Reliability and Validity
研究工具的信、效度

張秀如老師

護理系

hsiuju@tmu.edu.tw

校內分機6309

學習目標

- 了解決定研究工具信度的方法
- 了解影響研究工具信度的因素
- 了解效度的類型

信度的概念

- 研究工具具有一致性與穩定性的程度。
- Moser & Kalton：一份量表或測驗具有信度，是指其在**不變的條件**下，重複施以**相同的測量**，而能獲致**相同的結果**。
- 以下列兩方面來看信度的概念：
  - 一項工具的可靠度如何？(一致性的程度)
  - 其不可靠的程度又如何？(不一致的程度)

決定研究工具信度的方法

- 外部一致性程序
    - 測試／再測試
    - 同樣形式的相同測試
- 內部一致性程序
    - 折半技術

Stability（穩定性）

- 它所關心的是重覆測量的一致性，一般指的是再測信度(test-retest reliability)。

- Relatively endurable traits such as personality, ability for which a test-retest approach is suitable

再測信度(Test-retest reliability)

- 有時再測信度在研究的應用不太理想
  - many traits of interest do change over time, independently of the stability of the measure.

    (moods, physical condition)

  - memory interference
- 再測時間的長短

Fictitious Data for Test-retest Reliability of Self-Esteem Scale

Subject Number	Time 1	Time 2	
1	55	57	
2	49	46	
3	78	74	
4	37	35	
5	44	46	
6	50	56	
7	58	55	
8	62	66	
9	48	50	
10	67	63	r = .95

Equivalence（對等性）

- 它所注意的是比較兩個或以上的觀察者對同一事件的測量，指的是評量者間一致性信度(inter-rater reliability)，常用在觀察性研究。

- The accuracy of observer ratings and classification can be enhanced by careful training, the development of clearly defined categories

Equivalence（對等性）

- Alternative forms
    - 例如考駕駛執照的筆試有不同套試題，每套試題的困難程度必須相同
    - 兩組題目給同一人試測，看兩組題目得分的相關係數

Homogeneity（同質性）

- 測量工具的內在一致性
- 較早以前，測試工具中不同項目間的同質性可用折半信度(split-half reliability)，類似再測信度。

Fictitious Data for Split-Half Reliability of the Self-Esteem Scale

Subject Number	Total Score	Odd-Numbers Score	Even-numbers Score	
1	55	28	27	
2	49	26	23	
3	78	36	42	
4	37	18	19	
5	44	23	21	
6	50	30	20	
7	58	30	28	
8	62	33	29	
9	48	23	25	
10	67	28	39	r =.80

Homogeneity（同質性）

- 目前，Cronbach's alpha coefficient(α係數)較常被用。
    - It gives an estimate of the split-half correlation for all possible ways of dividing the measures into two halves

Interpretation of Reliability Coefficients

- A measure that is unreliable interferes with an adequate testing of a researcher's hypotheses. If data fail to confirm a research prediction, one possibility is that the measuring tools were unreliable

- What an acceptable reliability coefficient should be?

Interpretation of Reliability Coefficients

- Reliability is the proportion of true variability to the total obtained variability. For example, the reliability coefficient were.85, then 85% of the variability in obtained scores could be said to represent true individual differences, and 15% of the variability would reflect random, extraneous fluctuations.

影響研究工具信度的因素

- 題目的用詞
- 訪談情境
- 受訪者的情緒
- 互動的本質
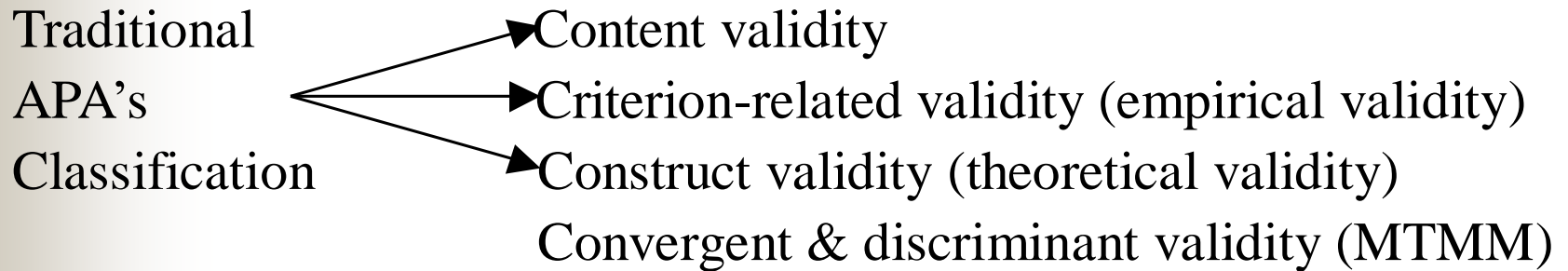- 工具的迴歸效應

Definition of Validity

- **the degree to which an instrument measures what it is supposed to be measuring.**

Validity vs. Reliability

- **low reliability**
- **high reliability**

Types of validity

Traditional

APA's

Classification

Content validity

Criterion-related validity (empirical validity)

Construct validity (theoretical validity)

Convergent & discriminant validity (MTMM)

Content Validity

- the adequacy with which a specified domain of content is sampled

- the degree to which the test contains an adequate amount of items from all aspects of the dimension

- this is probably the most important type of validity for classroom tests

- difficult to evaluate

Content Validity (c'd)

- **face validity: Dose the test appear to measure the trait of interest? The lay person's acceptance that an instrument appears to be relevant. A test need not have face validity to be valid.**

- **test blueprint: careful definition of the domain of behaviors to be measured by a test and a specification of the relative importance of each attribute**

Content Validity (c'd)

- **expert's judgment: if a committee of experts in the field agree that these items adequately tap the entire domain in the right breadth and proportions**

Index of Content Validity (CVI) (Waltz & bausell, 1981)

- **the rating of the content relevance of the items on an instrument using a 4-point ordinal rating scale, where 1 connotes an irrelevant item and 4 an extremely relevant item**

- **CVI is the proportion of items that received a rating of 3 or 4 by the experts**

Index of Content Validity (CVI)

專家使用四分量表評估item的relevance(相關性)

1- an irrelevant item,

2-unable to assess relevance without item revision

3-relevant but needs minor alteration

4 - extremely relevant item

Steps of content validation

- **Theoretical work (concept's domain)**
- **Stratify the domain into major facets**
- **Write items (building a large item pool)**
- **Pretest or content validation:**

(1) collect data and do a criterion-related validation or construct validation;

(2) use theory to examine the meaning of each item

Index of Content Validity (c'd)

- **number of experts: a minimum of five experts; the maximum is unlikely to exceed 10**

- **how many experts should agree on the item?**

- **the 4-point scale is preferable because it does not include the middle rating**

Criterion Validity

- a measure of the extent to which a particular test is related to an external criterion
- the degree to which the test corresponds to an accurate criterion of the construct
- predictive validity
- concurrent validity

Predictive Validity

- **the degree to which the test is able to predict (or forecast) a future criterion**

- **the external criterion is future performance at the task of interest**

Predictive Validity

- **how well does the test predict success and failure in the task it is being used to predict?**

- **the validity coefficient is the correlation between scores on the test and performance scores on the criterion**

Concurrent Validity

- the degree to which the test corresponds to a present time criterion

- the external criterion is simultaneous performance at some related task of interest

- difference between predictive and concurrent validity: the time of the second test

Concurrent Validity

- a substitute for predictive validity
- objective: diagnosis of existing status
- Is Joe schizophrenic? Vs. Is Joe likely to become schizophrenic?

Limitations of criterion validation:

- 1.Criterion validity is not only influenced by the true relationship between the measure and the construct, but also by the relationship between the criterion and the construct. The correlation could change due to factors not related to the validity.

- 2.For many measures, it is hard to find appropriate criteria.

Problems with Criterion Validity

- **difficulty in identifying a meaningful criterion: immediate vs. ultimate**

- **criterion contamination: test scores themselves influence an individual's criterion status**

Correct for Unreliability in Predictor & Criterion

$$\frac{r_{xy}}{r_{TxTy} = \sqrt{r_{xx'}\, r_{yy'}}}$$

- $r_{TxTy} = \dfrac{r_{xy}}{\sqrt{r_{xx'}\, r_{yy'}}}$
- if the reliability is very small (<.65), the correction should not be applied

Construct Validity

- **the degree to which the test corresponds to other assessments of the construct**

- **to the extent that a variable is abstract rather than concrete, we speak of it as being a construct**

- **not being directly observable**

- **examples: empathy, stress, personality, depression, quality of life**

Defines a construct on Two Levels

- **operational definition: a formal definition of the attributes comprising that construct, including the procedure used to measure them**

- **syntactic definition: postulation of the specific relationships between measures of the construct and certain measures**

Process to Verify a Construct Validity

- **formulate one or more hypotheses about the relationships between the construct and other construct (criteria)**

- **select (or develop) a measurement instrument which consists of items representing behaviors that are specific and concrete**

Process to Verify a Construct Validity (c'd)

- **gather empirical data which will permit the hypothesized relationship to be tested**

- **determine if the data are consistent with the hypotheses and explain the findings**

Evidence Needed for construct Validity

- correlations between construct and other variables

- differentiation between groups

- factor analysis

- multitrait-multimethod validity

Convergent Validity

- the extent to which a new test adheres to other, related indicators of the construct that the new test is designed to measure (this may be other tests designed to tap the same construct or tests of related constructs); *Ideally over .3 but not too high!*

Divergent/Discriminant Validity

- the extent to which a new test does not relate to indicators of different constructs which should not be associated with the intended construct; *Ideally under .3*

Summary

- 研究工具的信、效度牽涉資料的精準性，唯有可信且具有效度的研究工具方能測得誤差小且精準的資料，使資料能正確的呈現，以助研究者詮釋資料、解釋研究發現，而後完成研究主旨。

參考資料

- 潘中道、黃瑋瑩、胡龍騰合譯(2000)‧研究方法步驟化學習指南‧台北：學富文化。
- 徐南麗、王如華、黃璉華等合著(2003)‧護理研究導論：第二版‧台北：偉華。