methodologies in terms of standard quantitative research criteria—such as random sampling of respondents, respondent-sample size, and any sort of statistical analysis of the results other than ordinal recommendations of "degree"—is completely inappropriate. Internal-consistency reliability in AQR refers to the analyst's "confidence alphas" in making inferences from successive sampling of respondents' data; enough interviews have to be conducted to yield high confidence in all the main inferences

constituting the mini-theory. Test–retest reliability in AQR refers to the analyst's record of predictive validity over jobs, the most important consideration in buying qualitative research commercially.

3. There is only one relevant validity criterion for qualitative research: the predictive validity of the results. A corollary of this is that the inferences formulated by the analyst cannot be regarded as contributions to knowledge until those inferences demonstrate predictive validity when tested in the field. This means that "interpretive" qualitative research, on its own, is not knowledge. In AQR conducted by professionals, predictive validity only has to be demonstrated once, for the current brand or object and the current promotional campaign. This is done *not* by trying to "quantify" the findings but via direct practical applications of the qualitative mini-theory. Qualitative research measurement cannot be "validated"—or, worse, *replaced*—by quantitative research measurement.
4. Analytic qualitative research, AQR, it could be argued, will always be a professional domain. This is because qualitative researchers must prove themselves professionally to be expert analysts and also because qualitative research mini-theories require a field test (a campaign) to establish their validity. Social-science academics and Ph.D. students, if they are particularly skilled as analysts, can conduct AQR to propose a theory (this despite the fact that as doctoral students they will receive no training in qualitative research—see the survey of U.S. doctoral programs by Aiken, West, and Millsap (2008), and also the comment by Zimiles (2009)). They can then put the theory up for peer review, which may result in a theoretical article clearly labeled as such. This is not to be belittled, as the social-science fields cannot develop without promising theories, though less so mini-theories. But very few academics or doctoral students have the necessary ability to do analytic qualitative research and, given usual resources, academics cannot test the theory by applying it in a real-world behavior-change campaign. Realistically, only professionals can do this. Professional qualitative researchers are therefore in a unique position to contribute scientific knowledge, not just temporary and specific knowledge for a particular campaign, but also more enduring and general knowledge in the case of those theories that are worth repeated trials in the marketplace.

8.8 End-of-Chapter Questions

(8.1) Write a conceptual definition of analytic qualitative research and then explain in 2,000 words or fewer why AQR (as I abbreviate it) is the most important method of measurement. (10 points)

(8.2) What skills are needed to become an expert AQR interviewer and analyst? How might you become one? If as a qualitative researcher you don't pass muster, what would you do to get valid results for the social-science research topics you are pursuing? (7 points)

(8.3) What do “validity” and “reliability” mean in relation to AQR as contrasted with their meanings in conventional psychometric theory? You can look back at Chapter 2 for this answer. (About 1,500 words; 7 points)

(8.4) Pick an article that relies on “interpretive” qualitative research in a major journal—such as the interdisciplinary *Journal of Consumer Research*—and criticize it from the standpoint of C-OAR-SE-based AQR theory. (About 1,500 words; 7 points)

(8.5) Advanced readers should attempt this question, which depends on *cumulative* understanding of C-OAR-SE theory. Read the article by France and Bone in *Marketing Letters*, 2009, 30(4), 385–397 (it would be hard to find a more confused treatment of construct definition and measurement than this). The article examines the U.S. Federal Drug Administration’s approach to, and the researchers’ suggested remedy for, “strength of science” warnings on package labels for dietary supplement products. Now that you understand C-OAR-SE, explain how you would properly define this construct, measure it, and communicate the warning system to consumers. This should be a journal article-length answer and I hope you get it published. (25 points)

Chapter 9
DROAVR Application Checklist

> *Drover, n: One who drives cattle or sheep to market.*
> —Webster's Unabridged Dictionary

In this final—"proactive" and rather "agro"—chapter I am going to summarize the most important things to do when *applying* this new version of C-OAR-SE theory. The application sequence forms the imaginative acronym DROAVR (pronounced "drover," which figuratively I see myself as) while propounding C-OAR-SE theory. The DROAVR acrostic denotes the checklist of tasks, in *order of importance*, in my experience, that the researcher must perform to correctly apply C-OAR-SE measurement theory:

1. Definition (within theory) of the construct
2. Rater entity identified
3. Object correctly represented
4. Attribute components formatively specified
5. Validity (content validity) rationally argued
6. Reliability (precision) of scores reported

I will now summarize these tasks in action language. In doing so, I will take this opportunity to attack some of the critics of my measurement theory and simultaneously to reiterate its main principles.

9.1 Definition (Within Theory) of the Construct

Before you can design (or select) a measure, you must carefully define the construct to be measured—within the theory that you are proposing to test. The only exception to this is when you are conducting—or, better still, commissioning a professional to conduct—*analytic qualitative research*. Unless you are an extremely talented introspector—which very few researchers are—you will need to commission an AQR study to formulate a theory in the first place. The constructs and their functional relationships will then emerge by thoughtful inference from the mind of

J.R. Rossiter, *Measurement for the Social Sciences*,
DOI 10.1007/978-1-4419-7158-6_9, © Springer Science+Business Media, LLC 2011

the expert qualitative researcher (as explained in Chapter 8). But *you personally* are obliged to *finally* define each of the constructs in the theory. You should not "sheepishly" follow anyone else's definition, because that definition, if written before this book came out, will undoubtedly fail on the three C-OAR-SE criteria for construct definitions. In C-OAR-SE theory, the construct is defined in terms of the *object* to be rated, the *attribute* it is to be rated on, by a particular *rater entity*.

For abstract (multiple-meaning) constructs—containing either an abstract object, an abstract attribute, or both—you will also have to write a definition that includes the *object's constituents or components* (see Chapter 3) and the *attribute's components* (see Chapter 4). In the case of an abstract *psychological* attribute (see Chapters 1 and 4), you as the researcher will have to do this task entirely yourself. In the case of an *abstract formed* object, you may need to enlist the aid of a sample of the rater-entity respondents to help identify its main components, but the final selection of components for the definition of the construct must be again made by you, the researcher, and the selections must be rationally justified by you.

After the initial *full* definition is spelled out in the introduction to the article you are preparing, you can thereafter use a *summary label* for the construct. If the construct is abstract, the *summary* label does not need to include the specific constituents and components. For instance: AMAZON'S COMPONENTIAL SERVICE QUALITY AS RATED BY FIRST-TIME CUSTOMERS is a summary label. Your readers may then realize that there is no such *construct* as the simplistically stated SERVICE QUALITY. This is the "attribute-only fallacy" mentioned numerous times in this book. This fallacy pervades all articles in all the social science journals, including the leading journals in management, marketing, information systems, organizational behavior, psychology, and sociology.

The definition of the construct may also differ according to its role in the theory you are testing. Ideally, the theory should be evident from your labeling of the construct (such as a label mentioning COMPONENTIAL service quality as opposed to *OVERALL* service quality). The most ubiquitous example of construct definitions differing by theory in the social sciences is ATTITUDE, which may be a set of *unipolar beliefs* in one theory, a *bipolar overall evaluation* in another theory, and either a *predictor variable* or a *criterion variable* in either theory.

9.2 Rater Entity Identified

I am tired of hearing the argument, such as from my cordial (and beer-loving) European colleague, Diamantis Diamantopoulos, to the effect that the rater entity is not part of the construct (see Diamantopoulos 2005, Diamantopoulos and Sigauw 2006). In Chapter 1, I went to great lengths to explain why the *rater entity* is an essential element of the construct. As I said (in Chapter 5), failure to identify the rater entity in the definition of the construct—up front—is the reason for many of the "mixed findings" that plague "literature reviews" at the beginning of articles (and in books).

If you clearly identify the *rater entity* for the construct or constructs that you are planning to measure, you will save social scientists an astronomical amount of confusion. Witness my attack on the study by the experienced psychometricians Edwards and Cable (2009) in Chapter 5. This is why I have given the rater-entity identification task highest priority among the O–A–R decisions in the DROAVR application by making the order here, in DROAVR, R–O–A.

9.3 Object Correctly Represented

It was the late William J. McGuire—one of the greatest theorists and methodologists of all time and a genuine iconoclast who, like me, found it almost impossible to get his most radical ideas published in the ultra-conservative top journals—who reinforced in my mind the importance of representing the *object* correctly in the measure (in his "object-on-attribute" theory of construct definition, in McGuire 1989). But it was my ongoing experience as an *applied* marketing researcher—studying real consumers in the real world—that really drove this point home. (This industry experience also led to my insistence on *efficient* measures.) One of the few to understand the importance of object focus, I was the first in the advertising theory literature to divide the construct of BRAND AWARENESS into BRAND RECOGNITION (in which the object in the measure is usually represented visually) and BRAND RECALL (in which the object in the measure is usually represented verbally) and also to carefully measure BRAND ATTITUDE OR PREFERENCE based on a visual stimulus or a verbal stimulus in the questionnaire depending on what stimulus usually *elicits* the ATTITUDE OR PREFERENCE for the BUYER (see my advertising textbooks, Rossiter and Percy 1987, 1997, Rossiter and Bellman 2005). This fundamental distinction is ignored in almost every journal article written by off-the-planet advertising academics—most of whom have never had to design an actual advertising campaign, working with real clients.

Academic researchers continually make the "UFO error" of failing to identify the object in the construct, or else they misrepresent the object. Academics who have seriously misrepresented the object include previous *Journal of Consumer Psychology* editor the esteemed Bob Wyer, one of the researchers to whom I am greatly indebted for the insightful idea that polytomous options in answer scales may be responded to probabilistically (see Chapter 6, and Wyer 1974). I hope you realize that I am criticizing what researchers do, not the researchers themselves.

If you have identified the rater entity and correctly represented the object, you are well on the way to coming up with a content-valid measure of the construct. Everyone else wrongly focuses on the attribute alone, the A, ignoring the R and the O.

9.4 Attribute Components Formatively Specified

Measurement is of course also about *attributes*, and specification of the attribute and, if abstract, its *components*, is the most complex aspect of C-OAR-SE theory. Accordingly, this aspect of the construct gets the second-longest chapter in the book (Chapter 4), exceeded only by the item type-selection chapter, which covers many *different* constructs. Chapter 4 will be the most controversial—no doubt—since it is in this aspect of construct measurement that I differ most profoundly with the psychometricians, given that the attribute is the only element that they include!

The main applied outcome of my theory of the attribute element of constructs is to reject the psychometric sheep's main fodder—Spearman's *factor analysis* and Cronbach's *coefficient alpha*. Cronbach himself abandoned coefficient alpha in favor of "generalizability theory," which is just another illogically attempted *statistical* solution to the validity problem. Former colleague Ujwal Kayandé and his mentor Adam Finn (2005), two who rejected C-OAR-SE theory (the former was originally intending to be a co-author with me for the 2002 article—but lost faith), have followed Cronbach's new misdirection. Among other things, G-theory, like alpha-theory, continues to wrongly propose that multiple items are always needed to measure a construct, so that the composite score will be "generalizable," whatever that may mean.

A *concrete* attribute does *not* need a multiple-item measure. It just needs one good (content-valid) *single item*. An *abstract* attribute does require multiple items in the measure. But the items correspond to the predefined first-order components and *themselves* are *concrete* (actually "doubly concrete" because the object must also be concrete) and each should use only one good single item.

Most important is my realization—a change from the 2002 version of C-OAR-SE—that all *abstract* attributes are "formed," not "reflective," from a measurement standpoint. This realization rules out the otherwise thoughtful attempt by Dutch researcher Denny Borsboom—to whom I sent my C-OAR-SE article twice with arrogantly no reply—to address the validity problem (see Borsboom 2005). I don't expect him to do so, but he should read my recommendation on how to fix his theory to incorporate formed attributes (Rossiter 2005, p. 24, an article in which I now apologize profusely for mis-spelling his name). Borsboom's theory applies only to "reflective" attributes (as does Bagozzi's theory, e.g., Bagozzi 2007) and I now argue that "reflective" measures of them mis-measure the real attribute.

All so-called "latent attributes" are not real, and it is misleading for researchers to chase these artificial statistical phantoms. I write this on the day that the infamous female "G-spot" has been proven to be not real, a comparison that, though crude, might make this point memorable. There's not even a "latent" G-spot, so give up searching.

9.5 Validity (Content Validity) Rationally Argued

The DROAVR strives to drive out the multitrait-multimethod (MTMM) concept of validity, whose initials should stand for the "more the merrier mistake." MTMM is completely illogical. You cannot establish the validity of a measure by the *scores* it produces (I explained this in Chapter 2). The only *essential* type of validity is *content validity*—which consists of *item-content validity* and *answer-scale validity*. Content validity depends entirely on the semantic correspondence between the construct definition and the measure and not at all on the relationship between the measure and its score or scores (see my C → M → S structure-of-measurement model in Chapter 2).

My “Horse is a horse” retort (Rossiter 2005) to the few critics of C-OAR-SE who bothered to speak out also made this point. As I said in that article, an MTMM theorist trying to measure a Yak will end up with just the four legs as the measure. This measure will show convincing “convergent validity” with other four-legged measures—which may be the legs of a Horse, or a Camel (see Fig. 9.1). As I said—and was the first to say so as far as I know—“convergent validity” ignores the content-validity of the measure (indeed of both measures, the new and the old) and thus *cannot* prove that the measure is a good one.

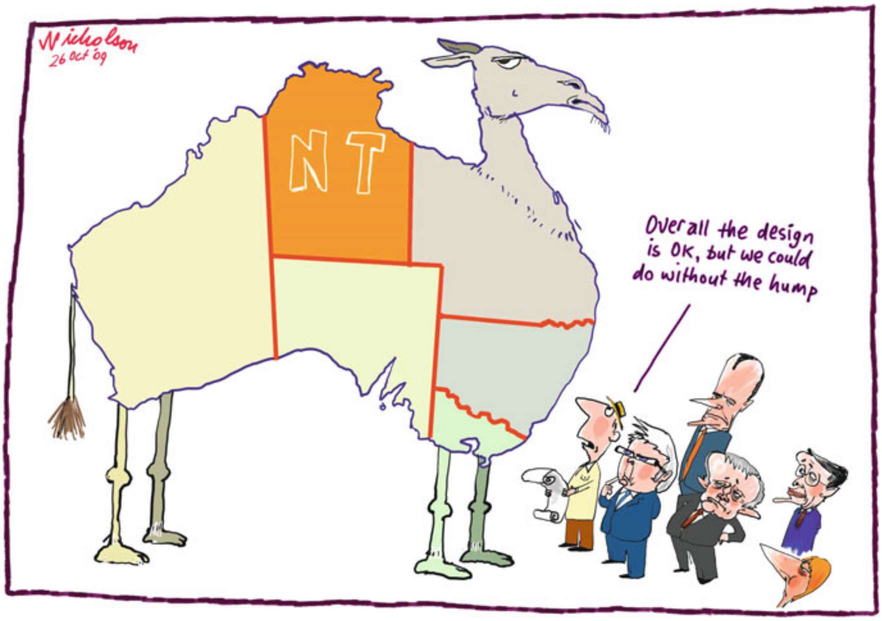

Fig. 9.1 This is a Yak, according to psychometricians. They would pick up the four legs as the only internally consistent (high alpha) multiple items forming a “factor.” They would not see it as a Camel. (Cartoon by Nicholson from *The Australian*, permission granted, and see www.nicholsoncartoons.com.au)

Content validity is all that matters and content validity is solely a matter of *rational argument*. Content validity requires a rational, nonempirical, nonstatistical argument to the effect that the item content in the measure has high semantic correspondence with the content of the construct definition, and that the categories in the answer scale for the item allow the great majority of raters to express what’s in their minds when answering the item, no more and no less.

For causal predictor variables in the theory, good *predictive validity* of the measure’s scores in predicting the scores on the criterion variable is additionally desirable, *given* prior high content-validity of all measures involved. “Good” predictive validity means coming close (say 95% confidence interval) to the *population validity coefficient*, R_{pop}; it does *not* mean *maximizing* the prediction.

9.6 Reliability (Precision) of Scores Reported

The term “reliability” means two things in C-OAR-SE theory: *stability*, which is short-interval test–retest reliability ($R_{\text{stability}}$) and *precision-of-score reliability* ($R_{\text{precision}}$).

Stability reliability ($R_{\text{stability}}$) has to be reported if you have developed an entirely new item and answer scale to go into your measure. If you have—and congratulations!—then you will need to conduct a short-interval (1-week or at most 2-week) within-person, test–retest study with a fairly large and representative sample of raters to establish the measure’s score stability. If the measure involves a unipolar attribute, you should use only the double-positive repeat rate (DPRR) statistic of stability (see Dolnicar and Rossiter 2008). If the measure involves a bipolar attribute, you should report DPRR, DNRR, and overall CONRR (consistent positive *and* negative repeat-rate stability—although the DPRR is usually the more important; see Rossiter et al. 2010). Do not use Pearson correlations because this statistic is wrongly confounded by noisy intensity judgments and, if the measure has a polytomous answer scale, it will be artificially inflated by *extreme* responding and by *evasion* responses if there is a midpoint offered.

If you use the new and superior DLF IIST Binary measure of beliefs, the most common construct in all social sciences, then you can take its $R_{\text{stability}}$ report from the hopefully forthcoming article by me, Dolnicar, and Grün—the two very important “et al.” above, and I can’t help feeling I’m channeling Paul Simon’s *Graceland* song here, “You Can Call Me Al”—or you can e-mail me for the 2010 working paper.

Precision-of-score reliability ($R_{\text{precision}}$) must be reported for every *use* of the measure. $R_{\text{precision}}$ is not something “inherent” in the measure but is a situational empirical *statistic*. As such, it says nothing at all about the measure’s *validity*, but rather *presumes* it to be high. The look-up tables in Appendix B, Tables B.1 and B.2, are sufficiently accurate for the $R_{\text{precision}}$ report. The tables are based largely on sample size and so they don’t require any tedious computations specific to your own data—nor three or more decimal places to falsely make them look impressive! To supplement or complement my account, you would do well to read the article by Sawyer and Peter (1983) defending small sample sizes. The *content validity* of the measures is far more important than sample size and this is what has been fatally overlooked in the social sciences.

Go DROAVR! With acknowledgement to Nike—the sports brand, not the Greek winged goddess of victory, whom I thought of often in my battle against psychometrics while writing this book—“Just do it!”

Appendix A

Comparison of Nunnally/Churchill Approach with C-OAR-SE Approach to Measurement

Measurement Theory and Procedural Steps	**Nunnally/Churchill**	**C-OAR-SE**
True-score theory	Based on old true-score model: *Observed score = True score + Random error*.	Based on new true-score model: *Observed score = True score + Measure-induced distortion + Rater error*.
Scope	Applicable *only* to "abstract" (multiple-item) constructs.	Applies to *all* constructs, "concrete" (single-item) and "abstract" (multiple-item).
Validity	*Content validity:* Acknowledged as essential, but inadequately defined and handled in the Nunnally/Churchill measure-development procedure.	*Content validity:* Essential, and consists of (a) item-content validity – semantic identity of the construct and the measure; and (b) answer-scale validity – freedom from measure-induced distortions. Established rationally by expert judgment.
	Construct validity: Seen as essential, though should be called *measure validity*. Measure validity is wrongly tested empirically by examining convergent correlations and discriminant correlations with other measures.	*Construct validity:* Meaningless, because you cannot validate – that is, prove the truth of – a *construct*. You can only validate a *measure* of a construct, and then only by a rational argument as to its high content validity, not by any empirical means.

J.R. Rossiter, *Measurement for the Social Sciences*,
DOI 10.1007/978-1-4419-7158-6, © Springer Science+Business Media, LLC 2011

	Predictive validity: Essential, but not adequately explained.	*Predictive validity:* Desirable but not essential. Predictive validity applies only to predictor constructs. Criterion constructs depend completely on high *content* validity. Predictive validity requires an estimate of the *population correlation* between predictor and criterion scores.
Reliability	Defined as absence of *random (i.e., rater) error* in observed scores, following the "old" true-score model. But operationalized only as *internal-consistency* reliability (coefficient alpha), which assumes a *multiple-item* measure. Both Nunnally and Churchill mention *test–retest* reliability (stability) but advise against using it.	*Stability reliability:* Essential, observed score(s) must be highly repeatable on a short-interval retest. *Precision reliability:* Accuracy of *observed* score(s), which depends mainly on *sample size* and *presumes* a highly content-valid measure. Precision reliability should be reported for observed scores on all the main measures in the study.
1. Define the construct	Both Nunnally and Churchill define the construct in terms of the *attribute only*. This mistake is made by almost all researchers.	C-OAR-SE construct definition requires specification of (1) the *object* to be rated, (2) the *attribute* it is to be rated on, and (3) the *rater entity*, who does the rating. Constructs are ultimately *researcher-defined*, with no empirical assistance other than pooled experts' judgments when the researcher is unsure.
2. Generate items	Candidate items are either borrowed from others' measures (of questionable content validity and unknown stability) or are generated from qualitative open-ended interviews, with the item content mainly decided by the *raters*.	Items must be decided on ultimately by the *researcher*. Raters' inputs are necessary only if the construct is *perceptual*. Raters' inputs are not used if the construct is *psychological*, i.e., not self-reportable.
3. Purify the measure	Items are deleted from the candidate pool if they don't correlate with other items and with a "latent" statistical factor and don't contribute to a high coefficient alpha.	Items are *never* deleted from the defined set of items. The items are based on *a priori* argued item-content validity, not derived from correlated scores *ex post*.

4. Assess reliability	Only *internal-consistency* reliability (coefficient α) is calculated. Coefficient α is legitimate (though unnecessary) for a multiple-item measure but meaningless for a single-item measure. Nunnally's (1978) minimum α of .8 for a final measure is very often ignored and the measure is used anyway.	*Stability reliability* is assessed by a short-interval test–retest. High stability (a "double-positive" repeat rate of 80% is the acceptable minimum) is required for the measure. *Precision reliability* can be estimated from the sample size of raters in a particular study by using "lookup" tables.
5. Assess construct validity	Construct validity is assessed by the multitrait-multimethod correlational procedure, which does not relate to the construct itself. In any case, construct validation can only mean *measure* validation.	*Constructs* are definitions, not empirically testable propositions. Only a measure can be validated (with regard to the defined construct). This is *content validity* (high item-content validity and high answer-scale validity) and high content validity is essential.
	Churchill also recommends empirically testing the measure for *known-groups discriminant validity*, which he doesn't realize is just another form of *predictive validity*.	*Predictive validity* (of the measure of a predictor construct) is *desirable* only, not essential. Predictive validity requires prior high content validity of the measure and a *population correlation* estimate against which to assess the observed predictive validity correlation.
6. Develop norms	Norms are misleadingly recommended as a solution to the problem of assessing whether you're getting true scores from different answer scales. Norms require a very large and representative rater sample – rarely attained in academic studies, which usually employ college students, a nonrepresentative rater entity.	Norms are needed in the form of *population correlations* to properly assess *predictive validity*. Norms based on measures with *low content validity*, and comparisons based on a *different measure* than the one in the norms, are useless.

Appendix B

$R_{\text{precision}}$ Tables

Table B.1 Single percentage

	Plus or minus error when the percentage is close to				
Sample size	10% or 90%	20% or 80%	30% or 70%	40% or 60%	50%
1,000	2	3	3	3	3
500	3	4	4	4	4
250	4	5	6	6	6
200	4	6	6	7	7
150	5	6	7	8	8
100	6	8	9	10	10
50	8	11	13	14	14
25	12	16	18	19	20

Example: A reported percentage of 30%, based on a random sample of 200 consumers, has an error rate of plus or minus 6%. That is, we could be "95% confident" that the actual population percentage, had everyone been surveyed, is between 24 and 36%.

Table B.2 Difference between percentages

	Difference needed when the average of the two percentages is close to				
Average of the two sample sizes	10% or 90%	20% or 80%	30% or 70%	40% or 60%	50%
1,000	4	4	5	5	5
500	4	5	6	6	6
250	5	7	8	9	9
200	6	8	9	10	10
150	7	9	10	11	11
100	8	11	13	14	14
50	12	16	18	19	20
25	17	22	25	27	28

Example: Suppose a TV commercial day-after-recall test, based on a random sample of 200 viewers, indicates a recall score of 20%. You are disappointed. You decide to repeat the test with a new random sample of 100 viewers, and the commercial now obtains a recall score of 30%. Are these reliably different scores? The average of the two sample sizes is 150. The average of the two recall scores is 25%. The conservative difference needed is 10% (from the table at the intersection of the 150 row and the 30% column). Yes, you can be "95% confident" that the second recall score is reliably higher than the first.

Source: Compiled from more detailed tables in the Newspaper Advertising Bureau publication, The Audience for Newspaper Advertising, New York: NAB, 1978, Appendix.

Appendix C

A Binominal Effect-Size Conversion of *r*

Value of the common language effect size indicator (CL_R) for corresponding values of Pearson's *r*

	CL_R									
r	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.00	.500	.503	.506	.510	.513	.516	.519	.522	.525	.529
.10	.532	.535	.538	.541	.545	.548	.551	.554	.558	.561
.20	.564	.567	.571	.574	.577	.580	.584	.587	.590	.594
.30	.597	.600	.604	.607	.610	.614	.617	.621	.624	.628
.40	.631	.634	.638	.641	.645	.649	.652	.656	.659	.663
.50	.667	.670	.674	.678	.682	.685	.689	.693	.697	.701
.60	.705	.709	.713	.717	.721	.725	.729	.734	.738	.742
.70	.747	.751	.756	.760	.765	.770	.775	.780	.785	.790
.80	.795	.801	.806	.812	.817	.823	.830	.836	.842	.849
.90	.856	.864	.872	.880	.889	.899	.910	.922	.936	.955

Example: To read the table entries as *binary odds*, select the Pearson *r* correlation to two decimal places (e.g., $r = .55$) and then locate the corresponding CL_R figure in the table (e.g., $CL_R = .685$), then move the CL_R's decimal place two figures to the right to get a percentage figure (e.g., 68.5%). That is, a correlation of .55 translates to odds of 69% successes and 31% failures, as do $r = .56$ and $r = .57$.
Source: Dunlap (1994).

References

Abelson RP (1995) Statistics as principled argument. Lawrence Erlbaum Associates, Hillsdale

Aiken LS, West SG, Millsap RE (2008) Doctoral training in statistics, measurement, and methodology in psychology. Am Psychol 63(1):32–50

Ajzen I (1988) Attitudes, personality, and behavior. Open University Press, Buckingham

Allport GW (1935) Attitudes. In: Murchison CM (ed) Handbook of social psychology. Clark University Press, Worcester

Allport GW (1985) The historical background of social psychology. In: Gardner L, Elliot A (eds) The handbook of social psychology, 3rd edn. Lawrence Erlbaum Associates, Hillsdale

Althuizen N, Wierenga B, Rossiter JR (2010) The validity of two brief measures of creative ability. Creativ Res J 22(1):53–61

Anastasi A (1981) Coaching, test sophistication, and developed abilities. Am Psychol 36(10):1086–1093

Arce-Ferrer AJ, Guzmán EM (2009) Studying the equivalence of computer-delivered and paper-based administrations of the Raven Standard Progressive Matrices test. Educ Psychol Meas 69(5):855–867

Armstrong JS (1967) Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. Am Stat 21(5):17–21

Armstrong JS (1998) Are student ratings of instructors useful? Am Psychol 53(11):1223–1224

Armstrong JS, Soelberg P (1968) On the interpretation of factor analysis. Psychol Bull 70(5): 361–364

Back MD, Schmukle SC, Egloff B (2009) Predicting actual behavior from the explicit and implicit self-concept of personality. J Pers Soc Psychol 97(3):533–548

Back MD, Stopfer JM, Vazire S, Gaddis S, Schmukle SC, Egloff B, Gosling SD (2010) Facebook profiles reflect actual personality, not self-idealization. Psychol Sci 21(3):372–374

Bagozzi RP (1994) Structural equation models in marketing research: basic principles. In: Bagozzi RP (ed) Principles of marketing research. Blackwell, Cambridge, pp 317–385

Bagozzi RP (2007) On the meaning of formative measurement and how it differs from reflective measurement: comment on Howell, Breivik, and Wilcox. Psychol Meth 12(2):229–237

Baldinger AL (1992) What CEOs are saying about brand equity: a call to action for researchers. J Advert Res 32(4):RC6–RC12

Bardi A, Lee JA, Hoffmann-Towfigh N, Soutar G (2009) The structure of intraindividual value change. J Pers Soc Psychol 97(5):913–939

Barrett LF (2004) Feelings or words? Understanding the content in self-report ratings of emotional experience. J Pers Soc Psychol 87(2):266–281

Bartoshuk LM, Duffy VB, Green BG, Hoffman HJ, Ko C-W, Lucchina LA, Marks LE, Snyder DJ, Weiffenbach JM (2004) Valid across-group comparisons with labeled scales: the gLMS versus magnitude matching. Physiol Behav 82(1):109–114

Bass FM (1969) A new product growth model for consumer durables. Manag Sci 15(1): 215–217

Bearden WO, Netemeyer RG (1999) Handbook of marketing scales: multi-item measures for marketing and consumer behavior research. Sage, Thousand Oaks

Belk RW (ed) (1991) Highways and buyways: naturalistic research from the consumer behavior Odyssey. Association for Consumer Research, Provo

Bell ST (2007) Deep-level composition variables as predictors of team performance: a meta-analysis. J Appl Psychol 92(3):595–615

Bellman S (2007) Theory and measurement of type 1 and type 2 emotions. Australas Mark J 15(1):14–22

Bergkvist L, Rossiter JR (2007) The predictive validity of multiple-item vs. single-item measures of the same constructs. J Mark Res 44(2):175–184

Bergkvist L, Rossiter JR (2009) Tailor-made single-item measures of doubly concrete constructs. Int J Advert 28(4):607–621

Bialik C (2010). "The power of lucky charms." The Wall Street Journal. Accessed online Apr 28

Blalock HM (1964) Causal inferences in nonexperimental research. University of North Carolina Press, Chapel Hill

Blau P, Duncan OD (1967) The American occupational structure. Wiley, New York

Bollen K, Lennox R (1991) Conventional wisdom on measurement: a structural equation perspective. Psychol Bull 110(2):305–314

Borgenau P, Ostendorf F (1998) The Big Five as states: how useful is the five-factor model to describe intra-individual variations over time? J Res Pers, 32(2):202–221

Borsboom D (2005) Measuring the mind. Cambridge University Press, Cambridge, England

Boswell WR, Shipp AJ, Payne SC, Culbertson SS (2009) Changes in newcomer job satisfaction over time: examining the pattern of honeymoons and hangovers. J Appl Psychol 94(4): 844–858

Brakus JJ, Schmitt BH, Zarantonello L (2009) Brand experience: what is it? How is it measured? Does it affect loyalty? J Mark 73(2):52–68

Brehm JW (1966) A theory of psychological reactance. Academic Press, New York

Breivik E, Thorbjørnsen H (2008) Consumer brand relationships: an investigation of two alternative models. J Acad Mark Sci 36(4):443–472

Brewer MB, Chen Y-R (2007) Where (who) are collectives in collectivism? Toward a conceptual clarification of individualism and collectivism. Psychol Bull 114(1):133–151

Brown RP, Day EA (2006) The difference isn't black and white: stereotype threat and the race gap on Raven's Advanced Progressive Matrices. J Appl Psychol 91(4):979–985

Burger JM (2009) Replicating Milgram: would people still obey today? Am Psychol 64(1):1–11

Burkley E, Burkley M (2009) "Mythbusters": a tool for teaching research methods in psychology. Teach Psychol 36(3):179–184

Buss DM (2009) How can evolutionary psychology successfully explain personality and individual differences? Perspect Psychol Sci 4(4):359–366

Cacioppo JT, Petty RE (1982) The need for cognition. J Pers Soc Psychol 42(1):116–131

Calder BJ (1977) Focus groups and the nature of qualitative research. J Mark Res 14(3):353–364

Calder BJ (1994) Qualitative marketing research. In: Bagozzi RP (ed) Marketing research. Blackwell, Cambridge, pp 50–72

Callick R (2009) "Australia second on UN Human Index." The Australian, Oct 6, p 2

Calver LA, Stokes BJ, Isbister GK (2009) The dark side of the moon. Med J Aust 191:692–694 (11 Dec)

Campbell DT, Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 56(2):81–105

Campbell DT, Stanley JC (1973) Experimental and quasi-experimental designs for research. Rand McNally, Chicago

Cattell RB, Eber HW, Tastuoka MM (1970) Handbook for the sixteen personality factor questionnaire (16PF). Institute for Personality and Ability Testing, Champaign

Ceci SJ, Williams WM (1997) Schooling intelligence, and income. Am Psychol 52(10): 1051–1058

Chan AM, Rossiter JR (1998) Construction of a Chinese values scale and a Chineseness ethnicity scale. Asia Pac Adv Consum Res 3:61–67

Chapman S, and seven others (2009) The content and structure of Australian television reportage on health and medicine, 2005–2009: parameters to guide health workers. Med J Aust Dec 11: 620–624

Churchill GA (1979) A paradigm for development better measures of marketing constructs. J Mark Res 16(1):64–73

Clark JK, Wegener DT (2009) Source entitativity and the elaboration of persuasive messages: the roles of perceived efficacy and message discrepancy. J Pers Soc Psychol 97(1):42–57

Cohen J (1977) Statistical power analysis for the behavioral sciences, Revised edn Academic Press, New York

Cohen D (2010) The escape of Sigmund Freud. JR Books, London

Coleman RP (1983) The continuing significance of social class to marketing. J Consum Res 10(3):265–280

Collins Editorial Staff (2002) Paperback dictionary and thesaurus. Harper Collins, Glasgow

Collins AM, Quillian MR (1972) Retrieval time from semantic memory. J Verbal Learn Verbal Behav 8(2):240–247

Coltman T, Devinney T, Midgley D, Venaik S (2008) Formative versus reflective measurement models: Two application of formative measurement. J Bus Res 61(12):1250–1262

Conway R (1978) Land of the long weekend. Sun Books, South Melbourne

Coombs CH (1964) A theory of data. Wiley, New York

Cresswell A (2010) "Demand for obesity surgery soars, but lifestyle training the best value". The Australian, Mar 29, p 4

Cronbach LJ (1946) Response sets and test validity. Educ Psychol Meas 6(4):475–494

Cronbach LJ (1950) Further evidence on response sets and test design. Educ Psychol Meas 10(1):3–31

Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3): 297–334

Cuneo AZ (1999) "New ads draw on hypnosis for branding positioning." Advertising Age, Jul 19, p 9

Dart J (2009) "Anti-smoking ads rekindle desire." Sydney Morning Herald, Nov 17, p 3

De Houwer J, Teige-Mocigemba S, Spruyt A, Moors A (2009) Implicit measures: a normative analysis and review. Psychol Bull 135(3):347–368

De Rue DS, Morgeson FP (2009) Stability and change in person-team and person-role fit over time: the effects of growth satisfaction, performance, and general self-efficacy. J Pers Soc Psychol 92(5):1242–1253

Dean GA, Nias DKB, French CC (1997) Graphology, astrology, and parapsychology In: Nyborg H (ed) The scientific study of human nature. Elsevier, Oxford

Deligonul S, Kim D, Roath R, Cavusgil E (2006) The Achilles' heel of an enduring relationship: appropriation of rents between a manufacturer and its foreign distributor. J Bus Res 59(7): 802–810

Denollet J (2005) DS14: Standard assessment of negative affectivity, social inhibition, and Type D personality. Psychosom Med 67(1):89–97

Deshpandé R, Farley JU, Webster FE (1993) Corporate culture, customer orientation, and innovativeness in Japanese firms: a quadrad analysis. J Mark 57(1):23–37

Deshpandé R, Webster FE (1989) Organizational culture and marketing: defining the research agenda. J Mark 53(1):3–15

Diamantopoulos A (2005) The C-OAR-SE procedure for scale development in marketing: a comment. Int J Res Mark 22(1):1–9

Diamantopoulos A, Sigauw JA (2006) Formative versus reflective indicators in organizational measure development: a comparison and empirical illustration. Br J Manag 17(4):263–282

Dichter ER (1964) Handbook of consumer motivations: the psychology of the world of objects. McGraw-Hill, New York

Dickens WT, Kane TJ, Schultze CL (1995) Ring true? A closer look at a grim portrait of American society. Brookings Rev 13(1):18–23

Dolnicar S, Grün B (2007) Cross-cultural differences in survey response patterns. Int Mark Rev 24(1):127–143

Dolnicar S, Rossiter JR (2008) The low stability of brand-attribute associations is partly due to market research methodology. Int J Res Mark 25(2):104–108

Dowling GR, Kabanoff B (1996) Computer-aided content analysis: what do 240 advertising slogans have in common? Mark Lett 7(1):63–75

Dudley NM, Orvis KA, Lebietki JE, Cortina JM (2006) A meta-analytic investigation of conscientiousness in the prediction of job performance: examining the intercorrelations and the incremental validity of narrow traits. J Appl Psychol 91(1):40–57

Dunlap W (1994) Generalizing the common language effect size indicator to bivariate normal correlations. Psychol Bull 116(3):509–511

Durgee JF (1985) Depth-interview techniques for creative advertising. J Advert Res 25(6):29–37

Edwards JR, Cable DM (2009) The value of value congruence. J Appl Psychol 94(3):654–677

Embretson SE (ed) (2010) Measuring psychological constructs: advances in model-based approaches. American Psychological Association, Washington

Emons WHM, Sijtsma K, Meijer RR (2007) On the consistency of individual classification using short scales. Psychol Meth 12(1):105–120

English HB, English AC (1958) A comprehensive dictionary of psychological and psychoanalytical terms. Longmans, London

Erlandson DA, Harris EL, Skipper BL, Allen SA (1993) Doing naturalistic inquiry: a guide to methods. Sage, Newbury Park

Etgar M, Fuchs G (2009) Why and how service quality perceptions impact consumer responses. Manag Serv Qual 19(4):474–485

Evanschitsky H, Iyer GR, Plassmann H, Niessing J, Meffert H (2006) The relative strength of affective commitment in securing loyalty in service relationships. J Bus Res 59:1207–1213

Eysenck HJ (1958) A short questionnaire for the measurement of two dimensions of personality. J Appl Psychol 42(1):14–17

Eysenck HJ (1978) An exercise in mega-silliness. Am Psychol 33(5):517

Eysenck HJ (1979) The structure and measurement of intelligence. Springer, Berlin

Eysenck HJ (ed) (1981) A model for personality. Springer, Berlin

Eysenck HJ (1990) Check your own I.Q. Penguin Books, London

Eysenck HJ (1994) Meta-analysis and its problems. Br Med J 309:789–792 (24 Sept)

Eysenck HJ (1997) Rebel with a cause. Transaction Publishers, New Jersey

Eysenck HJ, Wilson G (1976) Know your own personality. Pelican Books, Harmondsworth

Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ (1999) Evaluating the use of exploratory factor analysis in psychological research. Psychol Meth 4(3):272–299

Ferguson CJ, Kilburn J (2010) Much ado about nothing: the misestimation and overinterpretation of violent video game effects in Eastern and Western nations. Psychol Bull 136(2):174–178

Fern E, Monroe K (1996) Effect-size estimates: issues and problems in interpretation. J Consum Res 23(2):89–105

Ferrando PJ, Anguiano-Carrasco C (2009) Assessing the impact of faking on binary personality measures: an IRT-based multiple-group factor analytic procedure. Multivariate Behav Res 44(4):497–524

Finn A, Kayandé U (2005) How fine is C-OAR-SE? A generalizability theory perspective on Rossiter's procedure. Int J Res Mark 22(1):11–21

Fishbein M (1963) An investigation of the relationships between beliefs about the object and attitude toward that object. Human Relat 16(3):233–240

Fishbein M, Ajzen I (1975) Belief, attitude, intention, and behavior: an introduction to theory and research. Addison-Wesley, Reading

Fishbein M, Ajzen I (2010) Predicting and changing behavior: the reasoned action approach. Psychology Press, New York

Fishbein M, Hall-Jamieson K, Zimmer E, von Haeften I, Nabi R (2002) Avoiding the boomerang: testing the relative effectiveness of antidrug public service announcements before a national campaign. Am J Public Health 92(2):238–245

Fitzsimons GJ, Lehmann DR (2004) Reactance to recommendations: when unsolicited advice yields contrary responses. Mark Sci 23(1):82–94

Flynn JR (1987) Massive IQ gains in 14 nations: what IQ tests really measure. Psychol Bull 101(2):171–191

Follman J (1984) Cornucopia of correlations. Am Psychol 40(6):701–702

Fornell C, Johnson MD, Anderson EW, Cha J, Bryant BE (1996) The American Customer Satisfaction Index: nature, purpose and findings. J Mark 60(4):7–18

Fournier S (1998) Consumer and their brands: developing relationship theory in consumer research. J Consum Res 24(4):343–373

Freud S (1911). The interpretation of dreams (2009 reprint by IAP, Scotts Valley, CA.)

Freud S (1949). An outline of psycho-analysis. Norton, New York (1989 edition translated by J. Strachey with a foreword by P. Gay)

Fussell P (1983) Class: a guide through the American Status System. Summit, New York

"G-spot's all in the mind" (2010). The Australian (item from the U.K. Sunday Times), Jan 4, p 3

Gardner H (1983) Frames of mind: the theory of multiple intelligences. Basic Books, New York

Gardner DG, Cummings LL, Dunham RB, Pierce JL (1998) Single-item versus multiple-item measurement scales: an empirical comparison. Educ Psychol Meas 58(6):898–915

Gerzema J, Lebar E (2008) The brand bubble: the looming crisis in brand value and how to avoid it. Jossey-Bass, San Francisco

Glaser BG, Strauss AL (1967) The discovery of grounded theory: strategies for qualitative research. Aldine, Chicago

Goff K, Torrance EP (2002) The Abbreviated Torrance Test for Adults (ATTA). Scholastic Testing Service, Bensenville

Goldberg LR (1992) The development of markers for the Big-Five factor structure. Psychol Assess 4(1):26–42

"Good oil on Cadbury." (2010) Sunday Telegraph, Jan 3, p 17

Gordon W (1997). Is the right research being ill-used? Admap, 20–23 Feb.

Gould SJ (1991) The self-manipulation of my pervasive, vital energy through product use: an introspective-praxis approach. J Consum Res 18(2):194–207

Green BG, Shaffer GS, Gilmore MM (1993) A semantically-labeled magnitude scale of oral sensation with apparent ratio properties. Chem Senses 18(6):683–702

Greenwald AG, Brock T, Ostrom T (1968) Psychological foundations of attitudes. Academic Press, New York

Greenwald AG, McGhee DE, Schwartz JLK (1998) Measuring individual differences in implicit cognition: the Implicit Association Test. J Pers Soc Psychol 74(6):1464–1480

Griffin AJ, Hauser JR (1993) The voice of the customer. Mark Sci 12(1):1–27

Griffiths P (2006) New penguin dictionary of music. Penguin, London

Guilford JP (1936) Psychometric methods. McGraw-Hill, New York

Guilford JP (1950) Psychometric methods, 2nd edn. McGraw-Hill, New York

Harrell TW, Harrell MS (1945) Group classification of test scores for civilian occupations. Educ Psychol Meas 5(3):231–239

Harringman U (1990) Complete crossword dictionary. GAP Publishing Pty. Ltd, Norman park

Harrington DM, Block J, Block JH (1983) Predicting creativity in adolescence from divergent thinking in early childhood. J Pers Soc Psychol 45(3):609–623

Harrower M (1976) Rorschach records of the Nazi war criminals: an experimental study after thirty years. J Pers Assess 40(4):341–351

Hatfield E, Rapson RL (2000) Love and attachment processes. In: Lewis M, Haviland-Jones JM (eds) Handbook of emotions, 2nd edn. Guilford, New York, pp 654–662

Hedges LV (1987) How hard is hard science, how soft is soft science? The empirical cumulativeness of research. Am Psychol 42(2):443–455

Heise DR (1969) Some methodological issues in semantic differential research. Psychol Bull 72(6):406–422

Hesse H (1927). Der Steppenwolf (2001 Penguin Classics reprint, translated by Basil Creighton and revised by Walter Sorrell).

Hihouse S, Broadfoot A, Devendorf SA, Yugo JE (2009) Examining corporate reputation judgments with generalizability theory. J Appl Psychol 94(3):782–789

Hilgard ER (1956) Theories of learning, 2nd edn. Appleton Century Crofts, New York

Hoch SJ (1988) Who do we know: predicting the interests and opinions of the American consumer. J Consum Res 15(3):315–324

Hogan J, Barrett P, Hogan R (2007) Personality measurement, faking, and employment selection. J Appl Psychol 92(5):1270–1285

Holmes TH, Rahe RH (1967) The Social Readjustment Scale. J Psychosom Res 11(2):213–218

Homburg C, Wiescke J, Hoyer WD (2009) Social identity and the service-profit chain. J Mark 73(2):38–54

Hong S-M, Faedda S (1996) Refinement of the Hong psychological reactance scale. Educ Psychol Meas 56(1):173–182

Howard JA (1977) Consumer behaviour: application of theory. McGraw-Hill, New York

Hudson LA, Ozanne JL (1988) Alternative ways of seeking knowledge in consumer research. J Consum Res 14(4):508–521

Hull CL (1952) A behavior system: an introduction to behavior theory concerning the individual organism. Yale University Press, New Haven

Hung IW, Wyer RS (2008) The impact of implicit theories on responses to problem-solving print advertisements. J Consum Psychol 18(3):223–235

Hunter JE (1986) Cognitive ability, cognitive aptitudes, job knowledge, and job performance. J Vocat Behav 29(3):340–362

Hunter JE, Hunter RF (1984) Validity and utility of alternative predictors of job performance. Psychol Bull 96(1):72–98

Hurst G (2009) "UK unis face research freeze-out amid push for centralisation." The Australian, Jan 20, p 22

Ilies R, Fulmer IS, Spitzmuller M, Johnson MD (2009) Personality and citizenship behavior: the mediating role of job satisfaction. J Appl Psychol 94(4):945–959

James W (1884) What is an emotion?. Mind 9(34):188–205

James W (1892) Psychology: the briefer course. Harper, New York, 1961 reprint

Jensen AR (1970) Hierarchical theories of mental ability. In: Cuckrall WB (ed) On intelligence. Methuen, London

Judge TA, Hurst C, Simon LS (2009) Does it pay to be smart, attractive, or confident (or all three)? Relationships among general mental ability, physical attractiveness, core self-evaluations, and income. J Appl Psychol 94(3):742–755

Kabanoff B, Rossiter JR (1994) Recent developments in applied creativity. In: Cooper CL, Robertson IT (eds) International review of industrial and organizational psychology, vol 9. Wiley, London

Kahle LR (1983) Attitudes and social adaptation: a person-situation interaction approach. Pergamon, London

Kelley TL (1927) Interpretation of educational measurements. Macmillian, New York

Kerlinger FN, Lee HB (2000) Foundations of behavioral research, 4th edn. Wadsworth, Melbourne

Key WB (1974) Subliminal seduction. Prentice-Hall, Englewood Cliffs

Klaassen A (2009) "What's your brand's social score?" Advertising Age, Jul 13, pp 4, 24

Klimstra TA, Hale WW, Raaijmakers QAW, Branje SJT, Meeus WHJ (2009) Maturation of personality in adolescence. J Pers Soc Psychol 96(4):898–912

Kline P (2000) Handbook of psychological testing, 2nd edn. Routledge, London

Kotler P (1978) Marketing management: analysis, planning, and control. Prentice-Hall, Englewood Cliffs

Kover AJ (1995) Copywriters' implicit theories of communication: an exploration. J Consum Res 21(4):596–611

Krauss SJ (1995) Attitudes and the prediction of behavior: a meta-analysis of the empirical literature. Pers Soc Psychol Bull 21(1):59–75

Kusev P, van Schaik P, Ayton P, Dent J, Chater N (2009) Exaggerated risk: prospect theory and probability weighting in risky choice. J Exp Psychol Learn Mem Cogn 35(6):1487–1505

Lahey BB (2009) Public health significance of neuroticism. Am Psychol 64(4):241–256

Langer J (1984) Managing market research: the contribution of qualitative techniques. Mark Rev 40(2):25–31

Langner T, Fischer A, Rossiter JR, Kürten DA (2010). The behavioral consequences of "loving" versus "liking" a brand and a typology of the origins of "brand love." Paper presented at the 32nd INFORMS Marketing Science Conference, Cologne, Germany.

Laurent G (2000) Improving the external validity of marketing models: a plea for more qualitative input. Int J Mark Res 17(2,3):177–182

Law KS, Wong CS (1999) Multidimensional constructs in structural equation analysis: an illustration using the job perception and job satisfaction constructs. J Manag 25(2): 143–154

Lawson D (2009) "A rage to succeed, not talent. . ." The Independent, Jun 23, p 25

Levin G (1992) "Anthropologists in adland: researchers now studying cultural meanings of brands." Advertising Age, Feb 24, pp 3, 49

Levy SJ (1968) Social class and consumer behavior. In: Newman JW (ed) On knowing the consumer. Wiley, New York

Liao H, Toya K, Lepak DP, Hong Y (2009) Do they see eye to eye? Management and employee perspectives of high-performance work systems and influence processes on service quality. J Appl Psychol 94(2):371–391

Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 140:1–55

Locke J (1690) An essay concerning human understanding, Book II, Chapter XXXIII, para. 2. Oxford University Press, Oxford, 1979 reprint

Locke EA (2009) It's time we brought introspection out of the closet. Perspect Psychol Sci 4(1): 24–25

Lord FM (1980) Applications of item response theory to practical testing problems. Erlbaum, Hillsdale

Lord FM, Novick M (1968) Statistical theories of mental test scores. Addison-Wesley, Reading

Lumsden J (1978) On criticism. Aust Psychol 8(3):186–192

Lunt P, Livingstone S (1996) Rethinking the focus group in media and communications research. J Commun 46(2):79–98

Lykken DT (1982) Research with twins: the concept of emergenesis. Psychophysiology 19(4): 361–373

Lynn R (1997) Geographic variations in intelligence. In: Nyborg H (ed) The scientific study of human nature. Pergamon, Oxford, pp 259–281

Martin A (2007) The representative of object concepts in the brain. Annu Rev Psychol 58:25–45

Martineau P (1957) Social classes and spending behavior. J Mark 23(3):121–130

Maslow AH (1943) A theory of human motivation. Psychol Rev 50(4):370–396

Mautner T (2000) Dictionary of philosophy. Penguin, London

McClelland DC (1975) Power: the inner experience. Irvington, New York

McGuire WJ (1989) The structure of individual attitudes and attitude systems. In: Pratkanis AR, Breckler SJ, and Greenwald AG(eds) Attitude structure and function. Erlbaum, New Jersey, pp 37–68

McQuarrie EF (1989) Book review. J Mark 26(1):121–125

McQuarrie EF, McIntyre SH (1990) What the group interview can contribute to research on consumer phenomenology. Res Cons Behav 4:165–194

Merlo O, Auh S (2009) The effects of entrepreneurial orientation, market orientation, and marketing subunit influence on form performance. Manag Sci 20(3):295–311

Meyer GJ, Finn SE, Eyde L, Kay GG, Moreland KL, Dies RR, Eisman EJ, Kubiszyn TW, Reed GM (2001) Wanted: research on applied clinical judgment in personality assessment. J Pers Assess 86(2):226–227

Mick DG (1997) Semiotics in marketing and consumer research: balderdash, verity, please. In: Brown S, Turley D (eds) Consumer research: postcards from the edge. Routledge, London, pp 249–262

Milgram S (1963) Behavioral study of obedience. J Abnorm Soc Psychol 67(4):371–378

Miller D (1983) The correlates of entrepreneurship in three types of firms. Manag Sci 29(7): 770–791

Miller GA, Galanter E, Pribram KH (1972) Plans and the structure of behavior. Adams Bannister Cox, New York

Mills KI (2009). "More shocking results: New research replicated Milgram's findings." APA Monitor (March):13

Millsap RE (1994) Psychometrics. In: Sternberg RJ (ed) Encyclopedia of intelligence, vol II. Macmillan, New York, pp 866–868

Molenaar PCM, Campbell CG (2009) The new person-specific paradigm in psychology. Curr Dir Psychol Sci 18(2):112–117

Moran WT (1986) The science of qualitative research. J Advert Res 26(3):RC16–RC19

Moroney MJ (1951) Facts from figures. Penguin Books, Mitcham

Morrison MT, Haley E, Sheehan KB, Taylor RE (2002) Using qualitative research in advertising. Sage, Thousand Oaks CAPT, CAPI

Mosteller F, Youtz C (1990) Quantifying probabilistic expressions. Stat Sci 5(1):2–34

Mowrer OH (1960) Learning theory and behavior. Wiley, New York

Muckler FA, Seven SA (1992) Selecting performance measures: "objective" versus "subjective" measurement. Hum Factors 34(4):441–445

Munson JM, McIntyre SH (1979) Developing practical procedures for the measurement of personal values in consumer behavior. Adv Consum Res 15:381–386

Myers IB, McCaulley MH (1985) Manual: a guide to the development and use of the Myers-Briggs Type Indicator. Consulting Psychologists Press, Palo Alto

Narver JC, Slater SF (1990) The effect of a market orientation on business profitability. J Mark 54(4):20–35

Nisbett RE, Wilson TD (1977) Telling more than we can know: verbal reports on mental processes. Psychol Rev 84(2):231–259

Norman WT (1963) Toward an adequate taxonomy of personality attributes. J Abnorm Soc Psychol 66(6):574–583

Nunnally JC (1967) Psychometric theory, McGraw-Hill, New York

Nunnally JC (1978) Psychometric theory, 2nd edn. McGraw-Hill, New York

Osgood CE, Suci GJ, Tannenbaum PH (1957) The measurement of meaning. University of Illinois Press, Urbana

Ouelette JA, Wood W (1998) Habit and intention in everyday life: the multiple processes by which past behavior predicts future behavior. Psychol Bull 124(1):54–74

Overholser C (1986) Quality, quantity and thinking real hard. J Advert Res 26(3):RC7–RC12

Parasuraman A, Zeithaml VA, Berry L (1988) SERVQUAL: a multiple-item scale for measuring customer perceptions of service quality. J Retailing 64(1):12–40

Parasuraman A, Zeithaml VA, Malhotra A (2005) E-S-Qual: a multiple item scale for measuring electronic service quality. J Serv Res 7(3):213–233

Peabody D (1962) Two components in bipolar scales: direction and extremeness. Psychol Rev 69(2):65–73

Peissig JJ, Tarr MJ (2007) Visual object recognition: do we know more now than we did 20 years ago? Annu Rev Psychol 58:75–96

Perrault WD, Leigh LE (1989) Reliability of nominal data based on qualitative judgment. J Mark Res 26(2):135–148

Peterson RA (2001) On the use of college students in social science research: insights from a second-order meta-analysis. J Consum Res 28(3):450–461

Peterson RA, Albaum G, Beltrami RF (1985) A meta-analysis of effect sizes in consumer behavior experiments. J Cons Res 12(1):97–103

Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R (2000) Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. Br Med J 321:323–329 (August)

Petty RE, Fazio RH, Briñol P (eds) (2009) Attitudes: insights from new implicit measures. Psychology Press, New York

Pleyers G, Comeille O, Luminet O (2007) Aware and (dis)liking: item-based analyses reveal that valance acquisition via evaluative conditioning emerges only when there is contingency awareness. J Exp Psychol Learn Mem Cogn 33(1):130–144

Plug E, Vijverberg W (2005) Does family income matter for schooling outcomes? Using adoptees as a natural experiment. Econ J Roy Econ Soc 115(506):879–906

Reichheld FF (2003) The one number you need to grow. Harv Bus Rev 81(12):46–54

Reid S (1983). “Whiz kids from Japan, land of the rising IQ.” Sun-Herald, Dec 11, p 41

Revelle W (1979) Hierarchical cluster analysis and the internal structure of tests. Multivariate Behav Res 14(1):57–74

Reyna VF, Nelson WL, Han P-K, Dieckman NF (2009) How numeracy influences risk comprehension and medical decision making. Psychol Bull 135(6):943–973

Reynolds TJ, Jolly JP (1980) Measuring personal values: an evaluation of alternative methods. J Mark Res 17(4):531–536

Robinson JP, Shaver PR, Wrightsman LR (1991) Measures of personality and social psychological attitudes. Academic Press, San Diego

Rogers RW (1983) Cognitive and physiological processes in fear appeals and attitude change: a revised theory of protection motivation. In: Cacioppo J, Petty R (eds) Social psychophysiology. Guilford Press, New York

Rokeach M (1968) Beliefs, attitudes and values. Jossey-Bass, San Francisco

Rokeach M (1973) The nature of human values. Free Press, New York

Rossiter JR (1977) Reliability of a short test measuring children’s attitudes toward TV commercials. J Consum Res 3(4):179–184

Rossiter JR (1994) Commentary on A.S.C. Ehrenberg’s “Theory of well-based results: which comes first?” In: Laurent G, Lilien GL, Pras B (eds) Research traditions in marketing. Kluwer, Boston, pp 116–122

Rossiter JR (2001) What is marketing knowledge? Stage I: forms of marketing knowledge. Mark Theory 1(1):9–26

Rossiter JR (2002a) The C-OAR-SE procedure for scale development in marketing. Int J Res Mark 19(4):305–335

Rossiter JR (2002b) The five forms of transmissible, usable marketing knowledge. Mark Theory 2(4):369–380

Rossiter JR (2004) How to construct a test of scientific knowledge in consumer behavior. J Consum Res 30(2):305–310

Rossiter JR (2005) Reminder: a horse is a horse. Int J Res Mark 22(1):23–25

Rossiter JR (2007a) Toward a valid measure of e-retailing service quality. J Theor Appl Electron Commer Res 2(3):36–48

Rossiter JR (2007b) Identifying and measuring Australian values. Australas Mark J 15(1):7–13

Rossiter JR (2008a) Content validity of measures of abstract constructs in management and organizational research. Br J Manag 19(4):380–388

Rossiter JR (2008b) Defining the necessary components of creative, effective ads. J Advert 37(4):139–144

Rossiter JR (2009a) ER-SERVCOMPSQUAL: a measure of e-retailing service components quality. Serv Sci 1(4):212–224

Rossiter JR (2009b) Qualitative marketing research: theory and practice. Australas J Mark Soc Res 17(1):7–27

Rossiter JR, Bellman S (2005) Marketing communications: theory and applications. Pearson Prentice Hall, Sydney

Rossiter JR, Bellman S (2010) Emotional branding pays off. J Advert Res (forthcoming)

Rossiter JR, Bergkvist L (2009) The importance of choosing one good item for single-item measures of attitude towards the ad and attitude towards the brand and its generalization to all measures. Transf Werbeforschung Praxis 55(2):8–18

Rossiter JR, Danaher PJ (1998) Advanced media planning. Kluwer, Boston

Rossiter JR, Dolnicar S, Grün B (2010). The LFFB comparative judgment measure of brand-attribute beliefs. Working paper, Faculty of Commerce, University of Wollongong, Australia.

Rossiter JR, Foxall GR (2008) Hull-Spence behavior theory as a paradigm for consumer behavior. Mark Theory 8(2):123–141

Rossiter JR, Lilien GL (1994) New "brainstorming" principles. Aust J Manag 19(1):61–72

Rossiter JR, Overstead J (1999). Freudian symbolic ads: a failure to replicate their claimed effectiveness. Working paper no. 99-01, Faculty of Commerce, University of Wollongong, Australia.

Rossiter JR, Percy L (1987) Advertising and promotion management. McGraw-Hill, New York

Rossiter JR, Percy L (1997) Advertising communications and promotion management. McGraw-Hill, New York

Rossiter JR, Percy L, Donovan RJ (1991) A better advertising planning grid. J Advert Res 31(5):11–21

Rossiter JR, Robertson TS (1976) Canonical analysis of developmental, social, and experiential factors in children's comprehension of television advertising. J Genet Psychol 129: 317–327

Rossiter JR, Smidts A (2010) Print advertising: celebrity presenters. J Bus Res (forthcoming)

Rossiter JR, Thornton J (2004) Fear-pattern analysis supports the fear-drive model for anti-speeding road-safety TV ads. Psychol Mark 21(11):945–960

Rozin P (2009) What kind of empirical research should we publish, fund, and reward? A different perspective. Perspect Psychol Sci 4(4):435–439

Rungie C, Laurent G, Dall'Olmo Riley F, Morrison DG, Roy T (2005) Measuring and modeling the (limited) reliability of free choice attitude questions. Int J Res Mark 22(3): 309–318

Rushton JP (1997) (Im)pure genius–psychoticism, intelligence, and creativity. In: Nyborg H (ed) The scientific study of human nature. Elsevier, Oxford

Russell JA, Barrett LF (1999) Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. J Pers Soc Psychol 76(5):805–819

Russell JA, Weiss A, Mendelsohn GA (1989) Affect grid: a single-item scale of pleasure and arousal. J Pers Soc Psychol 57(3):493–502

Rust RT, Cooil B (1994) Reliability measures for qualitative data: theory and implications. J Mark Res 31(1):1–14

Rust RT, Zahorik AJ, Keiningham TL (1995) Return on Quality (ROQ): making service quality financially accountable. J Mark 59(2):58–70

Ruth WJ, Mosatche HS, Kramer A (1989) Freudian sexual symbolism: theoretical considerations and an empirical test in advertising. Psychol Rep 64(2):1131–1139

Sackett PR, Kunal NR, Arneson JJ, Cooper SR, Waters SD (2009) Does socioeconomic status explain the relationship between admission tests and post-secondary academic performance?. Psychol Bull 135(1):1–22

Salvado J (2009). "World marvels at greatest athlete of all time." Daily Telegraph, Aug 22, p 142

Saucier G (1994) Mini-markers: a brief version of Goldberg's unipolar Big-Five markers. J Pers Assess 63(3):506–516

Sauter D (2010) More than happy: the need for disentangling positive emotions. Curr Dir Psychol Sci 19(1):36–40

Sawyer AG (2009). Personal communication, August 7, e-mail.

Sawyer AG, Peter JP (1983) The significance of statistical significance tests in marketing research. J Mark Res 20(2):122–133

Schacter S, Singer J (1962) Cognitive, social, and physiological determinants of emotional state. Psychol Rev 69(5):379–399

Schutz A (1967) The phenomenology of the social world. Northwestern University Press, Evanston

Schmidt FL, Hunter JE (1998) The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. Psychol Bull 124(2):262–274

Schwartz SH (1992) Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. Adv Exp Soc Psychol 25:1–65

Scipione PA (1995) The value of words: numerical perceptions associated with descriptive words and phrases in market research reports. J Advert Res 35(3):36–43

Siegel S, Castellan NJ Jr. (1988) Nonparametric statistics for the social sciences, 2nd edn. McGraw-Hill, New York

Skinner BF (1935) Two types of conditioned reflex and a pseudo type. J Gen Psychol 12(1):66–67

Skinner BF (1938) The behavior of organisms. Appleton-Century, New York

Skinner BF (1959) Science and human behavior. Appleton Century Crofts, New York

Slater L (2004) Opening skinner's box: great psychological experiments of the twentieth century. Bloomsbury, London

Smith A (1999) Some problems when adopting Churchill's paradigm for the development of service quality measurement scales. J Bus Res 46(2):109–120

Smith B (2009) "Learning to read English is hardest – brand expert." Sydney Morning Herald, Nov 17, p 3

Spearman C (1904) General intelligence, objectively determined and measured. Am J Psychol 15(2):201–293

Speedy B (2009) "Seal of approval for Grange." The Australian, Dec 24, p 4

Staats AW, Staats CK (1958) Attitudes established by classical conditioning. J Abnorm Soc Psychol 57(1):37–40

Stahl G, Fowler GA (2009) "Amazon stock surpasses 1999." The Australian, Oct 26, p 25

Steenkamp J-BEM, Baumgartner H (1998) Assessing measurement invariance in cross-national consumer research. J Consum Res 25(1):78–90

Steiner I (1972) Group processes and productivity. Academic Press, New York

Su R, Rounds J, Armstrong PI (2009) Men, things, women, and people: a meta-analysis of sex differences in interests. Psychol Bull 135(6):859–884

Sundie JM, Ward JC, Beal DC, Chin WW, Geiger-Oneto S (2009) Schadenfreude as a consumption-related emotion: feeling happiness about the downfall of another's product. J Consum Psychol 19(3):356–373

Swain SD, Weathers D, Niedrich RW (2008) Assessing three sources of misresponse to reversed Likert items. J Mark Res 45(1):116–131

Tadajewsley M (2008) Incomensurable paradigms, cognitive bias and the politics of marketing theory. Marle Theory 8(3):273–297

Taft R, Rossiter JR (1966) The Remote Associates Test: divergent or convergent thinking? Psychol Rep 19(2):1313–1314

Taylor SA, Baker TL (1994) An assessment of the relationship between service quality and customer satisfaction in the formation of consumers' purchase intentions. J Retailing 70(2): 163–178

The Australian (2009). "Singer slams breast cancer guide." Nov 20, p 3

The Australian (2010). "Breast screening doesn't cut cancer rates." Mar 25, p 3

Thorndilce EL (1898) Animal intelligence: an experimental study of the associative processes in animals. Psychol Rev, Monogr Suppl 2(8).

Thoulless RH (1930) Straight and crooked thinking. Hodder and Stoughton, London

Tuccitto DE, Giacobbi PR, Leite WL (2010) The internal structure of positive and negative affect: a confirmatory factor analysis of the PANAS. Educ Psychol Meas 70(1):125–141

Tybur JM, Lieberman D, Griskevicius V (2009) Microbes, mating, and morality: individual differences in three functional domains of disgust. J Pers Soc Psychol 97(1):103–122

Urban GL, Hauser JR (1993) Design and marketing of new products. 2nd edn. Prentice-Hall, Englewood Cliffs

Uslaner EM (2008) Where you stand depends upon where your grandparents sat: the inheritability of generalized trust. Public Opin Q 72(4):725–740

van Rekom J, Jacobs G, Verleigh PWJ (2006) Measuring the essence of a brand personality. Mark Lett 17(3):181–192

Vaughan R (1980) How advertising works: a planning model. J Advert Res 20(5):27–33

Verhoef PC, Leeflang PSH (2009) Understanding marketing department's influence within the firm. J Mark 73(2):14–37

Viswanathan M, Sudman S, Johnson M (2004) Maximum versus meaningful discrimination in scale response: implications for validity of measurement of consumer perceptions about products. J Bus Res 57(2):108–125

Vroom VH, Yetton PW (1973) Leadership and decision-making. University of Pittsburg Press, Pittsburg

Walker R (1985b) Evaluating applied qualitative research. In: Walker R (ed) Applied qualitative research. Gower, Aldershot, pp 177–196

Walker R (1985a) An introduction to applied qualitative research. In: Walker R (ed) Applied qualitative research. Gower, Aldershot, pp 3–26

Wark P (2005) "Too negative for your own good?" The Times, Oct 17, pp 32–33

Warner WL, Meeker M, Eells KE (1949) Social class in America. Science Research Associates, Chicago

Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect. The PANAS scales. J Pers Soc Psychol 54(6):1063–1070

Webb EJ, Campbell DT, Schwartz RD, Sechrest L (1966) Unobtrusive measures: nonreactive research in the social sciences. Rand McNally, Chicago

Weiss RF (1968) An extension of Hullian learning theory to persuasive communications. In: Greenwald AG, Brock TC, Ostrom TM (eds) Psychological foundations of attitudes. Academic Press, New York, pp 147–170

Wells WD (1986) Truth and consequences. J Advert Res 26(3):RC13–RC16

Wells W (1993) Discovery-oriented consumer research. J Consum Res 19(4):489–504

West SG, Duan N, Pequegnat W, Galst P, Des Jarlai Holtgrave D, Szapocznik J, Fishbein M, Rapkin B, Clatts M, Mullen PD (2008) Alternatives to the randomized controlled trial. Am J Public Health 98(8):1359–1366

Westen D (1998) The scientific legacy of Sigmund Freud: toward a psychodynamically informed psychological science. Psychol Bull 124(3):333–371

Westen D, Weinberger J (2004) When clinical description becomes statistical prediction. Am Psychol 59(7):595–613

White KR (1982) The relationship between socioeconomic status and academic achievement. Psychol Bull 91(3):461–481

Wilson TD (2009) Know thyself. Perspect Psychol Sci 4(4):384–389

Wilson TD, Dunn DS, Kraft D, Lisle DJ (1989) Introspection, attitude change, and attitude-behavior consistency: the disruptive effects of explaining why we feel the way we do. Adv Exp Soc Psychol 22:287–343

Witte K, Allen M (2000) A meta-analysis of fear appeals: implications for effective public health campaigns. Health Educ Behav 27(5):591–615

Wyer RS (1974) Cognitive organization and change. Lawrence Erlbaum Associates, Potomac

Zaichkowsky JL (1994) The personal involvement inventory: reduction, revision, and application to advertising. J Advert 23(4):59–70

Zimiles H (2009) Comment. Ramifications of increased training in quantitative methodology. Am Psychol 64(1):51

Index