Neuendorf

Reliability and Validity

(For more info., see Carmines & Zeller, 1979)

<u>Reliability vs. Validity</u>

Reliability: The extent to which a measuring procedure yields the same results on repeated trials; for scales, it is the internal consistency among individual measures; for coding (e.g., content analysis), it is the agreement between or among coders. A measure is unreliable and therefore "no good" if it can be conducted only once, by one person, in one place, etc.

Statistical tests for each of the reliability types include:

1. Repeated trials: Test-retest (r), Alternative-form (r), Split-halves (Spearman-Brown)
2. Internal consistency: Cronbach's alpha
3. Intercoder agreement/covariation: Percent agreement, Scott's pi, Cohen's kappa, Krippendorff's K or alpha

Validity: The extent to which a measuring procedure represents the intended, and only the intended, concept; "Are we measuring what we want to measure?"

Random error: A threat to reliability. C&Z say that random error is "endemic to social research." It includes, but is not limited to, such sources as coding errors, data entry errors, ambiguous instructions, interviewer bias (e.g., differential emphases on words), and fatigue (of the interviewer or the respondent).

Nonrandom error: A threat to validity. Nonrandom error may also be called "bias."

<u>External vs. Internal Validity</u>

External validity: Generalizability. Can we generalize our findings to other people, settings, times, etc.? This includes such key considerations as the representativeness of the sample (e.g., whether it is randomly selected/a probability sample), and whether the measurement context is true-to-life (i.e., if there is ecological validity).

Internal validity: The correspondence of conceptual and operational definitions. Are we measuring/manipulating what we want to measure/manipulate? (The classic work on experimental design by Campbell and Stanley itemizes key threats to internal validity in experiments.)

<u>Types of Internal Validity</u> (mostly a la Carmines & Zeller)

1. Face validity: “On the face of things,” does the measure tap what we want? “Common sense” seems to be the test for this type of validity.

2. Criterion validity: The extent to which a measure taps an important behavior that is external to the measure; may be either concurrent validity or predictive validity. This may be tested via any appropriate bivariate statistic.

3. Content validity: The extent to which the measure reflects a specific domain of content. This generally applies to a scale (set of multiple measures), and is tested via simple inspection.

4. Construct validity: The extent to which a measure relates to other measures in ways that are consistent with theoretically derived hypotheses. This may be tested via appropriate bivariate or multivariate statistics.

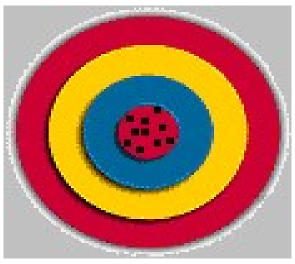
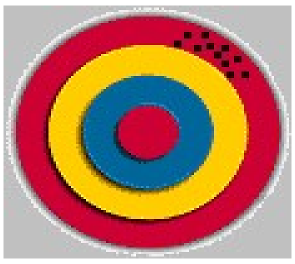
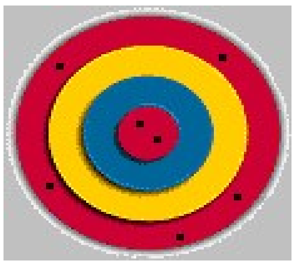
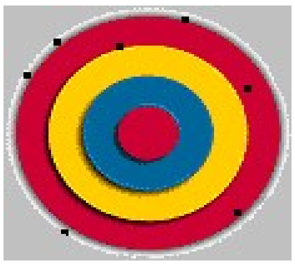
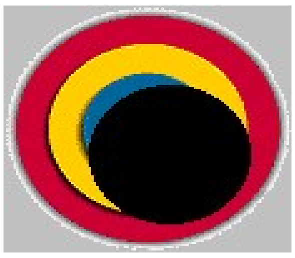
(Question: Is “internal consistency” reliability for a scale at odds with its content validity?)

Reliability, Accuracy, & Precision–Another model to consider

Internal Validity may be thought of as including several components:

1. Reliability
2. Accuracy--lack of bias (nonrandom error)
3. Precision–Fineness of distinction made between categories or levels of a measure

A graphical representation:

	
A) High Reliability High Accuracy High Precision	B) High Reliability Low Accuracy High Precision
	
C) Low Reliability Moderate Accuracy High Precision	D) Low Reliability Low Accuracy High Precision
	
E) High Accuracy Low Precision	

2/17