SCALE DEVELOPMENT AND VALIDATION

SEAMEO
REGIONAL CENTRE FOR EDUCATION IN SCIENCE & MATHEMATICS

THIEN LEI MEE, Ph.D.
R & D Specialist
SEAMEO RECSAM
Penang
leimee@recsam.edu.my/
crystalmee@yahoo.com
Mobile: +60194752541

Researchers need to have a fairly well-developed knowledge of conceptual & methodology/technical procedure (e.g., structural equation modeling).

Issue 1

(1) Fail to adequately discuss how to develop appropriate conceptual definitions of the focal construct.

"…*many researchers think they have a clear idea of what they wish to measure, only to find out that their ideas are more vague than they thought. Frequently, this realization occurs after considerable effort has been invested in generating items and collecting data—a time when changes are far more costly than if discovered at the outset of the process.*"

(DeVellis, 1991, p. 51)

"*There is no way to know how to test the adequacy with which a construct is measured without well specified domain… .*"

(Nunnally & Bernstein, 1994, p. 88)

Consequences (MacKenzie et al., 2005)

1. Confusion about what the construct does and does not refer to, and the similarities and differences between it and other constructs that already exist in the field

Consequences

(2) Indicators that may either be *deficient* because the definition of the focal construct is not adequately fleshed out, or *contaminated because the definition overlaps* with other constructs that already exist in the field.

Consequences

(3) Invalid conclusions about relationships with other constructs that later have to be rejected because the indicators of the focal construct are not really capturing what they are intended to capture.

(2) Often fail to properly specify the measurement model that relates the latent construct to its indicators.

Formative versus Reflective measurement model specification

Figure 1: A Reflective Construct

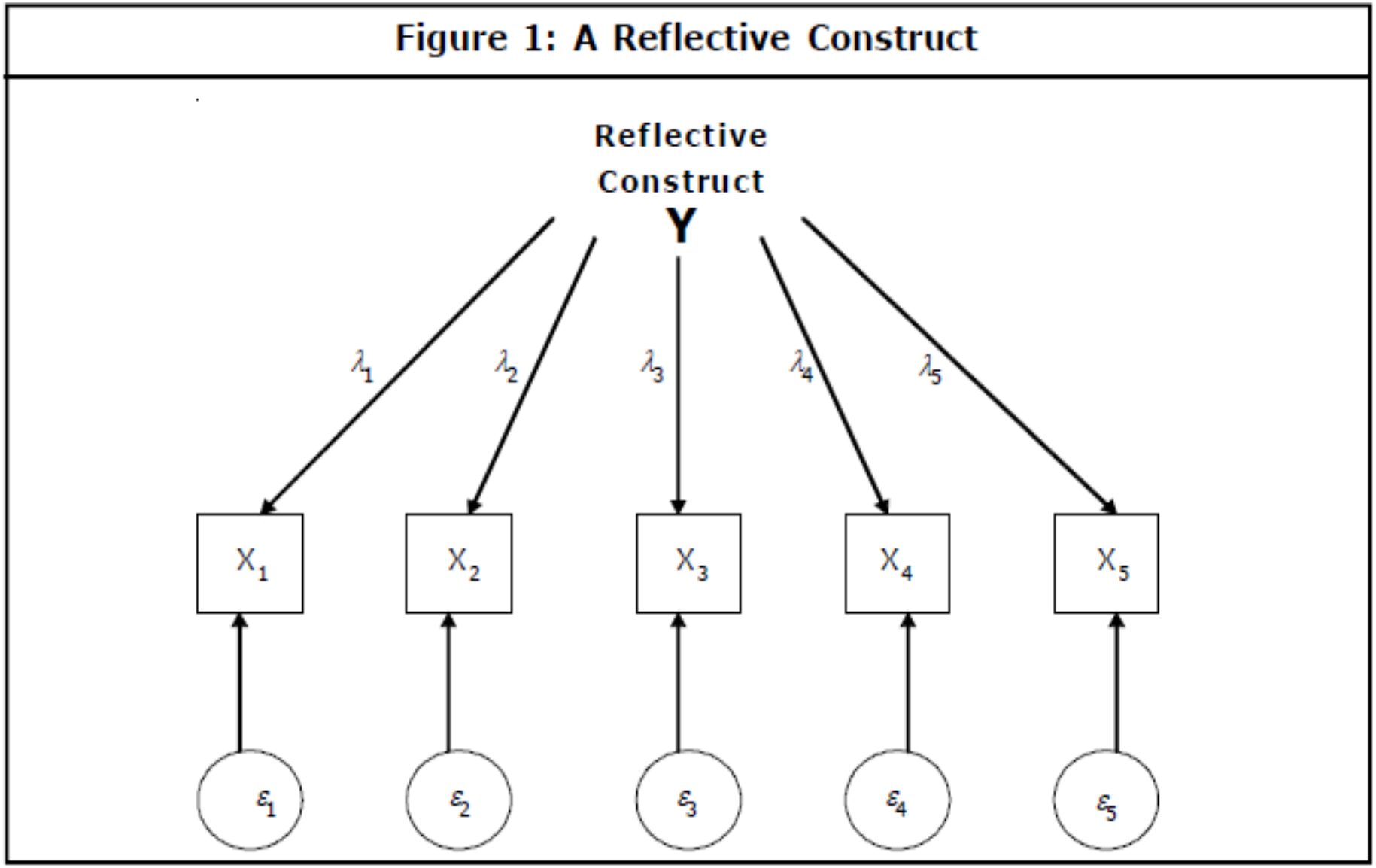
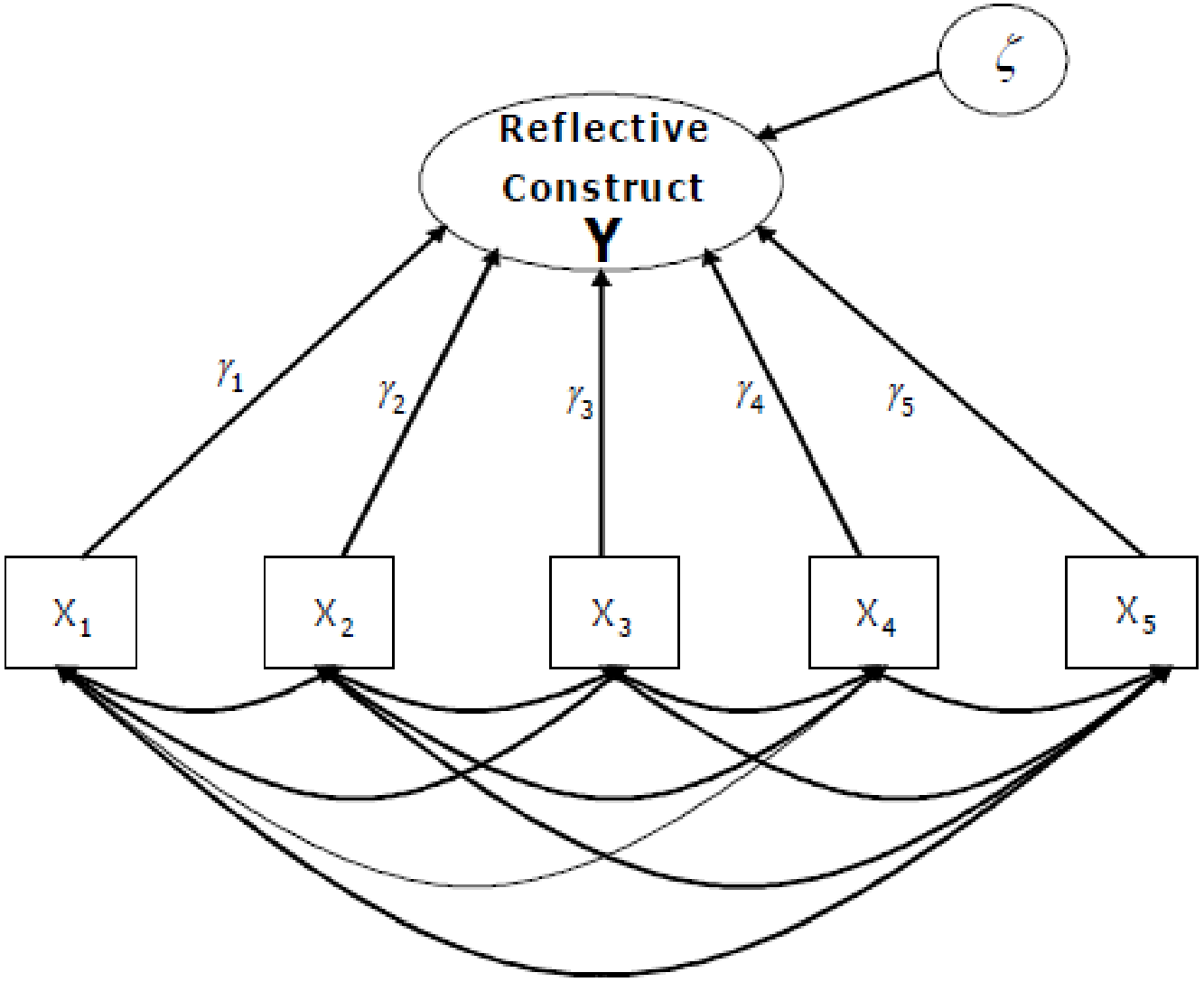


Figure 2: A Formative Construct



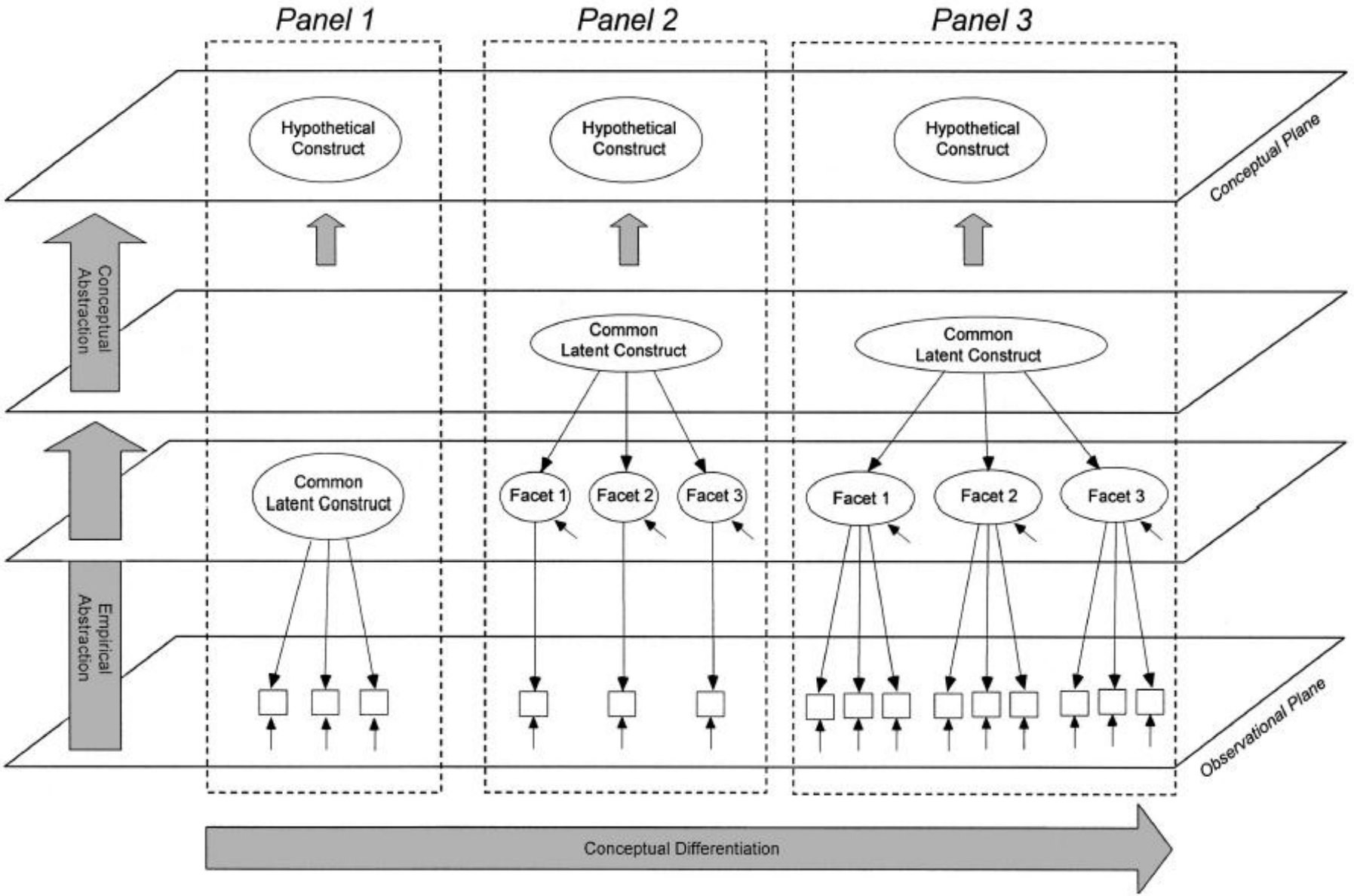


Figure 2. Reflective-indicator measurement models.

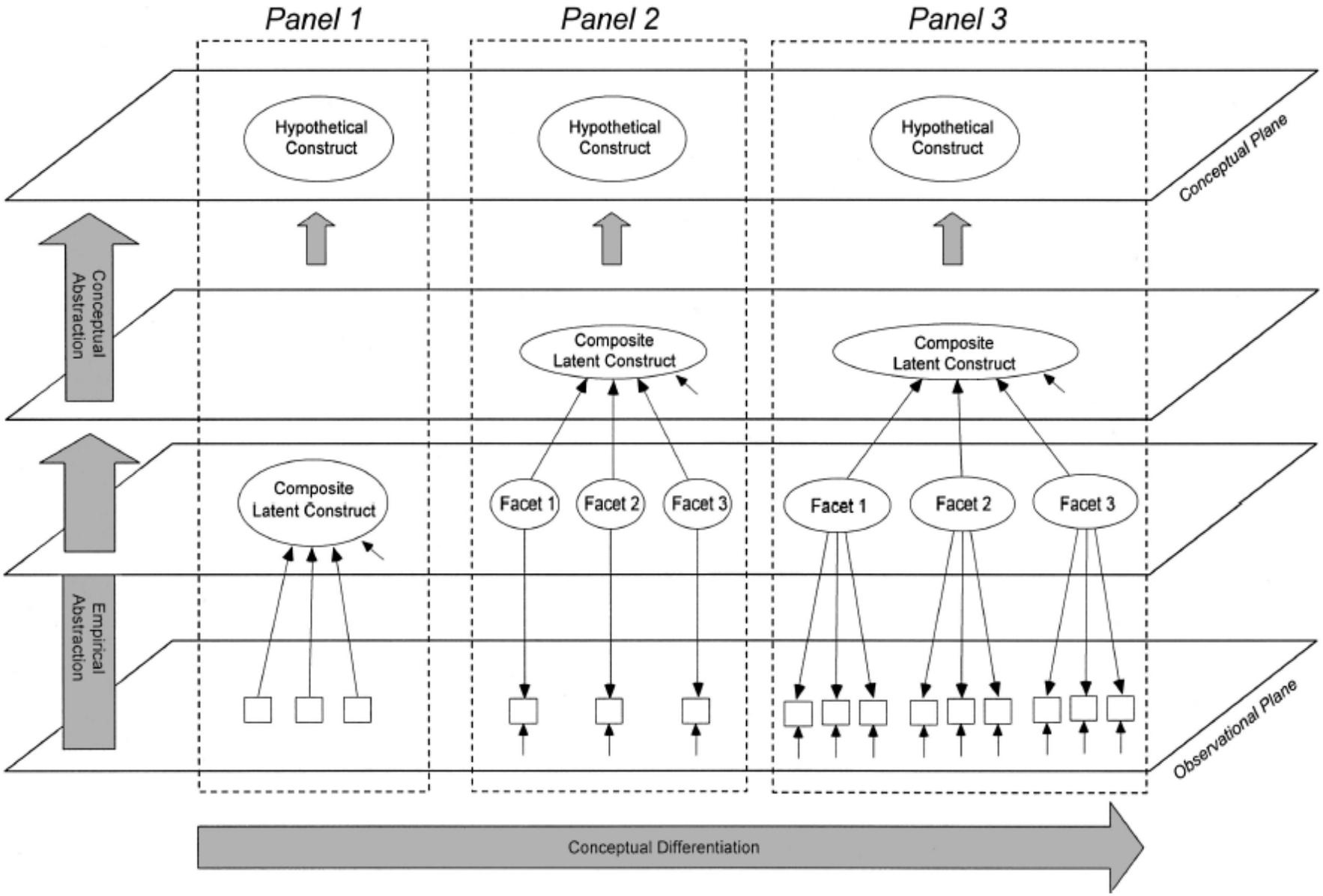


Figure 3. Formative- and mixed-indicator measurement models.

Reflective vs. Formative Construct

Reflective Construct	Formative Construct
1. Construct is reflected in the indicators	1. Construct is a composite of indicators
2. Account for observed variances in the outer model – error is assess at the item level.	2. Minimize residuals in the structural relationship-error is assessed at the construct level.
3. Identification achieved with three effect indicators	3. Identification is given only if the construct is embedded into a larger model
4. Important aspects: ▪ Internal consistency or reliability ▪ Positive correlation between measures ▪ Unidimensionality allows for removing indicators to improve construct validity without affecting the content validity	4. Important aspects: ▪ Indicators examine different dimensions ▪ Multicollinearity is a problem ▪ Removing an indicator affects content validity

- Thien, L. M., Ramayah, T., & Nordin, A. R. (*in press*). Specifying and Assessing Formative Measure of Hofstede's Cultural Values. *Quality & Quantity Journal* (2012 ISI Impact factor: 0.728). doi: 10.1007/s11135-013-9959-5.

(3) Underutilize techniques that provide evidence that the set of items used to represent the focal construct actually measures what it purports to measure.

MOVING BEYOND TRADITIONAL PSYCHOMETRIC APPROACHES

- *Researchers frequently employ outdated statistical procedures to evaluate their measure's psychometric properties….they simply fail to capture the full potential of the data…*

Sass & Schmitt (2013, p.)

First Issue

- Statistical and methodological decisions before factor analysis either EFA or CFA.

(1) An appropriate sample size;
(2) A factor model and estimation method;
(3) A valid method to determine the “correct” number of factors; and
(4) EFA rotation criterion.

(Schmitt, 2011; Sass, 2011, 2013)

Second Issue

- Modeling the latent constructs either EFA or/and CFA.
- Researchers should not assume CFA is the appropriate modeling approach simply because previous studies have provided evidence of simple structure (Schmitt, 2013).
- Consider also Exploratory Structural Equation Modeling (ESEM).

Third Issue

- Ignorance of assessing measurement equivalence Across Groups (e.g., gender, ethnics, types of primary/secondary schools)

Fourth Issue

- Only report Cronbach's alpha to help support a measure's internal consistency reliability.
- Should beware of tau equivalent.

(Sass & Schmit, 2013)

Conclusion

- Using appropriate & current statistical and methodological approaches;
- Estimating the correct measurement model;
- Testing for measurement invariance and
- Providing robust estimates of internal consistency reliability.

- *…relying solely on ‘traditional’ approaches not only precludes valid results, but also hinders the advancement of science.*

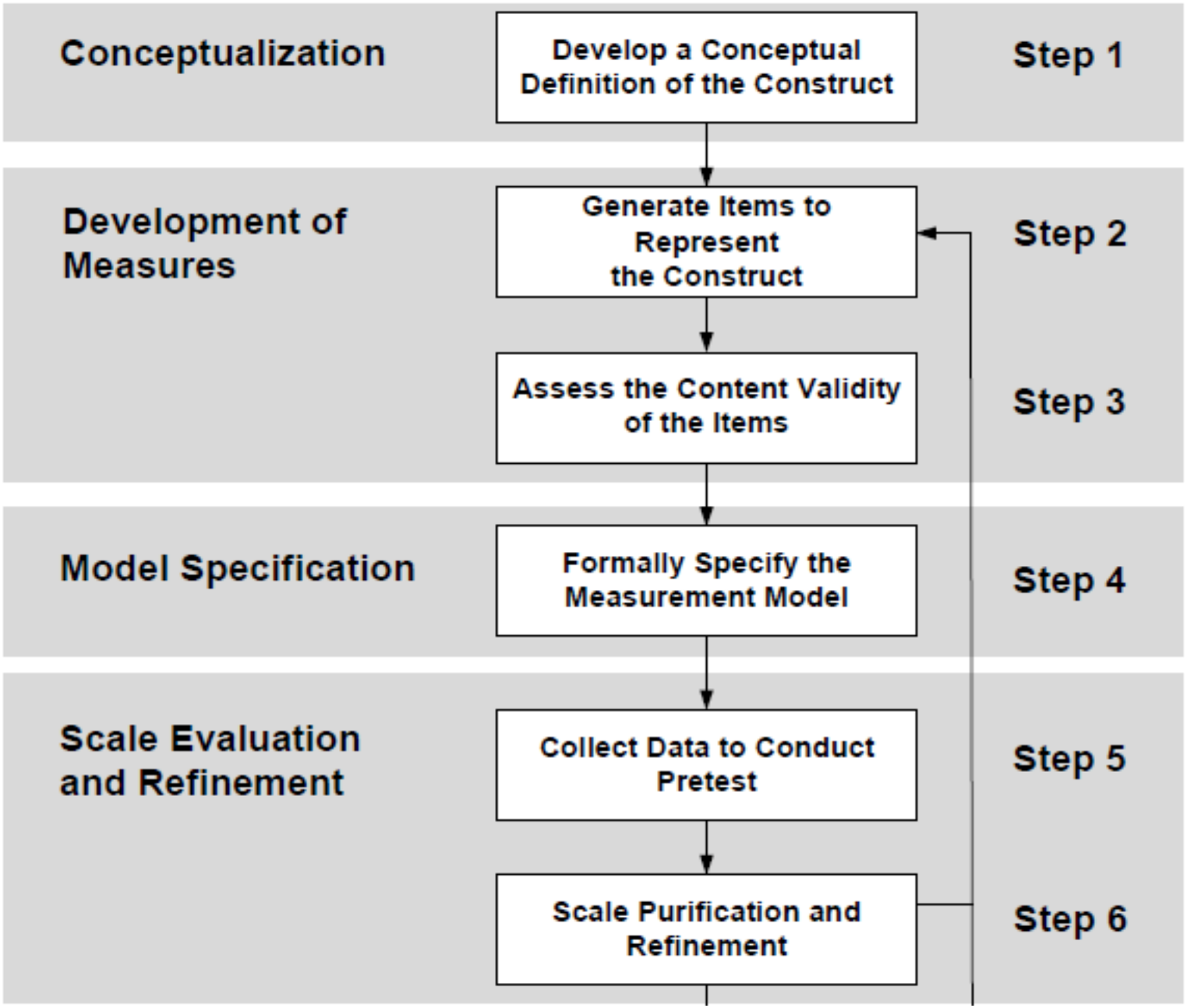
(Sass & Smith, 2013, p. 302)

However,…

Writing good construct definitions requires clear conceptual thinking and organization, the lack of which becomes apparent as soon as the researcher tries to write a tight conceptual definition of the construct (Petter et al., 2007)

How about if you want to explore the domain of a construct as it exists in a non-western culture and proceeds to develop a culturally sensitive multiple-item scale?

Scale Development and Validation


Conceptualization
Develop a Conceptual Definition of the Construct
Step 1
Development of Measures
Generate Items to Represent the Construct
Step 2
Assess the Content Validity of the Items
Step 3
Model Specification
Formally Specify the Measurement Model
Step 4
Scale Evaluation and Refinement
Collect Data to Conduct Pretest
Step 5
Scale Purification and Refinement
Step 6

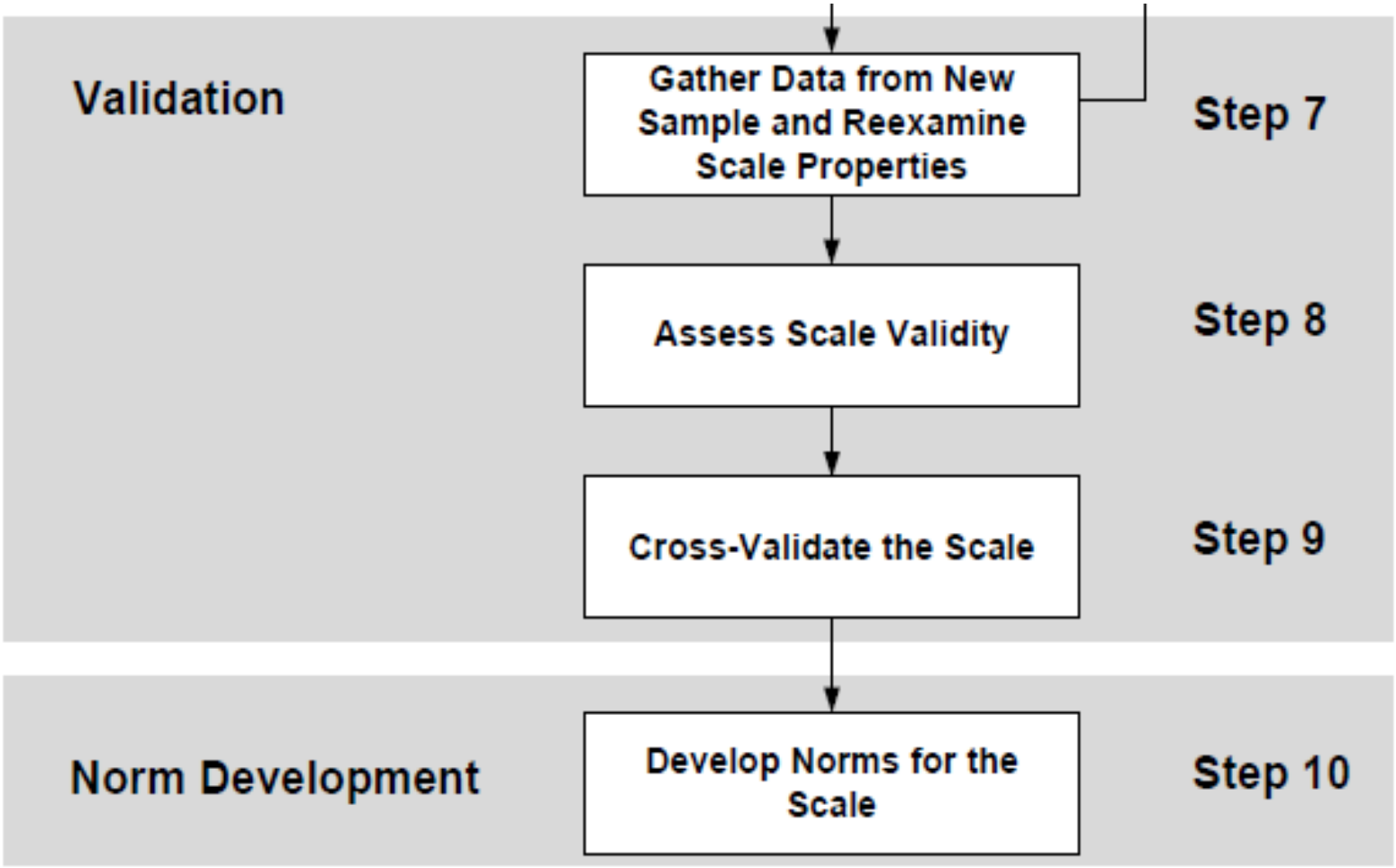


Figure 1 Scale Development & Validation

Source: MacKenzie, Podsakoff, & Podaakoff (2010)

Step 1: Construct Conceptualization

Requires the identification of what the construct is intended to conceptually represent or capture, and also a discussion of how the construct differs from other related constructs.

Step 1: Construct Conceptualization

Researcher should specify the nature of the construct and its conceptual theme in unambiguous terms and in a manner that is consistent with prior research (MacKenzie 2003).

Step 1: Construct Conceptualization

- However, this stage is the one that is often neglected or dealt with in a superficial manner (e.g., by assuming that labeling or naming the construct is equivalent to defining it).

- This leads to a significant amount of trouble later in the validation process.

Some considerations…

- Examine how the focal construct has been used in prior research or by practitioners;
- Specify the nature of the construct's conceptual domain;
- Specify the conceptual theme of the construct (unidimensionality/multidimensional); and
- Define the construct in unambiguous terms.

Step 2: Generate Items to Represent the Construct

- Items that fully represents the conceptual domain of the construct

Sources of Items generation:

1. Reviews of the literature,
2. Deduction from the theoretical definition of the construct,
3. Previous theoretical and empirical research on the focal construct,

Sources of Items generation:

4. Suggestions from experts in the field,
5. Interviews or focus group discussions with representatives of the population(s) to which the focal construct is expected to generalize, and
6. An examination of other measures of the construct that already exist.

Author	*Construct*	*Description (Scoring)*	*Items*
Gaski and Nevin (1985)	Perceived coercive power (in a marketing channel)	Supplier capability to take different kinds of action (0 = "no capability," 4 = "very much capability")	1. Delay delivery 2. Delay warranty claims 3. Take legal action 4. Refuse to sell 5. Charge high prices 6. Deliver unwanted products
McKee, Varadarajan, and Pride (1989)	Advertising expenditures (bank)	Spending on advertising compared with primary competitor (4 = "much more," 3 = "more," 2 = "same," 1 = "less")	1. Television 2. Radio 3. Newspaper 4. All media in total
Lumpkin and Hunt (1989)	Convenience (shopping)	Importance of different aspects (1 = "not important," 2 = "below average importance," 3 = "average importance," 4 = "above average importance," 5 = "very important")	1. Delivery to home 2. Telephone in order 3. Transportation to store 4. Convenient parking 5. Location close to home 6. Variety of stores close together
Burke (1984)	Company resource sharing	Extent of sharing of resources among business units (1 = "not all," 7 = "great")	1. Plant and equipment 2. Production personnel 3. Sales force 4. Distribution channels 5. Management services 6. Research and development facilities 7. Research and development personnel

Item should be written so that …

- Wording is as simple and precise as possible.

- Double-barreled items (e.g., “confident and smart”) should be split into two single-idea statements, and if that proves impossible, the item should be eliminated altogether.

Item should be written so that …

- Items that contain ambiguous or unfamiliar terms should be clarified,
- Items that possess complicated syntax should be simplified and made more specific and concise.
- Finally, efforts should also be made to refine or remove items that contain obvious social desirability (Nederhof 1985).

Step 3: Content Validity Assessment

- Content validity concerns “the degree to which items in an instrument reflect the content universe to which the instrument will be generalized.”

(Straub et al., 2004, p. 424)

Judgments of Content Validity

(1) Is the individual item representative of an aspect of the content domain of the construct?

(2) Are the items as a set collectively representative of the entire content domain of the construct/each dimensions of the construct?

How to assess the content adequacy of new measures?

- Content Validation Ratio (CVR)
- Q-Sorting
- Cohen’s Kappa
- Expert judgment

Hinkin & Tracy (1999)

Table 2. Hypothetical Example of Item Rating Task to Assess Content Adequacy

Rater Number = 001 Trustworthiness Scale Items†	***Benevolence*** is the degree to which the trustor believes that the trustee has goodwill or positive intentions toward the trustor (Serva et al. 2005, p. 630).	The other party's ***ability*** to accomplish a task important to the trustor, where ***ability*** is the set of skills or attributes that enable the trustee to have influence (Serva et al. 2005, pp. 629-630).	***Integrity*** is a trustor's perception that the trustee adheres to acceptable values, which could include issues such as consistency, honesty, and fairness (Serva et al. 2005, p. 630).
1. The management team really looked out for what was important to our development team.	4	2	1
2. Our development team felt that the management team was very capable of performing its job.	1	5	2
3. Our development team believed that the management team tried to be fair in dealings with others.	1	1	5

Table 3. Example of a Data Array for Content Adequacy Scores

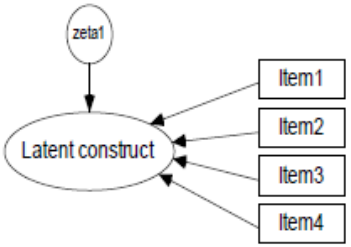
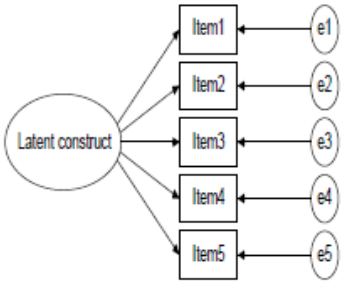
Rater Number	Aspects of Trustworthiness	Item #1	Item #2	Item #3	Item #4	Item #5	Item #6	Item #7	Item #8	Item #9
001	Benevolence	4	1	1	5	1	2	5	1	2
	Ability	2	5	1	1	5	1	2	5	1
	Integrity	1	2	5	2	2	4	2	1	5
002	Benevolence	5	2	1	5	2	1	5	1	1
	Ability	1	5	1	1	4	1	1	5	2
	Integrity	1	1	5	1	2	5	2	1	4
n...	Benevolence									
	Ability									
	Integrity									

Content Validation Assessment

- A one-way repeated measures ANOVA is used to assess whether an item's mean rating on one aspect of the construct's domain differs from its ratings on other aspects of the construct's domain.

Step 4: Measurement Model Specification

The Key Features of Formative and Reflective Measurement Model

Description	Formative Measurement Model	Reflective Measurement Model
Epistemic relationship		
Criteria		
(1) Direction of causality	Flow from measures to latent construct. Measures define latent construct.	Flow from latent construct to measures. Construct defines measures.
(2) Interchangeability of the measures	Measures should not be interchangeably. Measures need not to have the similar content or common theme. Dropping measures should not change the conceptual meaning of the latent construct.	Measures should be interchangeably. Measures should have the similar content or common theme. Dropping measures should not change the conceptual meaning of the latent construct.
(3) Correlation among the measures	Not necessary for measures to correlate with each other.	Measures are expected to correlate with each other.
(4) Nomological Network	Nomological net of the measures may differ.	Nomological net of the measures should not differ.

Source: Urbach and Ahlemann (2010)

Step 5: Collect data to conduct pre-test

- Data need to be obtained from a sample of respondents in order to examine the psychometric properties of the scale, and to evaluate its convergent, discriminant, and nomological validity.

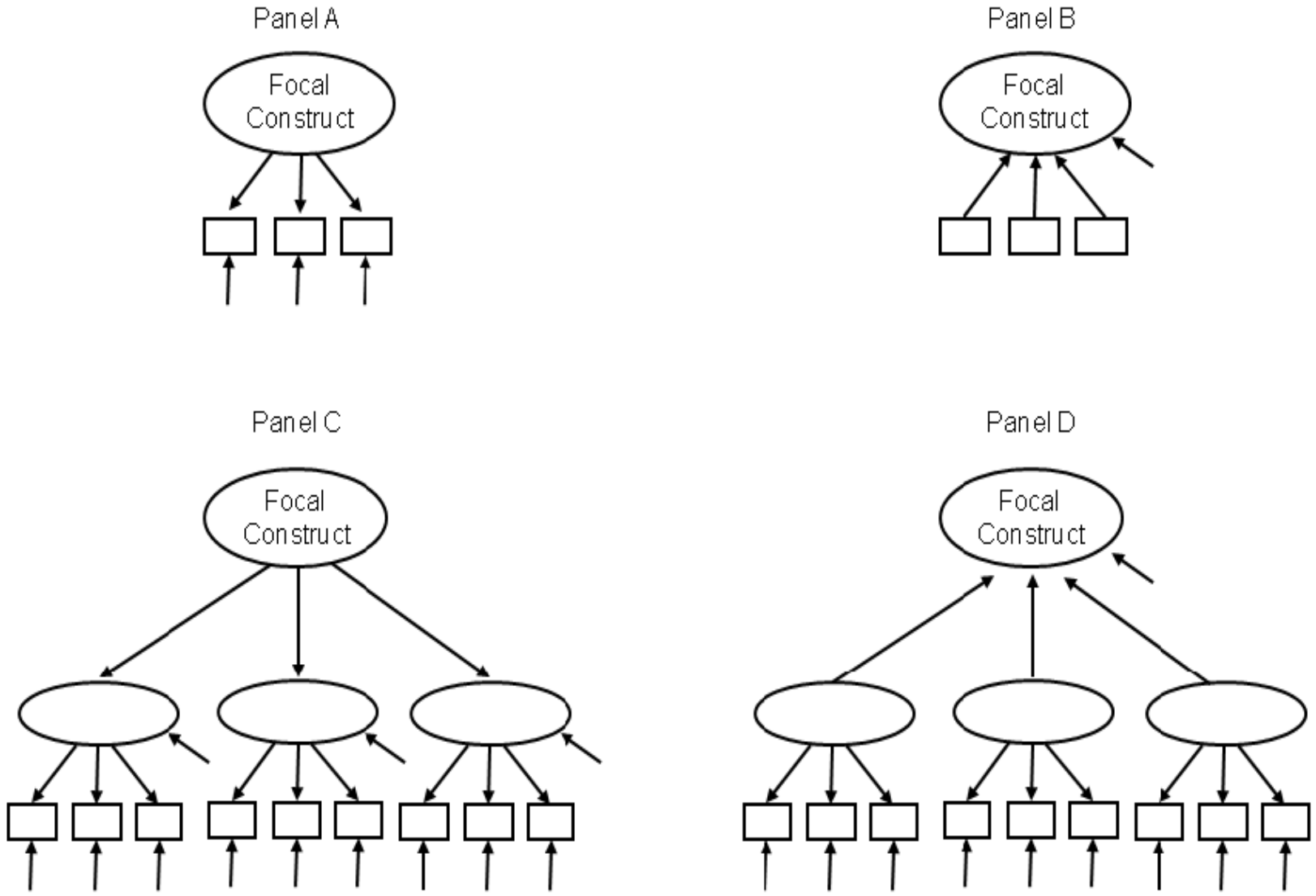
Sample

- Must be representative across ethnic, gender etc.
- Sample size:
- EFA -> from 100 to 500 (Comrey & Lee 1992; Gorsuch, 1983),
- The minimum ratio of the number of respondents to the number of items in the scale range from 3:1 to 10:1 (Cattell 1978; Everitt 1975).
- Also depends on types of analysis/ stability over time.

Step 6: Scale Purification & Refinement

(1) Methods for evaluating constructs with formative or/and reflective indicators

(Details can be refer MacKenzie , Podasakoff, & Podsakoff, 2011)



Note: The models in Panels B and D are not identified as shown, but can be estimated after resolving the identification problem as recommended in Step 4.

Figure 3. Examples of First-Order and Second-Order Measurement Models

Evaluation of the measurement model

Reflective measurement model	Formative measurement model
• Internal consistency/composite reliability	Collinearity among indicators
•Indicator reliability	Significance and relevance of outer weights/loadings
•Convergent validity (AVE)	Nomological net/external validity
•Discriminant validity	

Rules of Thumb for Evaluating Reflective Measurement Model (Hair et al., 2013)

International consistency reliability:

- CR> 0.708
- 0.60 < CR < o.70 acceptable for exploratory study

- Indicator reliability:

-outer loading > 0.708

-0.40 < outer loading < 0.70 consider for deletion only if the deletion leads to increase the AVE above the suggested threshold.

Rules of Thumb for Evaluating Reflective Measurement Model

- Convergent validity
- AVE > 0.50

- Discriminant validity

  Fornell-Larcker (1981) criterion – the square root of the AVE > the highest correlation with any other construct

Rules of Thumb for Evaluating Formative Measurement Model (Hair et al. 2013)

- Convergent validity
- Redundancy analysis ->examining its correlation with reflective measures or a global single items. The correlation should be 0.80 or higher.

Rules of Thumb for Evaluating Formative Measurement Model (Hair et al. 2013)

- Collinearity of indicators

-each indicator's tolerance/Variance Inflation factor (VIF) should be

- >0.1 (<10) (Diamantopoulos et al. 2008; Diamantopoulos & Winklhofer 2001) or
- > 0.20 (< 5) (Hair et al., 2013) or
- >3.3 (<3) (Petter et al., 2007).

Step 6: Scale Purification & Refinement

(2) Criteria for eliminating problematic indicators.

Reflective indicators :

- low validity,
- low reliability,
- strong and significant measurement error covariances, and/or
- non-hypothesized cross loadings that are strong and significant are candidates for elimination.

Formative indicators

- Outer weight significance testing using bootstrapping

(1) if significant, then continue with the interpretation.

Formative indicators

- Outer weight significance testing

(2) if not significant, then analyze the formative indicators' outer loadings

(i) If outer loading is < 0.5, test the significance of the formative indicator's outer loading

- If not significant, then delete the formative indicator
- If significant, then consider the removal of the indicator

Formative indicators

- Outer weight significance testing

(2) if not significant, then analyze the formative indicators' outer loadings

(i) If outer loading is > 0.5, keep the indicator even though it is nonsignificant.

A must read references:

- Diamantopoulos, A., & Siguaw, J. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management, 17*(4), 263–82.

- Diamantopoulus, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*, 269-277.

Evaluation of the measurement model

- Case study illustration - reflective measurement model

  Hair, J. F., Hult, G. T.M., Ringle, C. M., & Sarstedt, M. (2013*). A primer on Partial Least Square Structural Equation Modeling (PLS-SEM)*. Thousand Oaks: Sage. (pg. 107-112)

Evaluation of the measurement model

- Case study illustration – formative measurement model

  Hair, J. F., Hult, G. T.M., Ringle, C. M., & Sarstedt, M. (2013*). A primer on Partial Least Square Structural Equation Modeling (PLS-SEM)*. Thousand Oaks: Sage. (pg. 139-161)

References for higher-order measurement model

- Becker, J-M, Klein, K., & Wetzels, M. (2012). Hierarchical latent variable models in PLS-SEM: Guidelines for using reflective-formative type models. *Long range Planning, 45*, 359-394.

Step 7: Gather Data from New Sample and Re-examine Scale Properties

- Re-estimating the measurement model using a new sample of data is crucial because items are often added, dropped, or reworded in the scale purification process.

Step 8: Scale Validity Assessment

- To evaluate whether responses to the scale behave as one would expect if they were valid indicators of the focal construct.

Step 8: Scale Validity Assessment

To evaluate whether the indicators of the focal construct:

(1) are accurate representations of the underlying construct,

(2) adequately capture the multidimensional nature of the construct,

(3) are distinguishable from the indicators of other constructs (discriminant validity),

4) are related to the measures of other constructs specified in the construct' theoretical network (nomological validity).

Nomological Validity

(1) To specify the nature of the lawful relationships between the focal construct and other constructs, and

(2) to test whether the indicators of the focal construct relate to measures of other constructs in the.

Nomological validity

- The validity of one's measures of the focal construct should increase if they are related to measures of other constructs in a manner that is

  consistent with prior theory.

Nomological validity

- The other constructs refers to antecedents, correlates, or consequences of the construct of

  interest identified in previous research.

Nomological validity

- Antecedents are constructs that are hypothesized to cause the focal construct.
- Consequences are constructs that are hypothesized to be caused by the focal construct.
- Correlates are constructs whose conceptual definitions overlap with the focal construct.

* Researcher needs to collect the data of antecedent/consequence prior to data analysis.

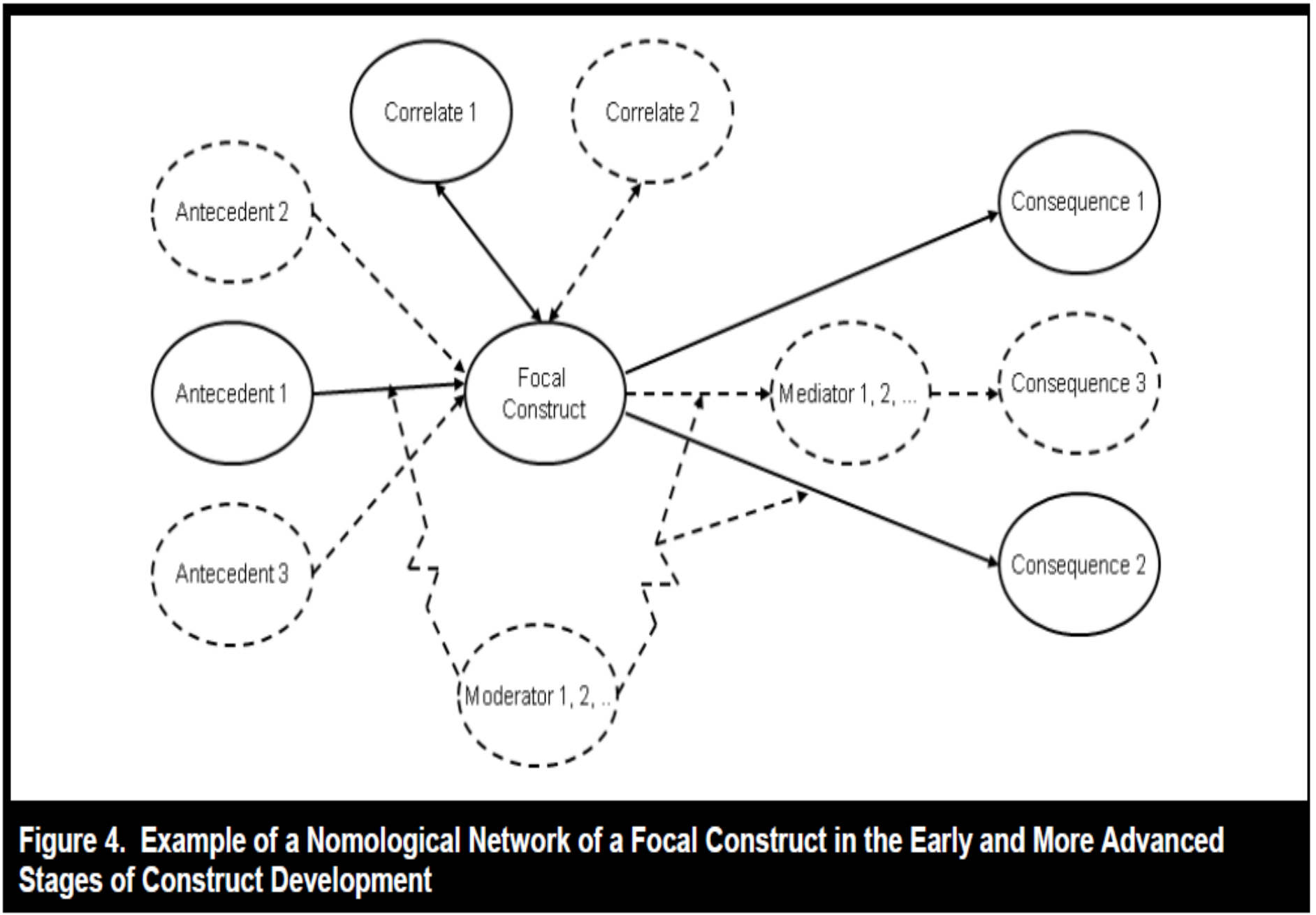


Figure 4. Example of a Nomological Network of a Focal Construct in the Early and More Advanced Stages of Construct Development

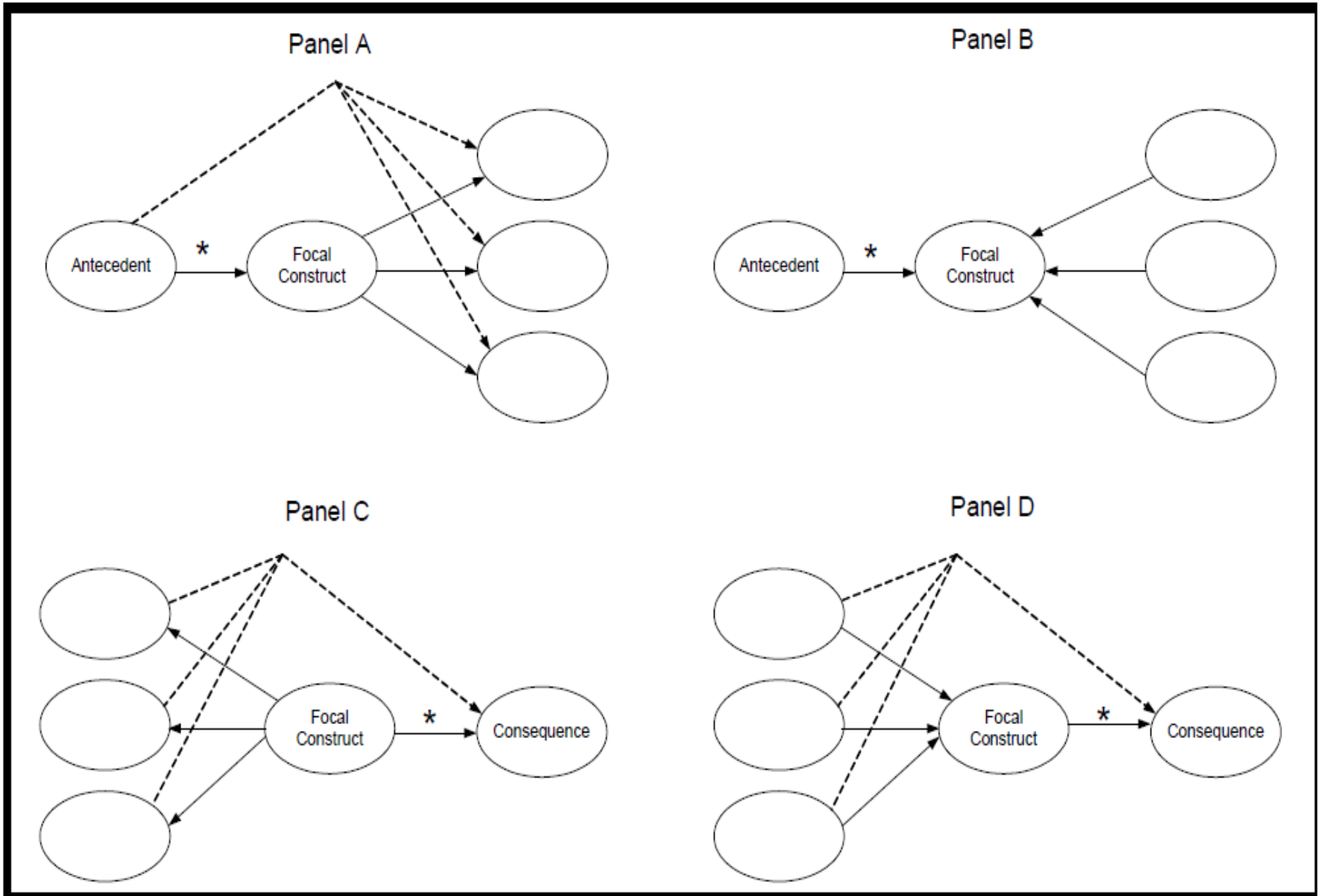


Figure 5. Illustration of Relationships Between a Multidimensional Focal Construct and its Antecedents or Consequences

Nomological validity

- In Figure 5, the relationship between the focal construct and one of its antecedents or consequences is marked with an asterisk (*).
- The statistical significance of the coefficients for these paths provides the key test of nomological validity of the focal construct's indicators.
- If these paths are significant, it suggests that the focal construct relates to other constructs as specified in the nomological network, thus increasing confidence in the validity of the indicators.

Step 9: Cross Validate the scale

- To cross validate the psychometric properties using new samples.

- This is particularly important if model modifications were made in the scale development and refinement process.

Step 9: Cross Validate the scale

- The new samples should be another population to which the construct would be expected to apply.

- For constructs with reflective indicators, the measurement estimates obtained from the developmental sample could be compared to the estimates obtained from the validation samples using the procedures recommended by Steenkamp and Baumgartner (1998) and Vandenberg and Lance (2000).

Step 9: Cross Validate the scale

- Steenkamp and Baumgartner (1998) and Vandenberg and Lance (2000) recommend using multigroup analysis to compare a series of nested models with systematically increasing equality constraints across groups to test:

(1) the equivalence of the covariance matrices,

(2) the configural equivalence of the factor structure,

(3) the metric equivalence of the factor loadings, and

(4) the scalar equivalence of the item intercepts.

** Refer Byrne & Stewart (2006) for second-order construct.*

Step 10: Develop Norm for the Scale

- To develop norms to aid in the interpretation of scores on the scale.

- "*The scale of measurement for most constructs in the social sciences is arbitrary. The meaning of a score can only be determined in relation to some frame of reference*" .

(Spector, 1992, p 67)

Step 10: Develop Norm for the Scale

- The biggest barrier to the development of scale norms is the difficulty of obtain in “representative” samples of the population to which one desires to generalize.

What we have learnt today?

(1) Without a clear definition, it is difficult to avoid contamination and deficiencies in the set of items used to represent the construct or to specify the relationship between the construct and its indicators.

(2) If the indicators do not adequately capture the domain of the construct, there may be little value in examining their psychometric properties or the relationships between these indicators and the indicators of other constructs.

What we have learnt today?

(3) If the measurement model is improperly specified, it may lead to inappropriately dropping items that are necessary to capture the complete domain of the construct, result in the use of inappropriate scale evaluation indices, and bias estimates of the relationships between the construct and other constructs.

What we have learnt today?

(4) If the researcher does not properly test the measurement model and evaluate the scale, it is difficult to determine whether the hypothesized measurement relationships are consistent with the data or to know how to refine the scale to improve its psychometric properties.

What we have learnt today?

(5) Unless the scale is cross-validated across subject populations, situations, and time, it will be difficult to evaluate the limits of its generalizability or its usefulness in other contexts.