Quality of Measurement

Philip N. Chase

Simmons University

Do you **trust** the data?

leefloyd.files.wordpress.com

- **Validity**-the extent to which the data reflect the **nature** of the behavior being measured
- **Accuracy**-the extent to which the data have reduced “error in measurement”
- **Reliability**-the extent to which the measure is consistent over time-”consistency”

Validity

- There are many kinds of validity:
  - Content-does the measure cover the content and nothing but the content
  - Predictive-does the measure predict what it is suppose to predict
  - Concurrent-Does the measure co-vary with another measure that it is suppose to co-vary with.
  - Social Validity-Do the stakeholders agree that the measure is considered important, inclusive, etc.
- Each kind suggests that the measure represents what it is purported to represent.

Predictive Validity

www.westga.edu

- Examples:
  - Do paediatrician tests of autism **predict** how well the child will learn to talk?
  - Does our assessment of performance during training allow us to **predict** whether staff will be able to adapt to changing conditions on the job?

Social Validity

- Does our definition of excellent fourth grade math performance correspond with the standards set by the government?
- Does our definition of disruption of the classroom correspond with the teacher’s definition of disruption?

Accuracy

- Accuracy is used here to talk about the extent to which the investigator has reduced error in measurement.
- For practical purposes, the more direct, continuous, and automatic the measurement, the less likely there will be errors.
- For observational measurement, we have to check for agreement between observers to determine accuracy.

Reliability

- Defined as the extent or degree to which a measure is consistent over time.
- Examples:
  - A ruler is very reliable for measuring centimeters-each time you use it you get the same number of centimeters
  - A beaker is very reliable for measuring liquid volume, each time you read the liquid up to a certain line on the beaker it gives you the same volume.
  - Number of students who receive a B or better on a multiple-choice quiz.
  - A thermostat.
  - Questions answered per minute on quiz.

Behavioral Reliability

- Defining behavior operationally as we discussed earlier is the number one means for improving reliability.

- A checklist improves the reliability of measuring a complex performance like writing an essay or making a swan dive.
- A checklist directs the observer to look for particular elements

- Makes it more likely the observer (or multiple observers) will rate the performance similarly.

View of a woman (Model Beverly Stevenson) in mid-air doing a swan-dive, ca. 1940s. Photograph by Philip Gendreau

Reliability and Validity

Reliability

- Does the value observed and recorded reflect consistency?
- Test by measuring the object under study multiple times:
  - Test/retest
  - Split halves
  - IOA

Validity

- Does the value observed and recorded reflect the behavior and dimension of interest for the behavior?
- Test by comparing with other data, expert judgments, stakeholder responses.

Validity, Accuracy, & Reliability

- I think of validity as the most important aspect of trustworthiness.
- If you measuring something different than what you think you are measuring, it does not matter whether it is accurate or reliable.
- You think you are measuring intelligence, but you are measuring only experience with what privileged people know-cultural bias!

Three Simple Truths:

- In order for a measure to be valid it must be reliable.
- In order for a measure to be reliable it must be accurate
- And in order for a measure to be accurate we must have agreement

Teams-Write it down

- State the definition of validity.
- State the definition of accuracy.
- State the definition of reliability.
- Try to think of one example for which my three basic truths do not hold:
  - In order for a measure to be valid it must be reliable.
  - In order for a measure to be reliable it must be accurate
  - And in order for a measure to be accurate we must have agreement

Threats to how much we trust the data.

www.pbase.com

Why do reliability, accuracy, and validity matter?

- We use data to make decisions and solve problems
  - Data is gathered through measurement procedures
  - The data only have meaning if they measure what they are supposed to measure (valid) and do so with accuracy and consistency (reliability).
- Evaluating whether data are accurate, reliable and valid is a key element in applying research findings.

First threat:Indirect measurement

- The extent to which descriptions have been inferred from the data.
- Describing what is actually measured is direct:
    - If you are interested in what a person will choose, giving them a choice is more direct than asking them what they like.
    - If you are interested in how well a person solves math problems, describing their math skills is more direct than describing their math ability
    - If you are interested in how many miles a person walks in a week, using a Fitbit is more direct then asking the person to report the number of miles per week.

Second Threat: Measuring the wrong dimension of behavior

- Investigators need to be careful to select the dimension that matches the interest:
    - If you are interested in how long a behavior lasts, you should **not** measure frequency.
    - If you are interested in how often a behavior occurs, you should **not** measure duration
    - If you are interested in how quickly the behavior occurs after a stimulus is presented, you should **not** measure percent correct.

Third Threat: Measurement artifacts

- Under or over estimations of the behavior
- Scheduling measurement at non-representative times may provide under or over estimations of behavior
- Using insensitive measures, like examining too few occurrences, may produce measurement artifacts

Fourth Threat: Poorly design Measurement Systems

- If a measurement system is tiring it will be produce more errors
- If a measurement system is complex it may produce more errors
- Reduce the complexity:

  The KISS principle.

KISS:

- Keep
- It
- Simple
- Stupid
- (although my mother taught me never to say stupid- so how about Keep It Simple , but Sophisticated).

Fifth Threat: Inadequate Observer Training

- Probably the most important is to make sure the observers are well trained on the behavioral definitions and measurement systems being used.
    - Select compulsive observers
    - Instruct them thoroughly
    - Have them practice
    - Give them feedback
    - Have them re-practice
    - Test them under realistic situations
    - Check from time to time to eliminate observer drift.

Teams-Write it down…

- State the five threats to measurement.
- State how one minimizes each of these threats.

Now we turn to agreement

- The official term for agreement in applied behavior analysis is interobserver agreement or IOA.
- IOA is the extent to which
  - two or more independent observers
  - report the same observed values
  - after measuring the same event
  - using the same measurement system.
- Typically reported in terms of Percent Agreement

Two basic techniques

- Total or Overall IOA
  - Smaller count/Larger count, shorter duration/longer duration X 100.
- Interval by Interval IOA
  - Number of intervals agreed/Total number of intervals X 100

Calculating Total Agreement

IOA charts adapted from: http://homepages.wmich.edu/~malavosi

OBSERVER 1

| | | | | | | | | |

| | | | | | | | | |

| | | | | | | | | |

| | | | | | | | | |

| | | | | | | | | |

50

OBSERVER 2

| | | | | | | | | |

| | | | | | | | | |

| | | | | | | | | |

| | | | | | | | | |

| | | | |

45

IOA ? **45/50 = 90%**

Interval-by-Interval Agreement

Scored (X)

Unscored(0)

4 agreements/7 total X 100=
57% agreement

O	O	X	X	X	O	X
X	O	O	X	O	O	X

When would you use each of these types of IOA?

- Overall
- Interval by Interval

Overall

- Overall is used with event recording, duration recording, frequency recording, and latency recording.
- One has to be careful because there is no guarantee that the observers are measuring the same instances of behaviors.
- This can be improved by having shorter time periods of measurement that are compared.

Interval by Interval

- Used when specific records can be compared across intervals
- Partial interval recording, whole interval recording, or momentary time sampling

As a whole class:

Using the interval data for touching and vocalizing below (adapted from Martin & Pear), calculate the interval-by-interval IOA.

Observer 1:

Ten second Intervals

	1	2	3	4	5
Touching:	+	0	+	+	0
Vocalizing:	+	+	0	+	+

Observer 2:

	1	2	3	4	5
Touching:	+	+	0	+	0
Vocalizing:	+	+	0	+	0

Teams: IOA From Video

- Calculate IOA on Data from Laughing Baby

YouTube Link: Baby laughing?

  - ***https://www.youtube.com/watch?v=RP4abiHdQpc***

- Define laughing using rate (see last lesson)
- Calculate percent agreement for first minute of video
- Which IOA will be the best for this behavior?
- Use this type of IOA to calculate agreement.