**­­­Week 1 （Basic Concepts）**

* Assessment: A purposeful and disciplined set of **procedures** (embodied in an instrument) aimed at describing, quantifying, and facilitating **inferences** about degrees to which particular characteristic(s) exist in given groups of persons, objects, events, organizations/ entities.
* Assessment tools: wide variety of measuring tools.
  + Test
  + Rating Sales
  + Checklist
  + Observation Form
  + Portfolios
* Evaluation: A judgement-based decision pertinent to whether a construct measure is “good” or “bad” for a given purpose.
* Construct: The **attribute** or **trait** that we attempt to assess with an instrument. All educational characteristics, such as the writing ability of students, are constructs.
* Operational definition: The definition of a construct with **observable indicators**. Identification of the actual responses, tasks, operations, behaviors, or other relevant indicators that represent a construct.
* Types:
  + Alternative assessment vs. authentic assessment
  + Minimum competency vs. standards-based assessment

**Week 2**

* Construct(s)
  + Representativeness
  + Multidimensional: overall health = physical + mental
  + Operational variables/indicators
    - Generally, one indicator cannot represent the construct.
* Population/population unit
  + Can be individual or group/population level
* Assessment purposes
  + Purpose can be more than one, but should not be too broadly misused since the constructs and the theories behind can be very different.
  + Primary assessment users
  + What link of inference: formative (which is good and which is not) and summative (pass/fail)?
    - Inference can be diagnostic or predictive.
  + External accountability
* Assessment operation 🡺 to narrow the error
  + Sampling (sometime, no need to include all indicators) a set of operational indicators and clustering them based on the domain.
  + Scale of the items
    - nominals, ordinal (Likert and rating scale), internal (standardized achievement tests, and scale of psychological attribute), and ratio (physic).
    - the differences between ordinal and internal is whether it can be adding and subtracting, and compare result data.
    - The differences between internal and ratio data is whether there is a meaningful zero value (able to multiply and divide)
  + Objective vs. subjective 🡺 self-report survey.
  + High-stack and low-stack (how precise the measure have to be).
  + Sometimes provide not applicable options.
  + Norm-reference score or raw score.
* Empirical methods (traditional) for instrumental design.
  + process
    - Based on the common sense
    - Testing the group of people
    - Use factor analysis to see how the item cluster
    - Looking at the loading of the methods
  + risk
    - Population-independent?
    - Theorical framework/foundations?
* Classical testing theory
  + Observed score = true score + error (X = T + E)
* Phase 4: measurement methods
  + Domain sampling approach
    - Review the literatures under the related topic/construct
    - Find every single indicator of construct
    - Sample the indicator
      * Content relevant
      * Content representativeness: the sampled indictors should be able to represent the whole domain
  + Validation
    - Clarify your purpose, use, and population
    - X can be raw score (sum of all numerically coded items), derived score (scores derived from the raw score using a statistical or psychometric procedure), composite score (sum of scores from more than one instrument scale/sub-scale, or data source), or aggregated/group scores (mean of score from more than one individual in a group or organization)
* Read chapter 3 and 13!!!

**Week 3**

* **Some notes for homework example**

1. the research question should end up with some constructs which can be quantitively measured, so that the research question can be doable. Do not do half quantitative and qualitive.

2. construct definitions can be very week. But only if it is clearly defined, we can find the resource in the corresponding domain to structure the operational indicators.

3. In measurement, we need to ensure the representativeness of sampling of item and the participants.

4. the measurement can be multiple units: e.g., individual team member and the team performance as whole. Thus, the constructs are also defined at multi-level.

5. the domain can be very different and should be separated.

6. program evolution: effectiveness, cognitive testing, and readiness.

7. some studies can start with cognitive response processes/test 🡺 qualitative study with interview.

* **General framework**

|  |  |  |  |
| --- | --- | --- | --- |
| Type of evidence | Validation Question | Empirical methods | Rational |
| Validity |  |  |  |
| Reliability |  |  |  |

**User-centered model of instrument design**

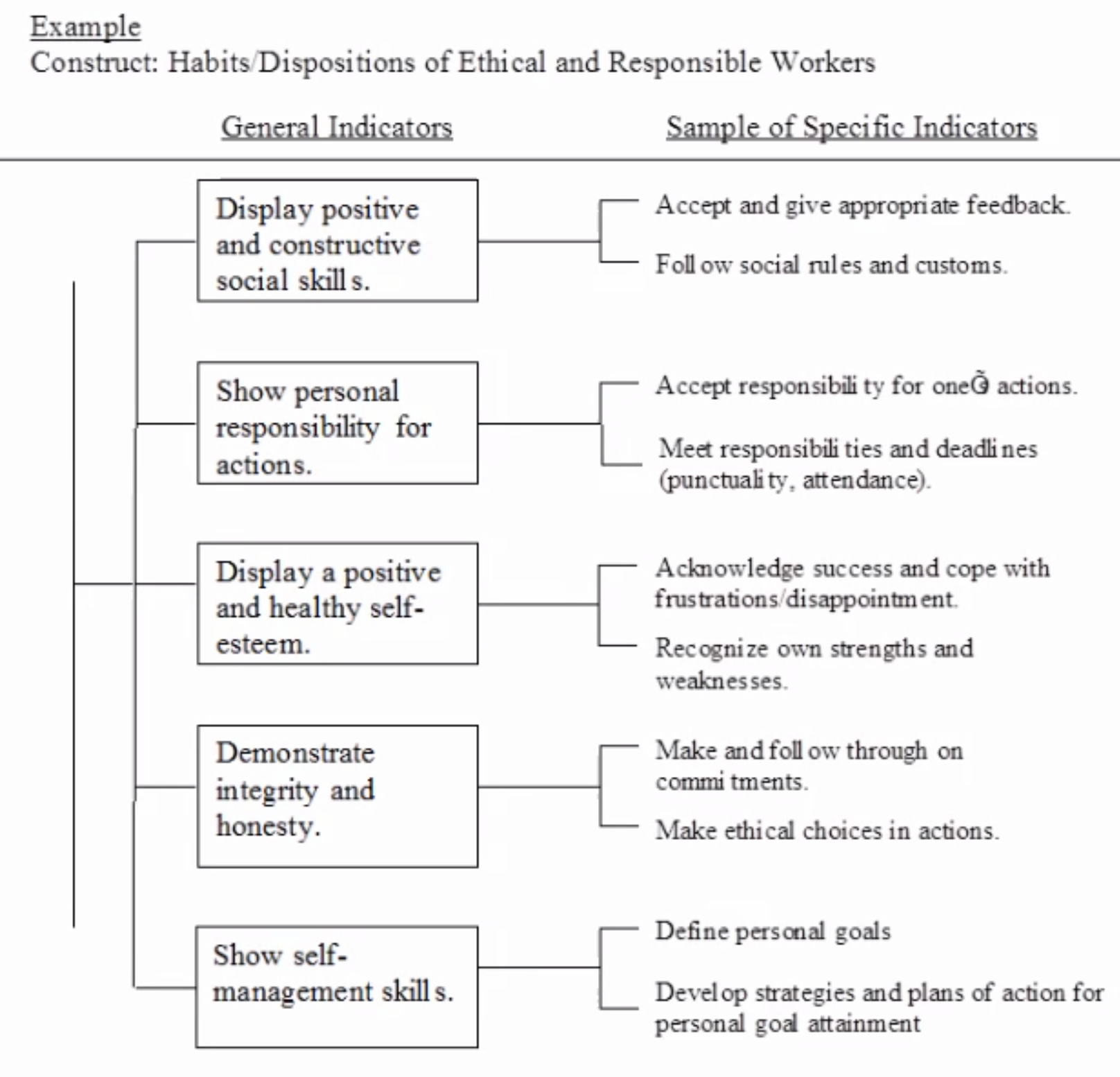
* Validity (construct validity)
  + To what degree are the **constructs (latent factors)** measures of interest, in the *population* units specified, and for the specific *inferences* and assessment *purpose* intended by primary *users*.
  + E.g., cultural, gender, or history biases. For example, inaccessible assessment modality or presentation format for some disabled population.
  + The meaningfulness of the score and measurement.
  + Systematical error of the measurement (easier to predict).
  + 3 Types (Trinitarian view) 🡺 construct validity (Unitarian view)
    - Construct validity
    - Predictive validity
    - Content validity
* Reliability
  + How precise and consistent are measures under different conditions.
    - Random error of the measurement: e.g., rater influence (cannot be clearly predicted that certain population will be affected).
    - Bring more item in the instrument usually reduce the random error.
  + E.g., waving focus or distractedness of respondents
* Utility
  + Usability and practicality of instrument and constructs measures for intended use and users.
    - E.g., is the instrument too long?
  + It will influence the validity and reliability as well.
* Classical test theory: X = T + E
  + E can be the error can be reliability (random errors) and validity (systematical errors).
* The primary user and population and construct should be commonly defined to cover the range.
* Types of validity evidence:
  + Content-based validity: expert view, CVI or content validity indices based on agreement
  + Correlational-based validity:
    - Convergent validity: overlap with the similar measures
      * Which can be used to decide whether we can replace one measurement (usually long test) with another one (usually short one test)
    - Discriminant validity: not correlated with the dissimilar measures
    - Criterion-related (predictive) validity: predict the **future** behaviors
      * How to find the correct criterion?
  + Evidence of internal structure: fit the conceptual structure of the construct
    - Factor analysis studies (SEM)
  + Controversial notion: face validity
    - Do the participants really believe the instrument measures what it says it does?

**Week 4**

* **Assessment assignment feedback**
  + The phase one of the process model is critical. We need to do it before the real instrumental design.
  + Key components: construct, population, and assessment purpose. These components are related to the design of the instruments.
  + Cognitive and non-cognitive construct:
    - Bloom et al. (1956) types of construct
      * Cognitive: ability
        + Method: testing
        + right/wrong answer 🡺 good or bad
      * Affective: attitude
        + Method: questionnaires
        + No right/wrong answer
        + Natural variability
      * Psychomotor ability
        + Meet the cut point for the development of a certain population in terms of a specific skill.
    - Eagly & Chaikin (1994) dimension of construct
      * Cognitive: belief, value, and opinion
      * Conative: behaviors and practices
      * Affective: emotional feeling
  + Indicator and construct are not the same thing
    - For example: belief of happiness is a construct. But level of happiness is indicator
* **Some major type of validity**
  + Content-based validity
    - Question:
      * Are the observable indicators of the construct (domains/sub-domains, items, tasks) defensible and consistent with what is known about the construct based on existing knowledge, theory or expert judgment about the construct?
    - Content relevance and content repetitiveness
    - When: Soon after design/selection of instruments, items or tasks
    - Data (Qualitative or quantitative):
      * Comments of expert (3-6 people) judges using structured reviews
      * Surveys
      * Checklists
  + Correlational evidence: Convergent and discriminant validity
    - Question:
      * Do the scores of my instrument correlate as expected (e.g., positively) with scores of other instruments tapping the **same or similar** constructs, as predicted by theory or the literature?
      * Do they show no relationship or correlate negatively with scores of other instruments tapping dissimilar constructs, as **predicted** by theory?
    - When: After an instrument has been content-validated and revised. First iteration/revision should be ideally completed
    - Data (Qualitative or quantitative):
      * Correlation matrix
      * inter-factor correlations
      * Multi-Trait Multi-Method (MTMM)
  + Internal structure and dimensionality of scale
    - Does an analysis of the person x item data matrix suggest one, two or more mathematically derived dimensions that are consistent with the theory about the construct?
    - After an instrument has been content-validated and revised. First iteration should be ideally completed
    - Data:
      * Factor Analysis (EFA, CFA)
      * Item Response Theory (IRT) models
  + Criterion-related validity
    - Question:
      * Do the scores predict some future behaviors in target population? 🡺 predictive validity
      * Do the scores relate to current behaviors in the target population? 🡺 concurrent/convergent validity
    - When: After content validation, internal factor structure evaluations, and convergent and discriminant validity tests are completed
    - Data:
      * Longitudinal (predictive validity)
      * Cross-sectional (concurrent validity)
  + Validity based on group difference
  + Validity based on item quality

**Week 5**

* Validity plan example (Chatterji & Lin, 2018)
  + Constructs: M-SE, M-SC, and M-ANX
  + Validity:
    - Internal structure, convergent validity (correlation between the similar constructs and miscorrelation with the dissimilar constructs), and reliability
    - Directional relationship with each other and with achievements (Nomological network of expected relationship)
    - Replicated and cross-validated evidence in different samples
* Put a few reverse-orientated items but not too much
  + get the information about whether the participants carefully attend the survey
  + do not bring too much cognitive load into the testing, which decrease the reliability.
* Validity vs. Reliability:
  + validity is the systematic error while the reliability is the random error
  + reliability investigates the internal correlation. Validity investigates the external correlation.

****

PS-TP1: Conative (part 1 & part 2)

PS-TP2: Affective

PS-EOY: Affective (part 1) + Conative (part 2)

* condition + behavior + content (double check each indicator)