

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.

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 *n*th 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., $S \rightarrow M$), 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 $C \rightarrow M \rightarrow S$ 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 (CV_{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 inferable 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_{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 $CV = CV_{\text{item}} \times CV_{\text{answer}}$. The two content validity terms, CV_{item} and CV_{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 (CV_{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 CV_{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 <i>lower</i> 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_{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 ($C \rightarrow M$ in the Construct \rightarrow Measure \rightarrow 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 $S \rightarrow M$ inference according to my $C \rightarrow M \rightarrow S$ 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 Tatsuoka (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* (CV_{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 $C \rightarrow M \rightarrow S$ 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 Quiñones, 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)

Measurement for the Social Sciences
The C-OAR-SE Method and Why It Must Replace
Psychometrics

Rossiter, J.R.

2011, XV, 169 p., Hardcover

ISBN: 978-1-4419-7157-9