1. Scale Construction

(see Stangor, 1998, Chapter 4)

Scale Construction refers to the creation of empirical measures for theoretical constructs; these measures usually consist of several items.

The process of measurement involves the assignment of numbers to empirical realisations of the variables of interest.

Relations between these numbers reflect relations between different empirical realisations. Depending on the kind of relations that are meaningfully reflected in the numbers of our empirical measure, we speak of *nominal*, *ordinal*, *interval*, and *ratio* scales.

Can you provide examples for each scale level?

Psychologists usually try to achieve *interval scale* level, and the assumption of an interval scale underlies most computational techniques for assessing reliability and validity.

1.1 Conceptual and Measured Variables

Conceptual variables (or constructs) form the basis of research hypotheses and theories. Examples are *reading time*; *attitudes toward the Euro*; *self-esteem*; *depression*; *autism*.

Measurement turns conceptual variables into measured variables; these consist of numbers (and sometimes a unit of measurement).

The more abstract a construct, the greater the variety in possible measures.

Can you provide examples for this principle?

Operational definitions specify the procedures how to turn a construct into a measured variable.

Converging operations: No single research instrument or method in psychology will probably ever be free of systematic error (see below). Because different methodologies have different weaknesses, however, it seems wise to use "multiple measures that are hypothesized to share in the theoretically relevant components but have different patterns of irrelevant components" (Webb et al., 1981, pp. 34-35). By using different measures (multiple operationalisation), a researcher can *triangulate* on a construct of interest.

1.2 Self-Report Measures of Individual Differences and of Attitudes

Free-format measures allow research participants to express their thoughts or feelings relatively free of constraints imposed by the research instrument.

Examples are *think-aloud techniques* (e.g in research on problem solving); *free associations* (in projective testing); *thought-listing protocols* (in persuasion research).

These free-format answers are usually transformed into numerical data (= measured variables) by trained coders or raters who use a coding system. This process is called *content analysis* (for an introduction, see Krippendorf, 1980).

Fixed-format measures are more widely used, mainly because they are more economical in application. They usually consist of a set of questions or *items*, each accompanied by a *response scale* that limits the type of responses that a participant can give.

Sometimes measures consist of only one item, for example:

“How would you describe your sexual orientation (tick one):

___ heterosexual

___ homosexual

___ bisexual.”

In most cases, however, the conceptual variable of interest is so complex that single-item measures would produce unstable (= unreliable) outcomes if the construct was measured repeatedly.

Therefore, *multi-item scales* are used; these achieve greater reliability in two ways: (a) in the construction phase, inappropriate items, which do not meet certain measurement criteria, are eliminated; (b) the final score is the sum or mean of *all items*, which compenstaes for the unreliability of any single item.

The most widely used multi-item scale is the Likert scale (Likert, 1932). It can be used to assess *individual differences* (e.g., self-esteem) and *attitudes*.

Other variants of attitude scales are the Semantic Differential (Osgood, Suci, & Tannenbaum, 1957) and the Thurstone scale (Thurstone, 1928).

Guttman scales (Guttman, 1944) are sometimes used to assess the degree to which a person "possesses" a certain variable of interest (e.g., gender constancy in children).

[Discuss examples!]

Semantic differential scale assessing attitudes toward Germans

Germans

dirty :_____:_____:_____:_____:_____:_____:_____: clean

(-3) (-2) (-1) (0) (+1) (+2) (+3)

friendly :_____:_____:_____:_____:_____:_____:_____: unfriendly

(-3) (-2) (-1) (0) (+1) (+2) (+3)

bad :_____:_____:_____:_____:_____:_____:_____: good

(-3) (-2) (-1) (0) (+1) (+2) (+3)

beautiful :_____:_____:_____:_____:_____:_____:_____: ugly

(-3) (-2) (-1) (0) (+1) (+2) (+3)

Likert scale items assessing sexist attitudes toward women (from the Neosexism Scale; Tougas, Brown, Beaton, & Joly, 1995)

(Items with an asterisk are reverse-scored.)

Discrimination against women in the labor force is no longer a problem in Canada.

totally disagree 1 2 3 4 5 6 7 totally agree

I consider the present employment system to be unfair to women.*

totally disagree 1 2 3 4 5 6 7 totally agree

It is difficult to work for a female boss.

totally disagree 1 2 3 4 5 6 7 totally agree

In order not to appear sexist, many men are inclined to overcompensate women.

totally disagree 1 2 3 4 5 6 7 totally agree

In a fair employment system, men and women would be considered equal.*

totally disagree 1 2 3 4 5 6 7 totally agree

Thurstone scale items assessing attitudes toward euthanasia (Tordella & Neutens, 1979)

(Each item's scale values, here ranging from 1 to 5, are given in parentheses. These are not presented to respondents.)

A person with a terminal illness has the right to decide to die. (4.15)

Inducing death for merciful reasons is wrong. (1.65)

A person should not be kept alive by machines. (2.44)

Euthanasia gives a person a chance to die with dignity. (4.29)

The taking of human life is wrong no matter what the circumstances. (1.36)

(For an example of a **Guttman scale**, see Stangor, 1998, p. 71.)

In both Thurstone and Guttman scales, each item represents different degrees of the variable of interest (sometimes called the difficulty of an item).

In Likert and Semantic Differential scales, each item represents to the same degree the construct to be measured (i.e., strong agreement with any one item would be seen as an equivalent indication of the construct).

These differences between scales affect the computation of reliability (see below).

1.3 Reactivity and Nonreactive Measures

The validity of direct attitude measures may be threatened by *reactivity* effects. These are changes in the to-be-measured variable due merely to the fact that a measurement has taken place.

The organisation of an attitude questionnaire, the context of a study, or subtle cues in the experimenter's behaviour may indicate to participants that certain hypotheses are being tested. Participants may then choose to respond to these *demand characteristics* (Orne, 1962) in a fashion that either confirms or disconfirms these hypotheses. They may also engage in *"impression management"* (Tedeschi, 1981), trying to present themselves favourably rather than responding truthfully.

Strategies proposed to reduce such reactivity biases range from temporarily misinforming participants about the purpose of a study to asking for their cooperation by emphasizing the importance of truthful responses (see Aronson, Ellsworth, Carlsmith & Gonzales, 1990; Rosenthal & Rosnow, 1991).

However, not all biases stem from respondents' intention to sugarcoat their attitudes. Bias may result simply because research participants follow the *rules of natural conversation* (Schwarz, 1994). Respondents may use the context of a question to interpret its meaning, especially if they know little about the attitude object.

Example: Strack, Schwarz, and Wänke (1991) assessed German students' attitude toward a (fictitious) "educational contribution" in two different contexts: The preceding question either referred to the average tuition fees that US students have to pay, or it concerned financial support that Swedish students receive.

What do you think: Which of the two contexts produced more favourable attitudes toward the ominous "educational contribution"?

An important lesson can be learned from studies that deal with response biases in attitude measurement: There probably is no such thing as a person's "true" response; rather, all responses are context-dependent, and in interpreting empirical findings, it is usually a good idea to take contextual factors into consideration.

As an alternative to direct self-report measures, various *indirect* and *non-reactive measures* have been suggested. The latter include *behavioural observation measures*, *archival records* and *psychophysiological measures* (e.g., skin conductance measures, facial EMG). For overviews, see Himmelfarb (1993); Bohner (1995).

2. Reliability and Validity

Does a scale we use actually assess the construct of interest? If it does, we would say that the scale has high *validity*.

A necessary, but not sufficient, condition for high validity is high *reliability*, the extent to which a scale measures with high precision whatever it does measure.

2.1 Random and Systematic Error

You should be familiar with these concepts from our discussion of experimentation and the analysis of variance. In the context of scale construction, both sources of error constitute threats to a measures validity.

The scale value obtained is a function of (a) the construct of interest, (b) random error and (c) systematic error.

With a reliable and valid measure, components (b) and (c) should be small in comparison to (a) þ as long as there is no change in the construct of interest, repeated measurements should ideally yield identical measures.

Random error: Chance fluctuations in measurement (e.g. due to coding errors, variations in participants' attention to the items, misreading of questions, etc.).

In the long run, sources of random error should cancel each other out.

Systematic error: Influence of other conceptual variables that are not part of the construct to be measured.

These do not cancel each other out but rather may systematically increase or decrease scores.

[see Stangor, 1998, Figure 5.1, p. 81]

2.2 Forms of Reliability: Test-Retest, Internal Consistency, Interrater Agreement

Reliability of a measure is high to the extent that the measure is free from random error.

One way of assessing reliability that can in principle be used with all types of scale is conducting the measurement twice with the same sample and correlating the scores obtained. If the construct of interest can be assumed not to change much over time (which would be the case for *personality traits* and – to a lesser degree – *attitudes*) and the scale is perfectly reliable, this *test-retest reliability* will have a score of $r_{tt} = 1$.

Problem: Reactivity effects may be increased by repeated administration of the same scale.

Would this artificially increase or decrease r_{tt} ?

Alternative: Using *equivalent forms* instead of the same test twice.

Another – highly popular – way of assessing reliability is computing the *internal consistency* of a test. This is particularly useful when the construct of interest is not assumed to be stable over time (in this case, retesting would not yield high correlations, but not because of low reliability). Examples for such *state variables* are: mood, stress level, biological needs.

Internal consistency is assessed with a single administration of the scale, on the basis of the *intercorrelations among the scale items*. **Note that this approach requires the assumption that each item, in principle, reflects the construct of interest to the same extent.** Thus, internal consistency can be computed for Likert and Semantic Differential scales but would not be a meaningful indicator of reliability for Thurstone or Guttman scales.

The basic idea is that each of the items in a scale measures the construct of interest (the "true score") to some degree, but that the error components are uncorrelated across items. Thus, the more items are averaged or added to yield a scale score, the more reliable the measure will be.

"*Internal consistency* refers to the extent to which the scores on the items correlate with each other and thus are all measuring the true score rather than random error" (Stangor, 1998, p. 84).

One way of computing internal consistency known as *split-half reliability* is to correlate the score on one half of a scale's items with the score on the other half. If the scale is highly reliable, the correlation of the two halves will be close to 1.

However, the split-half coefficient will vary somewhat depending on which items are used to define the two halves of the scale (e.g. odd- versus even-numbered items, first half versus second half, random selection, etc.).

The most widely used index of internal consistency, ***Cronbach's Alpha***, avoids this problem – it is equivalent to the average of all possible split-half correlation coefficients.

Its formula is

$$\alpha = \frac{k}{k-1} \times \frac{\sigma_y^2 - \sum \sigma_i^2}{\sigma_y^2}$$

with k = number of items; $\mathsf{F}^{2}{}_{\mathrm{y}}$ = variance of the sum of all items; and $\mathsf{F}^{2}{}_{\mathrm{i}}$ = variance of the i-th item.

There is no need to memorise this equation. But note that, all else being equal, an increasing number of items leads to a higher reliability estimate. As more and more items are added, however, the impact of each single item is attenuated.

A final form of reliability that is applied with the coding of qualitative, free-format responses or with observational data, is *interrater reliability*.

If judges' ratings have been assessed with a numeric scale (e.g., "How aggressive was the child's behaviour?" – 1, *not at all*, to 7, *very*), correlational techniques (including Cronbach's Alpha) can be used, and independent judges' ratings can in principle be treated like items of a scale.

Sometimes, however, judges have to assign responses to categories – a nominal scaling (e.g. deciding whether children primarily played *alone*, *in a pair*, or *in a group*, see Stangor, 1998, pp. 343-344). Here, a Pearson correlation coefficient would be meaningless, and a different way of assessing agreement must be used: The most widely accepted coefficient for this application is ***Cohen's Kappa***, an index of agreement between two judges that can range from 0 (indicating only random agreement) to 1 (indicating complete agreement (see below).

2.3 Types of Construct Validity: Face, Content, Convergent and Discriminant Validity

Even if a measure is perfectly reliable, i.e. free from random error, it may still not measure what it is supposed to measure, i.e. contain systematic error.

Construct validity is high to the extent that a variable in fact measures the construct it is supposed to measure. It can be assessed in various ways.

Face validity is high to the extent that a variable apparently measures what it is supposed to measure. For example, the item "I think that women are generally inferior to men" would have high face validity as a measure of sexism. However, it would probably not be a valid measure because people may not answer it honestly (see for comparison, the less face-valid, but overall more valid items in the Tougas et al. scale above).

Generally, when reactivity problems are likely to arise, less face-valid measures are often preferable.

Content validity is high when a measure adequately captures the range of phenomena associated with the construct in question. For example, a test that only assessed knowledge of multiple regression techniques would be a poor measure of general mathematical ability.

Convergent validity means that a measure should be highly associated with other measures designed to assess the same variable or theoretically related variables.

Discriminant validity means that a measured variable should be unrelated to measures designed to assess other, conceptually unrelated variables.

For example, a self-report scale of self-esteem would be considered high in convergent validity to the extent that its scores are positively correlated with an observational measure of self-esteem-related behaviour and with close friends' ratings of the participants' self-esteem. It would be considered high in discriminant validity to the extent that its scores are uncorrelated with social desirability scores.

The complex pattern of relationships between a measured variable of interest and other measured variables, both self-report and other types, which constitutes an overall estimate of construct validity, has been called a "nomological net" (Cronbach & Meehl, 1955).

Criterion validity refers to a special way of assessing construct validity: by relating a self-report measure to a behavioural criterion or to the membership in groups that are known to differ in the variable of interest.

For example, to validate their measure of “need for cognition” (NC; a tendency to engage in and enjoy thinking), Cacioppo and Petty (1982) compared academics and blue-collar workers and found, as predicted, that the latter scored considerably lower on NC than the former.

3. Computing Reliability and Validity

3.1 Internal Consistency: Cronbach’s Alpha; SPSS RELIABILITY

The internal consistency of Likert scales or semantic differential scales can be computed by the SPSS procedure RELIABILITY. It can be found in the menu under Analyze – Scale – Reliability Analysis.

The items to be included in the scale should be defined as variables in your SPSS file and – importantly – should all be scored in the same direction (if they are not, odd results may be found such as negative Cronbach’s Alphas).

To reverse-score items use the recode command. For three items (variable names: item1, item2, item5) that were measured on a 7-point scale from 1 to 7, such a recode command (to be typed in a a syntax window) would read:

```
RECODE item1, item2, item5
          (1=7)(2=6)(3=5)(5=3)(6=2)(7=1).
```

Once the “Reliability Analysis” window is open, paste the variables to be included in the scale into the “Items” window.

Then, in the “Model” pull-down list, you may select the coefficient you want to estimate (choices are Cronbach’s *Alpha*, *Split-half* and three additional coefficients). Usually you want to estimate Cronbach’s Alpha (the default).

The “Analyze” window allows you to look at several useful statistics at the item level and the scale level. I suggest to tick all entries in the “Descriptives for” and “Summaries” boxes, as well as inter-item correlations.

For a reliable scale, all inter-item correlations should be positive, alpha should be high (> .80 would be considered satisfactory for most purposes), and the item-total correlations should not be too low.

It may be that by removing an item with very low item-total correlation, the overall alpha would increase. This information can be gleaned from the last column of the “Item-total statistics” table that is part of the output (“Alpha if item deleted”).

3.2 Interrater Reliability with Nominal Data: Cohen’s Kappa; SPSS CROSSTABS

When two independent coders assign *N* observations to *k* nominal categories, an obvious measure of reliability seems to be their relative agreement, i.e. the number of observations they assigned to the same category divided by the total number of observations. But is it?

Consider the following example: Two coders observed N = 100 motorists who approached a pedestrian crossing while a person was waiting to cross the street. Each coder assigned each motorist’s behaviour to one of three categories: (a) stopped; (b) decelerated but did not stop; (c) passed without decelerating.

Results:

	Coder 2			
Coder 1	(a)	(b)	(c)	Total
(a)	**50**	6	4	60
(b)	3	**15**	2	20
(c)	9	9	**2**	20
Total	62	30	8	100

The number of observed agreements between Coders 1 and 2 (bold numbers in main diagonal) is

$$\sum f_o = 50 + 15 + 2 = 67$$

The raw proportion of agreement would thus be .67 – however, this would not take into account the fact that coders would agree on some of the codings by chance alone.

As in a chi-square analysis, we can compute expected chance agreements by multiplying the row and column marginals and dividing by *N*:

$$\sum f_e = \sum \frac{\text{row marginal} \times \text{column marginal}}{N}$$

yielding

$$\sum f_e = \frac{60 \times 62}{100} + \frac{20 \times 30}{100} + \frac{20 \times 8}{100} = 44.8$$

To compute the coefficient kappa, we correct both numerator and denominator by this chance expectation:

$$\kappa = \frac{\sum f_o - \sum f_e}{N - \sum f_e} = \frac{67 - 44.8}{100 - 44.8} = 0.402$$

Thus, Cohen's Kappa (Cohen, 1960) yields an estimate that is lower than simple proportion of agreement, because some agreement would always be expected by chance.

Note that kappa can reach its maximum of 1 only if both coders have identical marginals – otherwise there cannot be perfect agreement, because some off-diagonal entries must be unequal to zero.

Cohen's Kappa can be produced with SPSS procedure CROSSTABS, which can be found under Analyze – Summarize – Crosstabs. Paste the variables that represent the codings of coder 1 and coder 2 in the "Rows" and "Columns" window, respectively. Click on "Analyze" and tick "Kappa".

Your SPSS output will contain the table of agreement, kappa, an estimate for the standard error of kappa, and a significance test based on t with df = N-2.

Note that a significant result simply means that agreement is better than chance. As with other measures of reliability, you should aim for kappa approaching 1 and accept kappa > .80 as sufficient for most purposes.

3.3 Test-Retest Reliability and Construct Validity: SPSS CORRELATION

To compute test-retest reliability and construct validity, simple Pearson correlations can be performed. In the case of test-retest reliability, we would report the correlation between measurements at time 1 and time 2 and the time interval between measurements.