Program/Policy Evaluation and Assessment

Measurement for Program Evaluation and Performance Monitoring

Nhial T. Tutlam, PhD, MPH
September 27, 2018

Reliability and Validity of Measures

- Constructs
  - Words or phrases that convey meanings of constituents of logic model

- Measurement
  - Translation of construct observable, countable variables

Reliability of Measures

- What happens to constructs?
- Reliability
  - Are measurement results repeatable?
- About reducing random error in your study
- Index of random measurement error

Analytic Techniques

- Cronbach’s alpha (continuous data)
  - Statistic based on the extent to which the responses to closed-ended survey items correlate with each other, taking into account the number of items being assessed for their collective reliability

- Cohen’s Kappa (nominal data)
  - Measures inter-rater agreement

Reliability

- Study reliability: If a study is repeated under identical conditions, will you get the same result?
  - Precision of estimates of association
  - Sampling Error (partially accounted for in analysis with your p-value)
- Measurement reliability: How much error is in your individual measurements?
  - Precision of measures
  - INTER-individual variation

Measurement reliability

- Degree to which you can depend on your measurements

- Depends on how well you can discriminate between subjects
  - Overall variability between subjects

  - Error in the instrumentation

Measurement reliability

- If your population is homogenous, it's difficult to detect a difference between people. If your population is heterogeneous, it's easier to tell people apart.

- Examples
  - Adult vs. child height

  - population IQ vs. IQ among cognitively disabled

Reliability

$$\text{Reliability} = \frac{\text{True Variability of Observations}}{\text{True Variability of Observations + Random Error}}$$

- Range 0 – 1
  - 1 if all variability is due to true differences between subjects
  - 0 if all variability is due to random error
- So what is ideal?
- And what do we mean by true variability of observations?

Types of Reliability

- Inter-Rater or Inter-Observer Reliability
- Test-Retest or Stability Reliability
- Parallel-Forms Reliability or Split-half reliability
- Internal Consistency Reliability

Inter-Rater Reliability

- Interviewer 1 and Interviewer 2 both take a bp measure on a study participant

- So we’re “calibrating” the interviewers

- Assumptions: The interviewers “see” bp the same way (this may be harder for less concrete assessments)

Test–Retest Reliability

- Interviewer 1 takes a participant's bp at 10 AM and again at 4 pm

- We assess the consistency of measure over time.
    - But how much time? Too much time and things may have changed. Not enough and people might remember their answers

- Assumptions: No secular changes in the measure

Parallel Forms Reliability

- We’re exploring driving habits via questionnaire

- Let’s pretend we have two alternate universes

- We give you the exact same questionnaire in each of the two universes and see if the two questionnaires match exactly

Parallel Forms Reliability

- In real life, we don’t have two universes, AND people are never exactly the same
- One way of coming close to this would be to give the questionnaire on Monday and then to the same person on Tuesday. How well do you think that would work?
- Another way would be to include multiple questions about one aspect of driving habits, say road rage, interspersed throughout the questionnaire.

Internal Consistency Reliability

- What we just described has evolved into internal consistency reliability

- Think of this one as the “practical” way of using multiple measures to try to get at one theoretical concept.

- If we consistently get items that classify people as having road rage or not, we have a reliable measure

Measurement Validity

- Validity
  - Degree to which an instrument measures what it is supposed to measure
  - Degree of concordance between measure and underlying theoretical variable

Four Kinds of Validity in Research Design

- **Internal Validity:** the extent to which extraneous variables have been controlled for (change in the outcome might not be because of the experiment)

- **External Validity**: the extent to which the interaction effect of a treatment and other external effects have been controlled (the change was not only because of the treatment but also some other factors facilitated the treatment)

Four Kinds of Validity in Research Design

- **Construct validity**: extent to which variables used to measure construct convincingly represent the construct in logic model

- **Statistical Conclusion Validity**: extent to which we are confident that statistical requirements have been met to conclude existence and strength of association between dependent and independent variables

Biases/Threats to Internal Validity

- There are generally 8 common threats (biases) to internal validity:

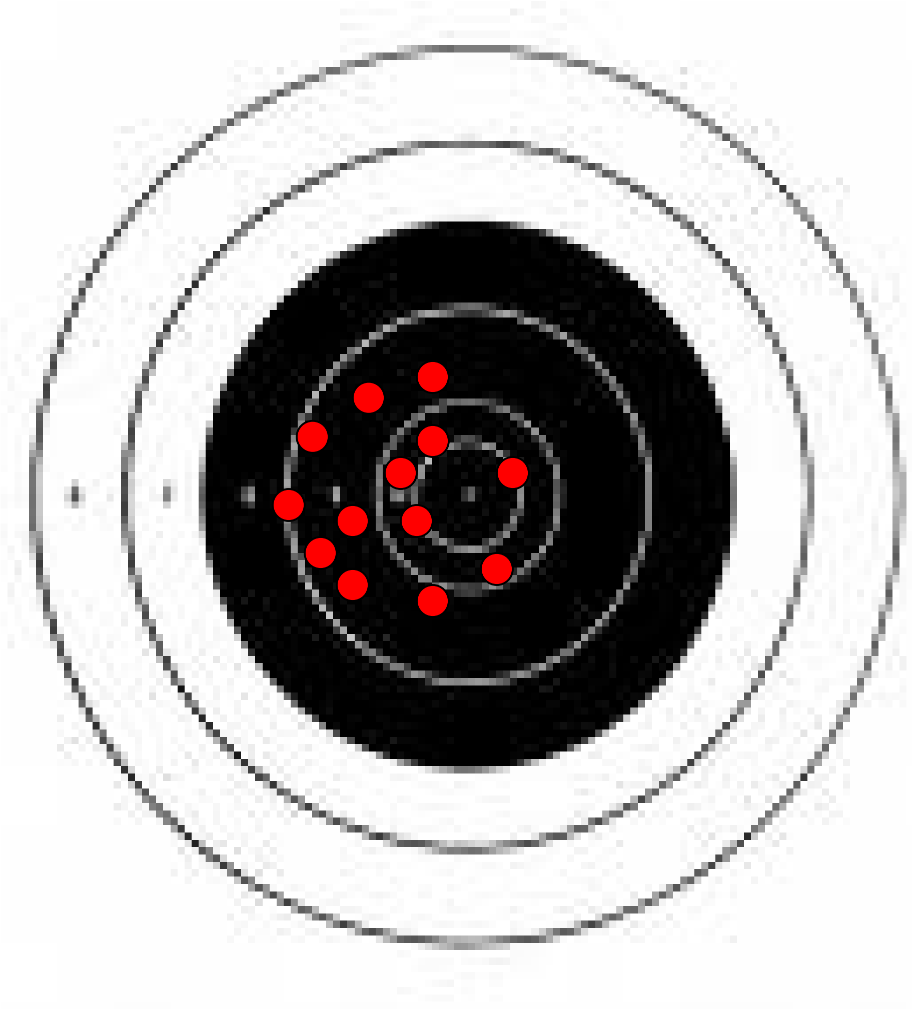
1. History
2. Maturation
3. Testing/Sensitizing
4. Instrumentation
5. Statistical Regression
6. Selection
7. Attrition/Mortality
8. Interactive Effect

Biases/Threats to External Validity

1. *Interaction between the causal results of a policy or program and the participants*
2. *Interaction between causal results of a policy or program and the treatment variation*
3. *Interaction between causal results of a policy and outcome variations*
4. *Interaction between the causal results of a policy or program and the setting*
5. *Context-dependent mediation*

Bias/precision (aka validity/reliability)

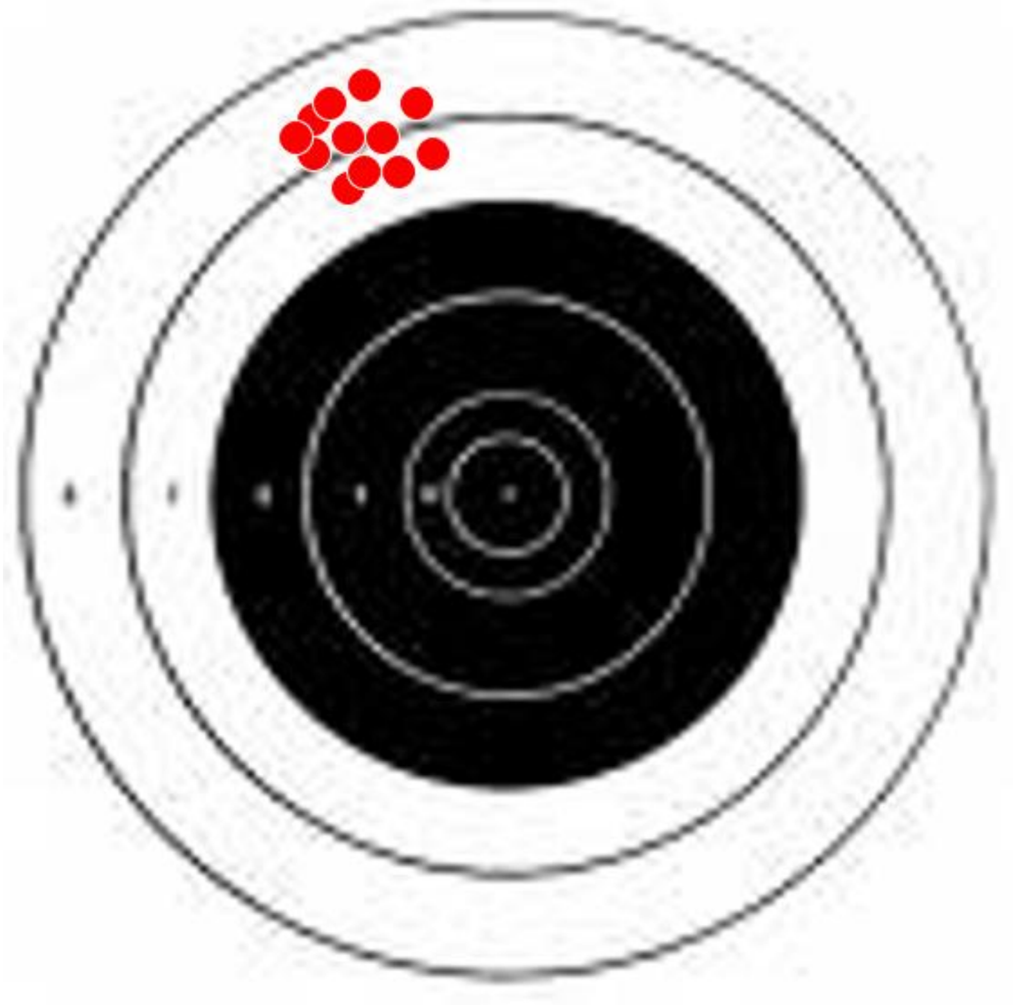
- Low bias, low precision
- Pretty valid, but not very reliable



Bias/precision (aka validity/reliability)

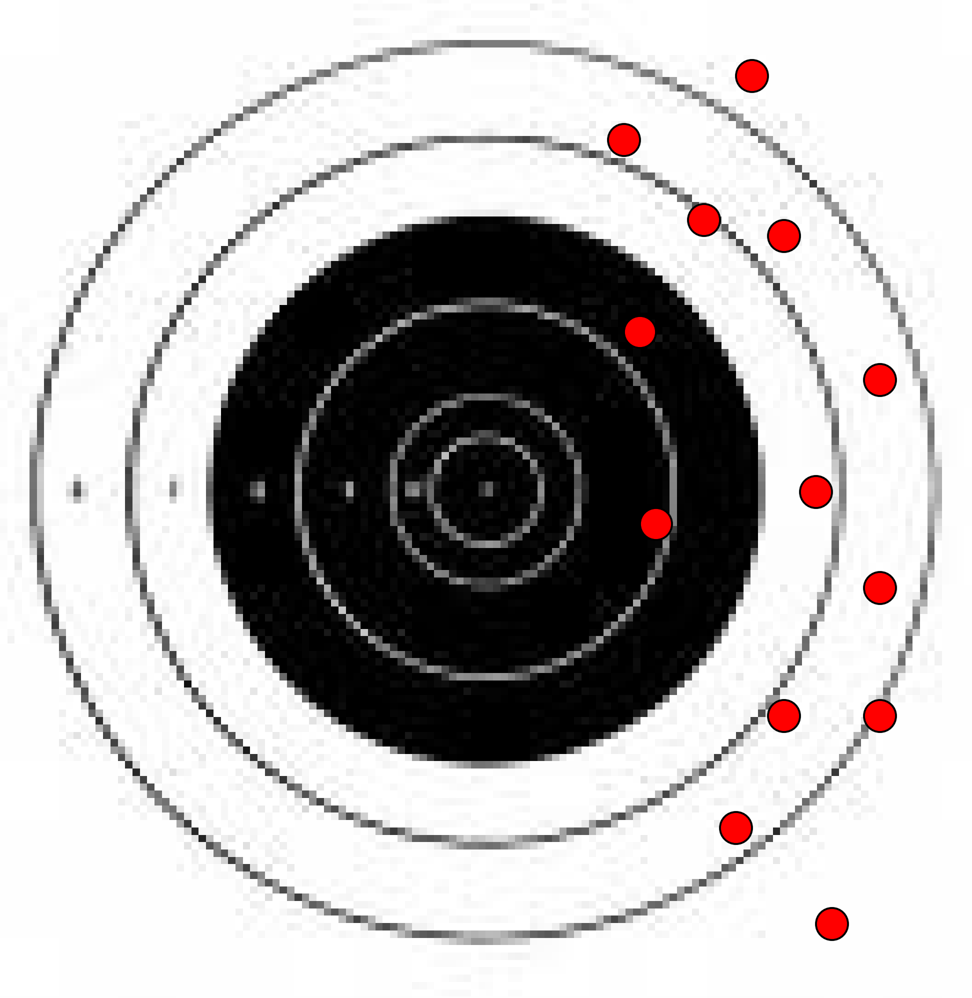
- Low bias, High precision

- Not very valid, but very reliable



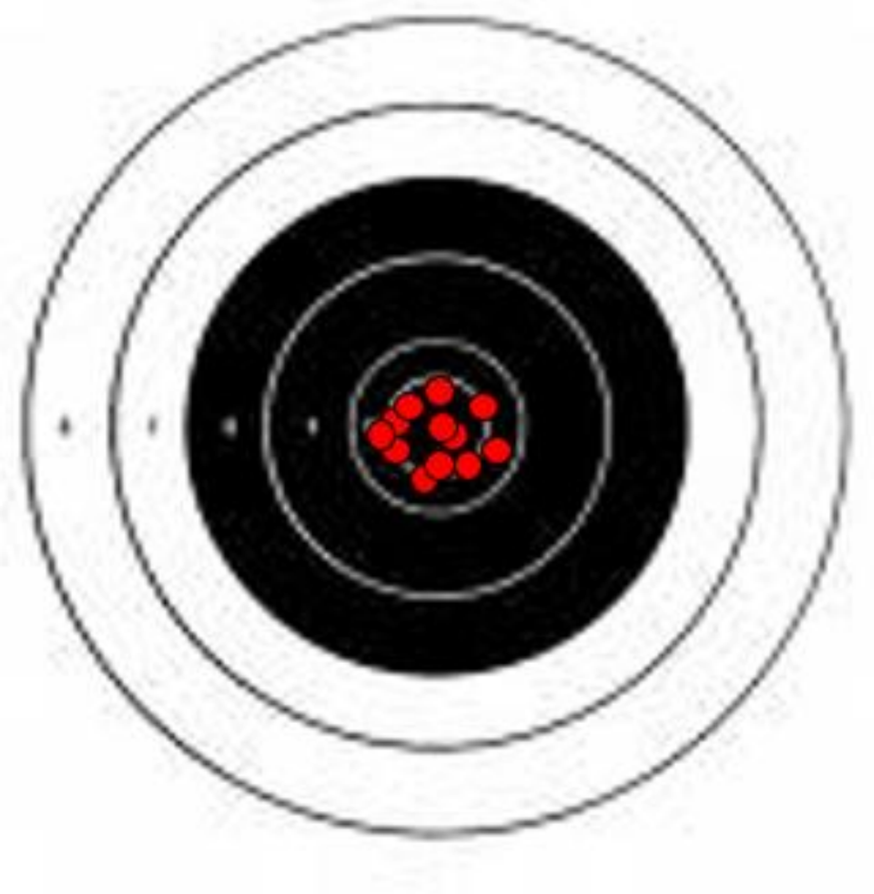
Bias/precision (aka validity/reliability)

- High bias, low precision
- Not valid or reliable
- Worst possible outcome!



Bias/precision
(aka validity/reliability)

- Low bias, high precision
- Pretty valid and reliable
- Best possible outcome!



Types of Validity

- One Measure of One Construct
  - **Face Validity**
  - **Content Validity**
  - Response Validity

- Multiple Measures of One Construct
  - Internal Structure Validity

- Multiple Measures of Multiple Constructs
  - Concurrent Validity
  - Predictive Validity
  - Convergent Validity
  - Discriminant/Divergent Validity

Types of Validity

- Face Validity
- Content validity
- Criterion/concurrent validity
- Construct validity

Face Validity

- On the face of it, does this measure/study appear to be valid?
- Often not based on empirical evidence
- “Lowest” form of validity
- Respondents will be more likely to accept an instrument with face validity
- Example: Measuring the flu season by school absenteeism

Content Validity

- Evaluates whether the measure incorporates all aspects of the construct
- Has “foundational understanding” of the construct
- Assessed by “experts” in the field

Criterion Validity

- Correlation with some other measure of the trait under study, ideally the “gold standard”
- Three types, depending upon temporality of criterion:

Criterion Validity

- **Concurrent Validity**
  - Check if a patient has the flu by administering a questionnaire about symptoms and taking a nasal swab at the time
- **Predictive Validity**
  - Good diet and exercise habits in 20’s predicting heart disease status in 40’s
- **Postdictive Validity**
  - Job skill level as measure of educational attainment

Evaluating Criterion Validity

- Can use statistical tests to evaluate!

- For binary outcomes
  - Sensitivity & Specificity
  - Phi Coefficient (similar to a correlation coefficient)
  - Think 2X2 tables

- For continuous outcomes
  - 1 sample t-test (parametric) or sign-rank test (nonparametric)
  - Correlations – Pearson (parametric) or Spearman (nonparametric)

Construct Validity

- Most difficult type of validity to understand conceptually
- Also takes the most work to implement
- The basic idea is that we validate our construct by seeing how well it corresponds to theoretically related concepts

Construct Validity

- Assesses theory and empirical data at the same time
- Attempts to link the attribute you are measuring to some other attribute by a hypothesis or construct (discriminant or convergent validity)
- Explore the difference between 2 or more populations who would be expected to have differing amounts of the property assessed by your instrument

Construct Validity

- Test the hypothetical construct by appropriately analyzing your instrument

- If expected relationship is found, then hypothesis and measure are sound; if no relationship found, there could be fault in measure or hypothesis

- Approach is non-specific, so evidence should be accrued in several experiments, not just one

Assessing Construct Validity

- Step 1: Go to the literature
  - Decide which measures should be associated (convergence)
  - And which measures should NOT be associated (divergence)

- Step 2: Calculate measures of association between your variable and the other variables you hypothesized would/would not be associated
  - Do this for several variables to accrue evidence of construct validity for your overall tool/study

Examples of Construct Validity

- **Convergent Validity**
  - Height and gender
  - Weight and age
  - BMI and incidence of MI
- **Divergent/Discriminant**
  - Always tough, because theoretically a LOT of things are associated
  - Study ID is usually a good variable for divergence
- Measures of association would be the same ones we used for criterion validity

Units of Analysis and Levels of Measurement

- Unit of Analysis
  - The cases that are the main focus of evaluation

- More than one type of unit of analysis
  - Clients
  - Service providers
  - Business persons who hired clients

- Units of analysis have attributes

Levels of Measurement

- **Nominal**
  - Most basic
  - Can have more than two categories
  - categorical with no order
  - Two basic Features:
    - Classification into one and only one category
    - All observations/responses must fit into existing categories

Levels of Measurement

- **Ordinal**
  - Incorporates features of nominal
  - Retains one basic feature of nominal level of measurement:
    - All cases must fit in one and only one category
  - Level of measurement with ordered categories but no defined distance between categories
    - Example: Likert scales

Levels of Measurement

- **Interval/Ratio**
  - Most sophisticated; has features of both nominal and ordinal
  - Numbers, differences between intervals are meaningful but no meaningful “zero” (most common are dates or temperature)
  - Numbers for which there’s meaningful zero
  - Parametric statistics

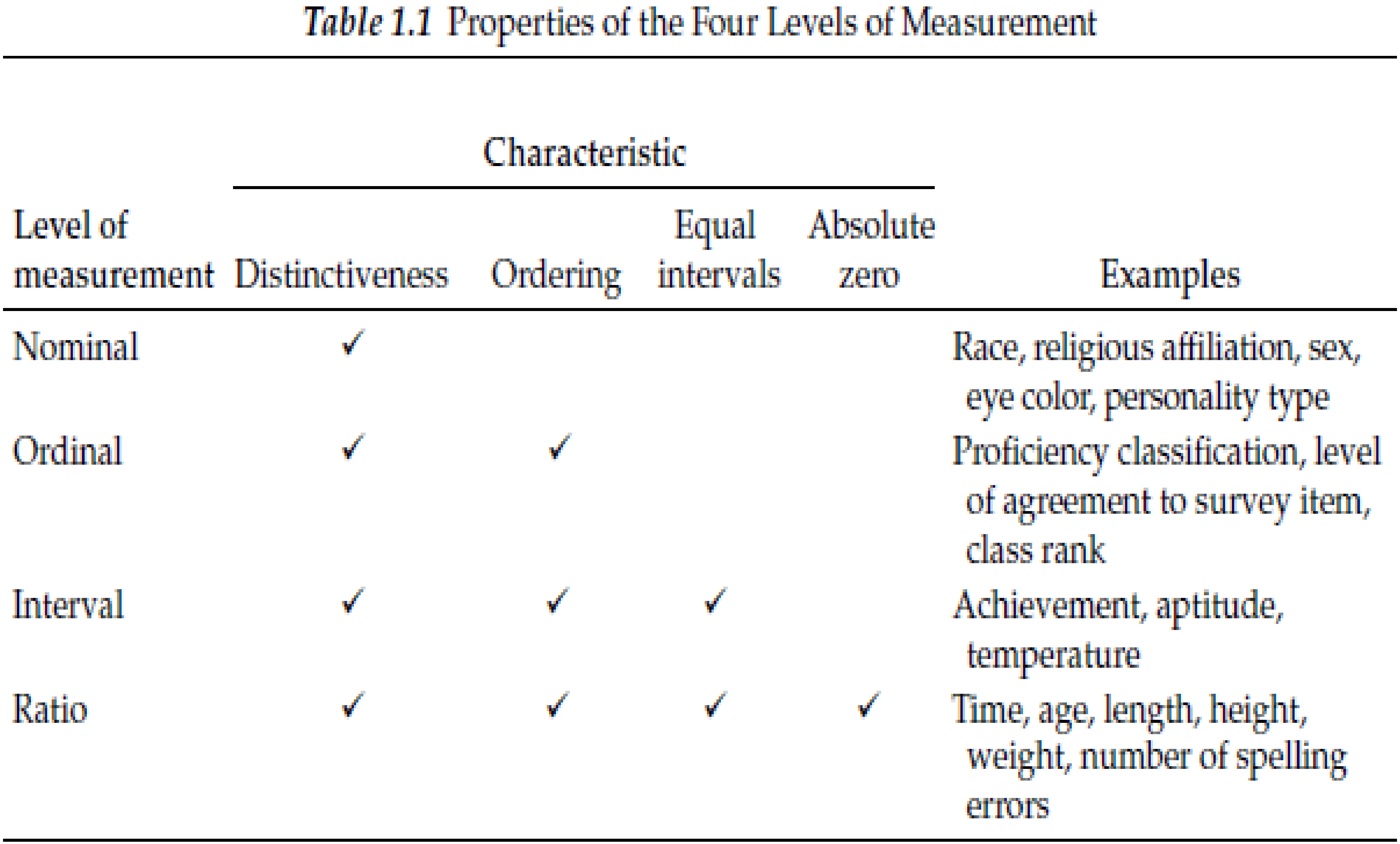


Table 1.1 Properties of the Four Levels of Measurement

	Characteristic				
Level of measurement	Distinctiveness	Ordering	Equal intervals	Absolute zero	Examples
Nominal	✓				Race, religious affiliation, sex, eye color, personality type
Ordinal	✓	✓			Proficiency classification, level of agreement to survey item, class rank
Interval	✓	✓	✓		Achievement, aptitude, temperature
Ratio	✓	✓	✓	✓	Time, age, length, height, weight, number of spelling errors

Sources of Data in Program Evaluation and Performance Measurement Systems

- Existing sources of data
  - Agency records
  - Government databases
  - Client records
- Output measures as proxies for outcomes

Sources of Data in Program Evaluation and Performance Measurement Systems

- Data Collected by program Evaluators
  - Through interaction with program managers
  - Personal experience of the evaluators
- Survey as a Data Source in Evaluation
  - These elicit information from respondents about program

Survey Design

- Likert Statements in Surveys
  1. Ensure the reliability of the construct-specific clusters
  2. Statements should be balanced
  3. Other variations are possible (7 – 9-point scale)
  4. Mix negatively and positively worded statements
  5. Measure individual constructs by clusters of Likert

Survey Design

- Designing and Conducting Surveys
  - In-person interviews
  - Mailed surveys
  - Internet-based surveys
  - Mixed approach surveying

Survey Design

- Specific Survey Design Considerations – Questions
  - Why include a particular question?
  - What is the question intended to measure?
  - How will the survey question help answer the evaluation question that motivated the evaluation?

Survey Design

- Specific Survey Design Considerations – sequence of question types
  - Begin with factual “warm-up” questions
  - Ask about program related experience
  - Assess experience about each phase of the process; rating of the experience
  - Assessment of overall program

Validity: Bias	Source of the Problem	Reliability: Random Error
Race, gender, appearance, interjections, interviewer reactions to responses	Interviewer (face-to-face, telephone)	Inconsistency in the way questions are worded/spoken
Age, gender, physical or psychological handicaps, suspicion, **social desirability, theory of change**	Respondent	Wandering attention
Biased questions, response set, question order, unbalanced Likert statements	Instrument	Single measures to measure client perceptions of the program
Privacy, confidentiality, anonymity	Surveying situation/ survey medium	Noise, interruption
Biased coding, biased categories (particularly for qualitative data)	Data processing	Coding errors, intercoder reliability problems

Using Surveys to Estimate the Incremental Effects of Programs

- Important tips in questionnaire design
  1. Answer every question after drafting instrument
  2. Respondent will use instrument to make sense of the questions
  3. Consult models of well-formed questions
  4. Pilot questions

Using Surveys to Estimate the Incremental Effects of Programs

- Important tips in questionnaire design

5. Familiarize yourself with basics on how people recall and report events

6. Give respondents enough time and be prepared to remind them accuracy is important

7. Consider using events calendars

8. Train interviewers to know intended meanings of questions

Survey Designs Are Not Research Designs

Survey Design	Research Design
‣ Measuring instruments ‣ Ways to measure constructs ‣ Used in wide variety of research designs	‣ Broader and include several ways of measuring constructs ‣ Focus on comparisons needed to determine whether program caused observed outcomes

Validity of Measures and Validity of Causes and Effects

- Validity of measures is part of establishing construct validity of an evaluation research design

- Validity of cause and effect focus on the combination of statistical conclusions and the internal validity of a research design

Discussion Questions

- Design four Likert scale items (with five points ranging from “strongly disagree” to “strongly agree”) that collectively measure patron satisfaction with restaurant dining experiences. Discuss the face validity and content validity of these measures.