

Classical Test Theory

Validity issues

Christophe Lalanne
ch.lalanne@gmail.com

November, 2009

Summary

“ abstract here. . . ”

Outline

1. *content validity*: Delphi's method, CVR, agreement
2. *construct validity*: CFA, MTS
3. *criterion validity*: ROC analysis, subgroup analysis
4. *concourant validity*: SEM, MTMM models
5. *cross-cultural issues*: multi-group CFAs, MIMIC model

*Various functions used throughout this chapter were collated in the package **Psychomisc**.*

Foreword

Validity is probably the most important issue that any researcher has to tackle from the start of his study, even before analysis of scores reliability.

However, for practical purpose, we approach this subject only now because validity issues are also linked to confirmatory analyses which we shall dwell on in the next chapters. Furthermore, there are many more facets related to validity than to measurement properties like reliability.

Following [14, chap. 4],

Validation of instruments is the process of determining whether there are grounds for believing that the instrument measures what it is intended to measure, and that it is useful for its intended purpose.

Nomenclature

As noted by [13, p. 48], several concepts of *validity* have been proposed so far. We shall consider the following definitions:

- *content validity* reflects the adequacy of the domains or dimensions spanned by the items;
- *criterion validity* demonstrates that scales have empirical association with external criteria, such as gold standards or other instruments purported to measure equivalent concepts;
- *construct validity* relates each inter-items and item-scale relationships from a theoretical point of view.

Nomenclature (Con't)

Convergent and discriminant validities are both subsumed in the general and theoretical concept of *construct validity*.

We previously defined *reliability* as the extent to which scores may be reproducible, and *sensitivity* as the ability for a test to detect differences between patients or groups of patients based on prognostic considerations.

Obviously, scores interpretation depends on both validity and reliability characteristics of the questionnaire, but good reliability properties without established validity don't mean anything!

A good starting point is [39].

Content validity

Content validity can be assessed by expert only, and generally this is done before releasing a given questionnaire, i.e. during the elaboration of the items. Afterwards, we merely have to deal with reliability issues based on the analysis of subjects' responses.

Content validity may be evaluated using expert judgments (variance, internal consistency and concordance) e.g. *inter-rater agreement* on the relevance of items.



Content validity should not be confused with *face validity*, which addresses the way a set of items is perceived or accepted by the respondents and has nothing to do with its statistical or content properties.

Why content validity is so important?

Content validity is a very important concept in all leading fields for questionnaire development (e.g. in clinical trials) where focus groups and cognitive pretesting play a particularly important role, see e.g. [27] and [20, pp. 32–34].

For example, a depression scale would lack content validity if it only assesses the affective dimension of depression but fails to take into account the behavioral dimension.

Content validity is supported by evidence from qualitative studies that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept, population, and use.

See FDA guidelines, under section *Drug*→*Guidance* at www.fda.gov.

Some caveats about clinical validity



In clinical setting, the validity of a *medical diagnosis* requires a clear *aetiology*. However, most of functional diagnosis of mental health-related disorders (e.g. personality disorder, schizophrenia) are defined in a circular fashion: The diagnosis is made on the basis of symptoms and the symptoms are accounted for by the diagnosis.

Likewise, although most psychiatrists will agree on what is depression, demonstrating categorically that clinical depression differs from dysphoria or everyday unhappiness is nearly impossible [32].

Finally, the problems of valid and reliable case identification in psychiatric epidemiology remain unsolved because no physical cause are generally associated to the disease under consideration.

Relative vs. absolute judgments

A simple solution would be to ask a clinician, or several clinicians, if a given item is able to cope with the construct under interest, and if so, whether it reflects specific symptoms associated to the depression syndrome.

However, one may wondering whether such absolute judgments are so adequate when in fact it is proved that we are unable to make reliable absolute judgment.

The Delphi method

The Delphi method is a systematic, interactive forecasting method which relies on a panel of experts [15]. The experts answer questionnaires in two or more rounds.

After each round, a facilitator provides an anonymous summary of the experts forecasts from the previous round as well as the reasons they provided for their judgments.

Finally, the process is stopped after a pre-defined stop criterion (e.g. number of rounds, achievement of consensus, stability of results) and the mean or median scores of the final rounds determine the results.

The Delphi method (Con't)

In summary, the main stages of the Delphi's method are [15]:

1. Formation of a team to undertake and monitor a Delphi on a given subject.
2. Selection of one or more panels (experts) to participate in the exercise.
3. Development of the first round Delphi questionnaire
4. Testing the questionnaire for proper wording (e.g., ambiguities, vagueness)
5. Transmission of the first questionnaires to the panelists
6. Analysis of the first round responses
7. Preparation of the second round questionnaires (and possible testing)
8. Transmission of the second round questionnaires to the panelists
9. Analysis of the second round responses (Steps 7 to 9 are reiterated as long as desired or necessary to achieve stability in the results.)
10. Preparation of a report by the analysis team to present the conclusions of the exercise

The Delphi method (Con't)

As can be seen, experts are encouraged to revise their earlier answers in light of the replies of other members of their panel. Indeed, it is believed that during this process the range of the answers will decrease and the group will converge towards the “correct” answer.

Pros	Cons
rapid consensus	depends on the level of expertise of experts
possibility of distant interaction	influence of the formulation
avoid focus group	influence of the mediator

A comprehensive survey is available to download at <http://www.is.njit.edu/pubs/delphibook/>.

Other alternative for qualitative evaluation

Again, the evaluation process relies on a voting procedure between expert in the domain under study [36], although criterion-based verbal agreement is not directly quantified. The idea is to evaluate congruence between item and objective [?].

A possible scoring rule is to consider +1 if item and objective agree, -1 if not, and 0 for uncertain cases. An agreement index, $-1 < I_{ik} < +1$, is then computed as follows:

$$I_{ik} = \frac{(N - 1) \sum_{j=1}^n X_{ijk} + N \sum_{j=1}^n X_{ijk} - \sum_{j=1}^n X_{ijk}}{2(N - 1)n}, \quad (1)$$

for the k th item, i th objective, with N dimensions and n experts.

Other alternative for qualitative evaluation (Con't)

Example:

Given a test of 36 items supposed to tackle 5 dimensions, here are the results observed for item 1 w.r.t. objective 2:

score	N
-1	1
0	1
+1	7

The item/objective agreement index is:

$$I_{ik} = \frac{(5 - 1)6 + 5(6) - 6}{2(5 - 1)} = \frac{48}{72} = 0,67$$

Other alternative for qualitative evaluation (Con't)

Such an index can be used

- as a *relative criteria*, when we want to compare two items one to each other (the closer I_{ik} is to 1, the better it is);
- as an *absolute criteria*, when the value for an item is compared to a reference or expected index (e.g. overall agreement of at least 7 rater out of 9, that is $I_{ref} = 0.78$).

Content validity criterion

Lawshe [24] proposes the Content Validity Criterion (CVR) as an index of validity of a measurement instrument. In this approach, a panel of subject-matter-experts (SMEs) is asked to indicate whether or not an item in a set of other items is “essential” to the operationalization of the theoretical construct. The question is formulated as follows:

Is the skill or knowledge measured by this item ‘essential’, ‘useful, but not essential’, or ‘not necessary’ to the performance of the construct?

According to Lawshe, if more than half the panelists indicate that an item is essential, that item has at least some content validity.

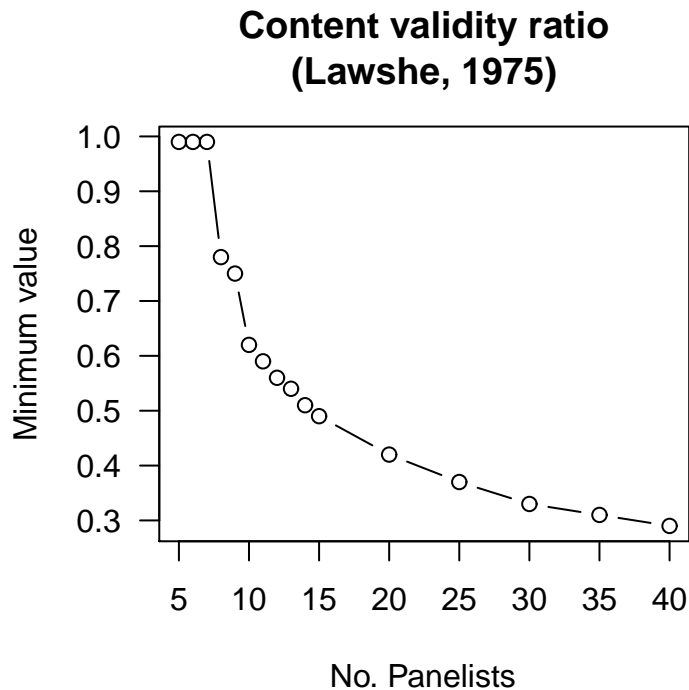
Content validity criterion (Con't)

For each item, the CVR is defined as

$$\text{CVR} = (n_e - N/2)/(N/2) \quad (2)$$

where n_e is the number of SMEs indicating “essential” and N is the total number of SME. Hence, a CVR of 0 means that 50% of the SMEs in the panel of size N believe that an item is “essential”. See [24, p. 568] for a Table of critical one-tailed values at $\alpha = 0.05$ corresponding to the minimum CVR's for different panel sizes.

The mean CVR across items may be used as an indicator of overall test content validity.



Content validity criterion (Con't)

If we ask the SMEs to sort the N items into a set of C *a priori* defined and mutually exclusive measurement scales for different constructs, we can use Cohen's κ to assess the degree of between-expert agreement as to the placement of these measurement items into their measurement scales.

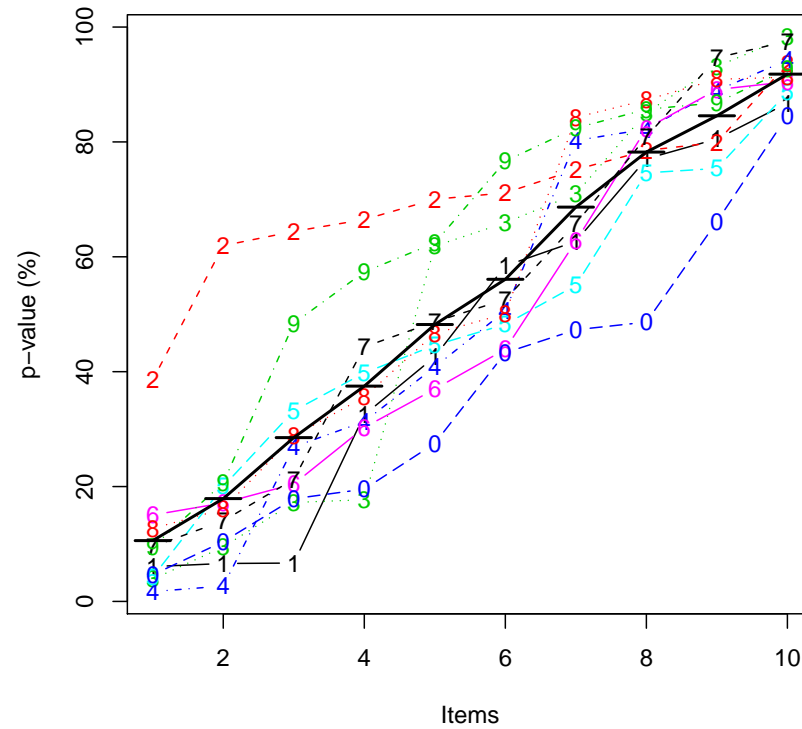
The Angoff's method

The different variations based on Angoff's method [4] are extensions of inter-rater agreement. Such methods aim at evaluating the minimal acceptable level for subjects to be able to answer correctly to an item.

In its simplest version, one ask several experts to decide who among n individuals are at the threshold level. The correponding p -values are added together and determine the *minimum passing level* (MPL).

The different MPL are viewed as minimal scores for a given test, and the average score is considered as the passing score.

The Angoff's method: Illustration



The Angoff's method: Variations

1. *Iterative approach*: the estimation of MPL is repeated several times (in general, 2–3), with interleaved discussion session including a moderator, which results in a small standard error [18];
2. *Normative approach*: normative data are shown before the last round and contribute to improve reliability by increasing the correlation between p -values and candidates' level [7];
3. *Yes/No approach*: although subject to critics, [21] have proposed to restrain the evaluation on one candidat only, hence a binary judgement;
4. *Weighted scores approach*: [18] extended the standard method to multidimensional items scored as polytomous items; experts have to determine the MPL and the weight of each dimension assessed in the test.

A critical review of these different approaches is given in [34].

The correlational approach

Inter-rater reliability for item adequation can be estimated with Kendall's coefficient of concordance, W , which is defined as:

$$W = \frac{\sum_{j=1}^n (R_j - 1/2k(n+1))^2}{1/12 \times k^2(n^3 - n)} \quad (3)$$

where n is the number of items, k the number of raters and R_j the rank sum for each item across raters.

When $n > 7$, $k(n-1)W \sim \chi^2(n-1)$ [38, pp. 269–270]. This asymptotic approximation is valid for moderate value of n and k [22], but with less than 20 items F or permutation tests are more suitable [25].

The correlational approach (Con't)

Kendall's W is an estimate of the variance of the row sums of ranks R_j divided by the maximum possible value the variance can take; this occurs when all variables are in total agreement. Hence $0 \leq W \leq 1$, a value of 1 representing perfect concordance.

- ✎ There is a close relationship between Spearman's ρ and Kendall's W statistic: W can be directly calculated from the mean of the pairwise Spearman correlations (for untied observations only):

$$W = \frac{(k-1)\hat{r}_s + 1}{k}.$$

For two raters, $W = (\rho + 1)/2$, and a permutation test of W for two variables is the exact equivalent of a permutation test of ρ [25].

Alternative testing strategies

The permutation test (raters are the permutation units under H_0) has a correct rate of type I error for all values of k and n . Likewise, the F -statistic

$$F = (m - 1)W / (1 - W)$$

with $\nu_1 = n - 1 - (2/k)$ and $\nu_2 = \nu_1(k - 1)$ degrees of freedom (or its Fisher transformation, $z = 0.5 \log_e(F)$) yields correct inference at the pre-specified α level.

Post-hoc tests, based on partial concordance index [9], might help to highlight a deviant rater (but not a subgroup of raters).

The correlational approach (Con't)



Although we headed this section with “correlational approach”, W is different from Spearman ρ or Kendall’s τ coefficient of rank correlation:

[...] whereas (both) express the degree of association between two variables measured in, or transformed to, ranks, W expresses the degree of association among k such variables, that is, the association between k sets of rankings.

Siegel & Castellan, [38, p. 262]

Illustration

Screening questionnaires

		True diagnosis		Total
		Positive	Negative	
Screening test	Positive	a	b	$a + b$
	Negative	c	d	$c + d$
Total		$a + c$	$b + d$	N

True diagnosis may be outcome of interview, result from biological analyses, etc. We will suppose this is an error-free result.

From this table of counts, we can define four useful measures of screening accuracy and predictive power.

Characteristics of screening “efficacy”

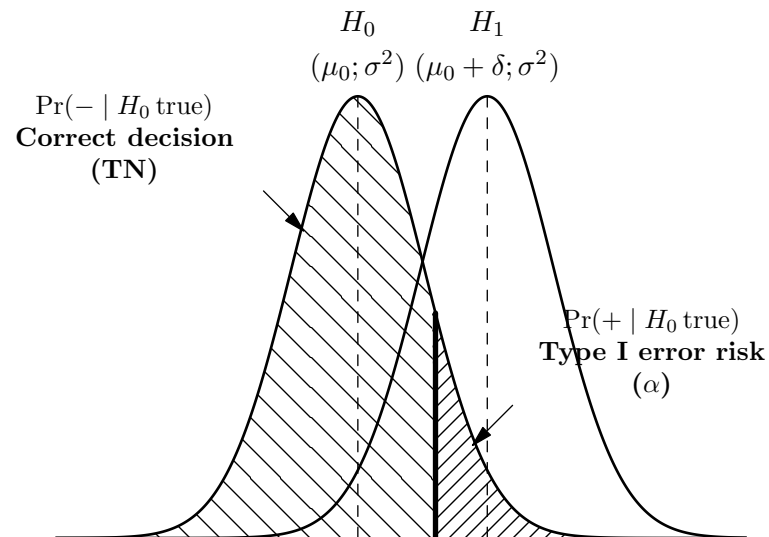
- Sensitivity (se), $a/(a + c)$, i.e. the probability of the screen providing a positive result given that disease is present;
- Specificity (sp), $d/(b + d)$, i.e. the probability of the screen providing a negative result given that disease is absent;
- Positive predictive value (PPV), $a/(a + b)$, i.e. the probability of patients with positive test results who are correctly diagnosed (as positive);
- Negative predictive value (NPV), $d/(c + d)$, i.e. the probability of patients with negative test results who are correctly diagnosed (as negative).

Screening vs. Hypothesis testing

Recasting the preceding Table in terms of the outcomes of a statistical test (but see Chapter 1), we see for instance that $Se = TP / (TP + FN)$, or in other words $Se = \Pr(\text{test } \oplus \mid \text{reality } \oplus)$ whereas $\alpha = \Pr(\text{test } \oplus \mid \text{reality } \ominus)$.

		True diagnosis		
		Positive	Negative	
Screening test	Positive	TP	FP (α)	\leftrightarrow PPV
	Negative	FN (β)	TN	\leftrightarrow NPV
		\updownarrow Se	\updownarrow Sp	

Screening vs. Hypothesis testing (Con't)



As shown above, the outcome of a screening assessment can be seen as a random experimental setting where the null hypothesis, H_0 , means a negative result. Therefore, a correct rejection of H_0 occurs with probability $1 - \alpha$ and may be considered as *test specificity*.

Application

Consider the exemple shown below. These are results for the CAGE questionnaire which was studied by [8], but see [12, pp. 31–32]. This study focused on $N = 518$ patients admitted to the orthopaedic and medical services of a community-based hospital (with a 6-month follow-up).

		Alcohol abuse		
		Positive	Negative	
CAGE	Positive	99	43	142
	Negative	5	97	102
		104	140	

For the moment, we will deliberately ignore the fact that these counts come from a two-stage sampling and consider only the 142 CAGE-positive and 102 CAGE-negative patients considered in the second phase.

Application (Con't)

We are interested in early detection of alcohol abuse using the CAGE¹. In this case, these are PPV and NPV that are of much interest.

Here, predictive values (with 95% confidence intervals) are found to be:

$$\text{PPV} = 99/142 = 69.7\%, [0.615; 0.771]$$

$$\text{NPV} = 97/102 = 95.1\%, [0.889; 0.984]$$

We would conclude that a patient classified as positive according to the CAGE is really a heavy drinker in 7 cases out of 10.

¹A person who answers “yes”, “sometimes”, or “often” to 2 or more of the questions may have a problem with alcohol.

Application (Con't)

To interpret the preceding result on PPV, it would be necessary to know the prevalence of alcohol abuse in the general population. For instance, [33] report that 4.1% of Canadians had an alcohol dependence in 1994.

Although of less interest here, sensitivity and specificity are also easily computed as:

$$Se = 99/104 = 95.2\%, [0.891; 0.984]$$

$$Sp = 97/140 = 69.3\%, [0.609; 0.768]$$

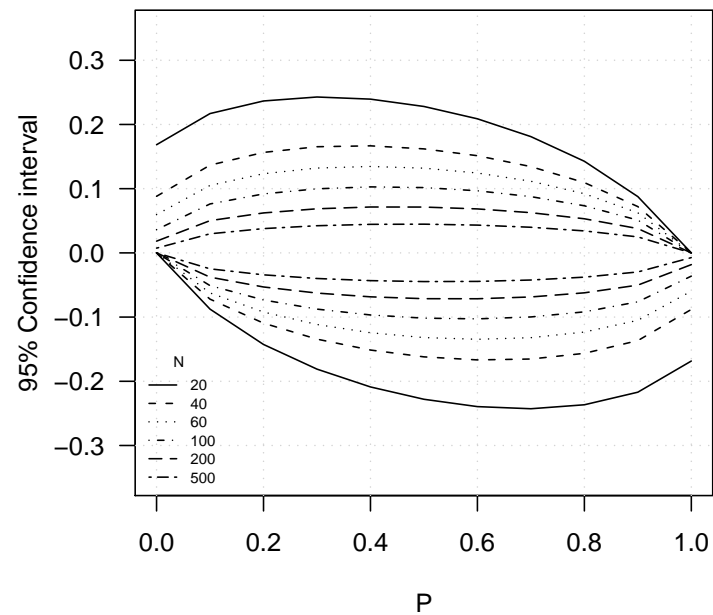
About computation of CIs for Se and Sp

✍ As sensitivity and specificity cannot exceed 100%, neither should their confidence intervals. Such impossible results arise when the standard large sample method for calculating confidence intervals for proportions is used when the proportion is near to zero or one or when the sample is small, or both [11].

Large and low values for proportions are a recurrent problem when dealing with binary variables. Since $\mathbb{V}(X) = npq$, where $X \sim \mathcal{B}(n, p)$ and $q = 1 - p$, the corresponding $(1 - \alpha)$ CIs will be large near $p = 0.5$, especially for small samples, but may exceed allowed values when $p = 0$ or 1 .

About computation of CIs for Se and Sp (Con't)

The next picture was provided by [19] and show how 95% CIs vary with both N and P .



A note on sensitivity analysis

The following is related to epidemiological studies and may be omitted during first reading.

The preceding calculations apply to screening, and in this case Se , Sp , PPV and NPV are useful to assess *criterion validity* in prospective sampling. In more general settings with discrete variables, e.g. case-control studies, misclassification may be seen as some form of measurement error.



Sensitivity analysis aims at quantifying such error [35, p. 347 ff.]. But it should be kept in mind that when using a screening questionnaire, misclassification errors do not necessarily result from the measurement process (reliability) because it may simply be measuring something different as compared to the gold standard (construct validity).

Sensitivity analysis: Exposure prevalence (1)

When interested in the corrected estimation of exposure (screen status in the preceding section), we in fact look horizontally at our 2×2 table of counts. Now, $a+b$ and $c+d$ are true exposition and unexposition frequencies (resp.), and $\hat{\bullet}$ will be their estimators.

	Diseased	Non-diseased
Exposed	\hat{a}	\hat{b}
Unexposed	\hat{c}	\hat{d}

Due to possible classification errors, we have

$$\hat{a} + \hat{b} = \text{Se}(a + b) + \text{FP}(c + d) \quad (4)$$

$$\hat{c} + \hat{d} = \text{FN}(a + b) + \text{Sp}(c + d). \quad (5)$$

Sensitivity analysis (1) (Con't)

The total number of patients, T , remains, however, unchanged even under possible (nondifferential) misclassification, because

$$\begin{aligned} T &= a + b + c + d \\ &= (\text{Se} + \text{FN})(a + b) + (\text{Sp} + \text{FP})(c + d) \\ &= \hat{a} + \hat{b} + \hat{c} + \hat{d}, \text{ since } \text{Se} + \text{FN} = \text{Sp} + \text{FP} = 1. \end{aligned}$$

Solving (4) and (5) yields the correct estimation of $a + b$ and $c + d$. Indeed,

$$\hat{a} + \hat{b} = \text{Se}(a + b) + \text{FP}[(\hat{c} + \hat{d}) - \text{FN}(a + b)]/\text{Sp},$$

Sensitivity analysis (1) (Con't)

which gives

$$\begin{aligned} a + b &= [\text{Sp}(\hat{a} + \hat{b}) - \text{FP}(\hat{c} + \hat{d})] / [\text{Se} \cdot \text{Sp} - \text{FN} \cdot \text{FP}] \\ &= (\hat{a} + \hat{b} - \text{FP} \cdot T) / (\text{Se} + \text{Sp} - 1) \end{aligned} \quad (6)$$

and $c + d = T - (a + b)$.

✍ These estimates are obtained under the assumption that the true sensitivity and specificity are Se and Sp . They may be negative if $\text{Se} \cdot \text{Sp} < \text{FN} \cdot \text{FP}$ which means that the classification method performs worse than random [35, p. 348].

Sensitivity analysis (1) (Con't)

Let's look at the exemple given by [35], p. 349, which comes from a CC study of cancer mortality ($X = 1$, for disease) described in [17].

	$X = 1$	$X = 0$
Cases	45	94
Controls	257	945

A crude estimate of the OR is 1.76, [1.17; 2.61] ($\chi^2(1) = 8.63$, $p = 0.003$).

Sensitivity analysis (1) (Con't)

Now suppose that $Se=0.9$ and $Sp=0.8$ for the cases, and that $Se=Sp=0.8$ for the controls. In other words, exposure detection is considered better for cases. Adapting the notation from cases (\bullet_1) and controls (\bullet_0), from (6) we have:

$$a_0 + b_0 = (257 - 0.2 \times 1202)/(0.8 + 0.8 - 1) = 27.67,$$

$$c_0 + d_0 = 1202 - 27.67 = 1174.33,$$

$$a_1 + b_1 = (45 - 0.2 \times 139)/(0.8 + 0.9 - 1) = 24.57,$$

$$c_1 + d_1 = 139 - 24.57 = 114.43.$$

This gives an estimated OR of 9.16, [4.94;16.93], which is much higher than the uncorrected estimate.

Sensitivity analysis: Disease incidence (2)

When the disease is of interest, the same table of counts should be look columnwise.

	Diseased	Non-diseased
Exposed	\hat{a}	\hat{b}
Unexposed	\hat{c}	\hat{d}

Now, we can use the same equations (4) and (5) to obtain:

$$A = (A^* - FP \cdot N) / (Se + Sp + 1)$$

Recap' on screening “efficacy”

Unlike Se and Sp, PPV and NPV give information about post-test probability of disease. Therefore, PPV is an important measure for a diagnostic method as it allows to quantify the probability that a positive test reflects the underlying condition being tested for.



However, its value depends on the prevalence of the disease, $P_e = (a + b)/T$. Using formula given in (6), then clearly for PPV [35, p. 354]:

$$\begin{aligned} \text{PPV} &= (\# \text{ correctly classified patients in } \hat{a} + \hat{b}) / (\hat{a} + \hat{b}) \\ &= \text{Se}(a + b) / [\text{Se}(a + b) + \text{FP}(c + d)] \\ &= \text{Se}[(a + b)/T] / [\text{Se}((a + b)/T) + \text{FP}((c + d)/T)] \\ &= \text{Se} \cdot P_e / [\text{Se} \cdot P_e + \text{FP}(1 - P_e)] \end{aligned}$$

Recap' on screening “efficacy” (Con't)

NPV and PPV should only be used if the ratio of the number of patients in the disease group and the number of patients in the healthy control group is equivalent to the prevalence of the diseases in the studied population, or, in case two disease groups are compared, if the ratio of the number of patients in disease group 1 and the number of patients in disease group 2 is equivalent to the ratio of the prevalences of the two diseases studied.

Otherwise, positive (PLR = $Se/(1 - Sp)$) and negative (NLR = $(1 - Se)/Sp$) likelihood ratios should be reported instead of NPV and PPV, for likelihood ratios do not depend on prevalence.

ROC analysis

blabla

Adjusting for covariates

[?]

Multi-trait scaling

In essence, Multi-trait scaling (MTS) is a confirmatory approach, like CFA but it can also be used during questionnaire reduction. Its aim is to study convergent and discriminant validity (construct validity).

Multi-trait scaling: Illustration

Items of the SF-36 have been shown to be more highly correlated with their own scales than with other scales [28]. Item scaling tests realized by these authors and summarized by [14, p. 119] are discussed below.

Since there are 36 items and 8 hypothesized scales in the SF-36 questionnaire, we have to summarize $k + (36 - k)$ correlation coefficients for each subscale of k items.

Multi-trait scaling: Illustration (Con't)

Scale	No. items	Conv. validity	Disc. validity	Scaling success *	Homogeneity [†]	Reliability (α)
PF	10	0.49–0.80	0.10–0.54	80/80	0.56	0.93
RP	4	0.65–0.70	0.12–0.58	32/32	0.57	0.84
BP	2	0.70	0.19–0.61	16/16	0.70	0.82
GH	5	0.38–0.72	0.09–0.58	40/40	0.42	0.78
VT	4	0.69–0.75	0.17–0.55	32/32	0.62	0.87
SF	2	0.74	0.20–0.62	16/16	0.74	0.85
RE	3	0.63–0.73	0.11–0.56	24/24	0.61	0.83
MH	5	0.65–0.81	0.11–0.59	40/40	0.64	0.90

* No. convergent correlations significantly higher than discriminant correlations/No. correlations; † Average inter-item correlation

As can be seen, scaling success is always 100% but average inter-item

correlation (as well as Cronbach's α) is lower for the General Health subscale.

Multitrait Multimethod analysis

We already presented the MTS approach, whereby we estimate the so-called success scaling. Here, we will not only focus on the way a given instrument measures one or more traits but compare it to other known instruments, also called methods.

There are various formulations of MTMM models, including *correlated uniqueness model* [26], *CFA model* for MTMM [2, 3], the *direct product model* [6], and the *true score (TS) model* [37].

Applications of MTMM matrix range from sociological [1] to psychological studies [5], including educational assessment [16] and quality of life [23]. More recently, it has been reframed in the multilevel structural modeling framework popularized by Muthén and coworkers [29, 30, 31].

Example of an MTMM matrix

An MTMM matrix would look like the correlation-like matrix shown below (Example is taken from www.socialresearchmethods.net):

		Meth. 1		Meth. 2		Meth. 3				
Meth. 1	trait 1	0.89								
	trait 2	0.51	0.89							
	trait 3	0.38	0.37	0.76						
Meth. 2	trait 1	0.57	0.22	0.09	0.93					
	trait 2	0.22	0.57	0.10	0.68	0.94				
	trait 3	0.11	0.11	0.46	0.59	0.58	0.84			
Meth. 3	trait 1	0.56	0.22	0.11	0.67	0.42	0.33	0.94		
	trait 2	0.23	0.58	0.12	0.43	0.66	0.34	0.67	0.92	
	trait 3	0.11	0.11	0.46	0.34	0.32	0.58	0.58	0.60	0.85

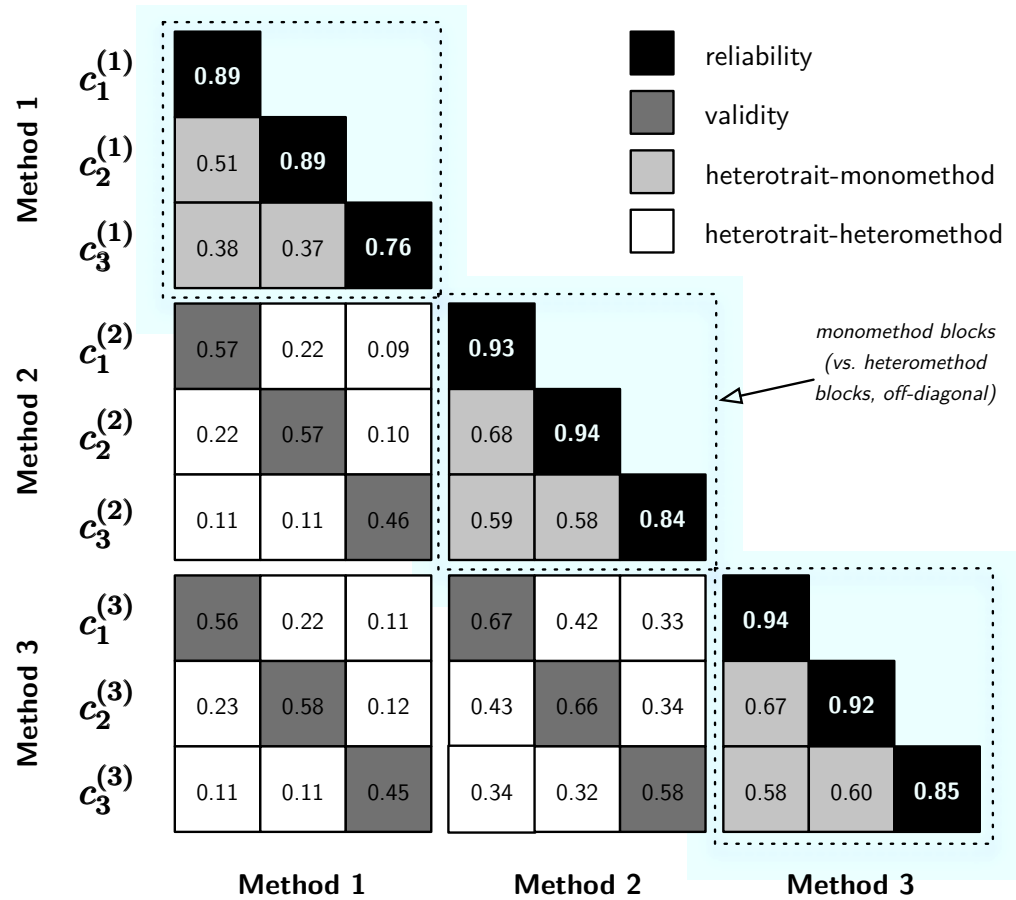
Example of an MTMM matrix (Con't)

Here, we are supposed to measure three *traits* (or constructs) by three *methods* (or instruments). The reliability of each scale is put on the main diagonal. The MTMM matrix summarizes different kind of information:

- *Reliability* of the measurement scales,
- *Validity* of the hypothesized (shared) constructs,
- Relations *between traits within method*, and *between traits across methods*.

Hence, MTMM provides an unique way to assess convergent and discriminant validity [10].

Overview of the MTMM matrix



Summary for the MTMM approach

Pros

rapid consensus
possibility of distant interaction
avoid focus group

Cons

depends on the level of expertise of experts
influence of the formulation
influence of the mediator

Rules of interpretation for the MTMM matrix

- Reliability coefficients should consistently be the highest in the matrix.
- Validity coefficients should be significantly different from zero and high enough to warrant further investigation.
- A validity coefficient should be higher than values lying in its column and row in the same heteromethod block.
- A validity coefficient should be higher than all coefficients in the heterotrait-monomethod triangles.
- The same pattern of trait interrelationship should be seen in all triangles.

The MIMIC model

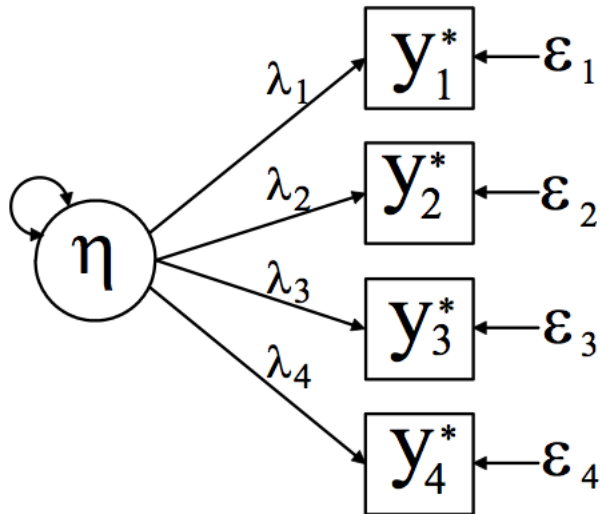
MIMIC stands for *multiple-indicator, multiple cause*, and this belongs to structural equation model. We shall restrict ourselves to a concise presentation of this 'hot' topic, reserving a more complete discussion about MIMIC and SEM in the next Chapters.

Like Multiple group CFA analysis, MIMIC model is used to study measurement invariance and population heterogeneity [] but it should be kept in mind that the MIMIC model can look at differences in intercepts and factor means only, whereas multiple group model can look at these parameters along with factor loadings, residual variances/covariances, factor means, and factor covariances.

In most cases, this kind of model is used when studying Differential Item Functioning (DIF) effects [].

The MIMIC model: Illustration

Consider the following factor model:



$$\mathbf{y}^* = \Lambda\eta + \boldsymbol{\varepsilon}$$

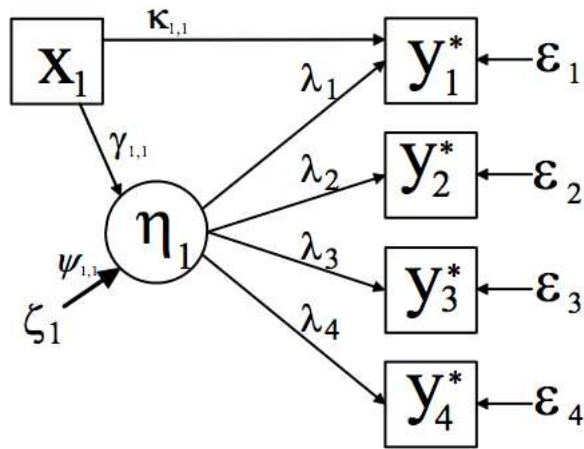
with $\mathbb{V}(\mathbf{y}^*) = \Lambda\Psi\Lambda' + \Theta$ and $\mathbb{V}(\eta) = \Psi = 1$.
The intercept and slope parameters are found to be:

$$a = \frac{\lambda}{\sqrt{1 - \lambda^2}}$$

$$b = \frac{\tau}{\lambda}$$

The MIMIC model: Illustration

Takin into account a grouping factor leads to a slight reformulation of the preceding model, where now:



$$\mathbf{y} = \Lambda \boldsymbol{\eta} + \mathbf{K} \mathbf{x} + \boldsymbol{\varepsilon}$$

with the following constraints: $\mathbb{V}(\boldsymbol{\eta}) = \Psi = 1$ and $\mu_{\boldsymbol{\eta}} = 0$. The intercept and slope parameters are found to be:

$$a = \frac{\lambda}{\sqrt{1 - \lambda^2}}$$

$$b = \frac{\tau - \kappa x}{\lambda}$$

Here, κ allows to account for *uniform* DIF.

References

- [1] R P Althausen and T A Heberlein. Validity and the multitrait-multimethod matrix. *Sociological Methodology*, 2:151–169, 1970.
- [2] R P Althausen, T A Heberlein, and R A Scott. A causal assessment of validity: The augmented multitrait-multimethod matrix. In H M Blalock, editor, *Causal Models in the Social Sciences*, pages 151–169. Chicago: Aldine, 1971.
- [3] F M Andrews. Construct validity and error components of survey measures: a structural modeling approach. *Public Opinion Quarterly*, 48:409–442, 1984.
- [4] W H Angoff. *Educational measurement*, chapter Scales, norms and equivalent scores, pages 508–600. Washington, DC: American Council of Education, 2nd edition, 1971.
- [5] V Benet-Martínez and O P John. Los Cinco Grandes across cultures and ethnic groups: Multitrait multimethod analyses of the big five in spanish and english. *Journal of Personality and Social Psychology*, 75(3):729–750, 1998.

- [6] M W Browne. The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology*, 37:1–21, 1984.
- [7] J C Busch and R M jaeger. Influence of type of judge, normative information, and discussion on standards recommended for the national teacher examinations. *Journal of Educational Measurement*, 27:145–163, 1990.
- [8] B Bush, S Shaw, P Cleary, T L Delbanco, and M D Aronson. Screening for alcohol abuse using the CAGE questionnaire. *American Journal of Medicine*, 82:231–235, 1987.
- [9] M De Cáceres and P Legendre. Associations between species and groups of sites: indices and statistical inference. *Ecology*, 90:3566–3574, 2009.
- [10] D T Campbell and D W Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56:81–105, 1959.
- [11] J J Deeks and D G Altman. Sensitivity and specificity and their confidence intervals cannot exceed 100%. *British Medical Journal*, 318:193, 1999.
- [12] Graham Dunn. *Statistics in Psychiatry*. Hodder Arnold, 2000.
- [13] Bruno Falissard. *Mesurer la subjectivité en santé. Perspective méthodologique et statistique*. Masson, 2008.

- [14] P M Fayers and D Machin. *Quality of life. The assessment, analysis and interpretation of patient-reported outcomes*. Wiley, 2000.
- [15] J Fowles. *Handbook of futures research*. Greenwood Press: Connecticut, 1978.
- [16] K F Gold and B O Muthén. Extensions of covariance structure analysis: Hierarchical modeling of multidimensional achievement data. In *Annual meeting of the American Educational Research Association*, Chicago, Illinois, April 1991.
- [17] S Greenland, A Salvan, D H Wegman, M F Hallock, and T J Smith. A case-control study of cancer mortality at a transformer-assembly facility. *International Archives of Occupational and Environmental Health*, 66:49–54, 1994.
- [18] R K Hambleton and B S Plake. Using an extended angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8:41–55, 1995.
- [19] R Harper and B Reeves. Reporting of precision of estimates for diagnostic accuracy: A review. *British Medical Journal*, 318:1322–1323, 1999.
- [20] Ron D Hays and Dennis Revicki. Reliability and validity (including responsiveness). In Peter Fayers and Ron Hays, editors, *Assessing Quality of Life in Clinical Trials*, chapter 1.3, pages 25–39. Oxford University Press, Second edition, 2005.

- [21] J C Impara and B S Plake. Teachers' ability to estimate item difficulty: A test of the assumptions in the angoff standard setting method. *Journal of Educational Measurement*, 35:69–81, 1998.
- [22] M G Kendall and B Babington Smith. The problem of m rankings. *Annals of Mathematical Statistics*, 10:275–287, 1939.
- [23] S Kuenstner, C Langelotz, V Budach, K Possinger, B Krause, and O Sezer. The comparability of quality of life scores: a multitrait multimethod analysis of the eortc qlq-c30, sf-36 and flic questionnaires. *European Journal of Cancer*, 38(3):339–348, 2002.
- [24] C H Lawshe. A quantitative approach to content validity. *Personnel Psychology*, 28(4):563–575, 1975.
- [25] P Legendre. Coefficient of concordance. In N J Salkind, editor, *Encyclopedia of Research Design*. SAGE Publications, 2010.
- [26] H W Marsh. Confirmatory factor analysis of multitrait-multimethod data: Many problems and few solutions. *Applied Psychological Measurement*, 13:335–361, 1989.
- [27] E McColl. Developing questionnaires. In P Fayers and R Hays, editors, *Assessing*

quality of life in clinical trials, pages 9–21. Oxford University Press, second edition, 2005.

- [28] C A McHorney, J E Ware, J F R Lu, and C D Sherbourne. The MOS 36-item short-form health survey (SF-36): (3) tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care*, 32:40–66, 1994.
- [29] B O Muthén. Latent variable modelling in heterogeneous populations. *Psychometrika*, 54:557–585, 1989.
- [30] B O Muthén. Means and covariance structure analysis of hierarchical data. Technical Report 62, Los Angeles: UCLA, 1990.
- [31] B O Muthén. Multilevel covariance structure analysis. *Sociological Methods & Research*, 22:376–398, 1994.
- [32] D Pilgrim. *Key Concepts in Mental Health*. Sage, second edition, 2009.
- [33] C Poulin, I Webster, and E Single. Alcohol disorders in canada as indicated by the cage questionnaire. *Canadian Medicine Association Journal*, 157:1529–1535, 1997.
- [34] K L Ricker. Setting cut scores: Critical review of angoff and modified-angoff methods. *CSEE*, 2003.

- [35] Kenneth J Rothman and Sander Greenland. *Modern Epidemiology*. Lippincott Williams & Wilkins, Second edition, 1998.
- [36] R J Rovinelli and R J Hambleton. On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research*, 2:49–60, 1977.
- [37] W E Saris and F M Andrews. Evaluation of measurement instruments using a structural modeling approach. In P P Biemer, R M Groves, L E Lyberg, N A Mathiowetz, and S Sudman, editors, *Measurement Errors in Surveys*, pages 575–597. New York: Wiley, 1991.
- [38] Sidney Siegel and Jr N John Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Second edition, 1988.
- [39] B D Zumbo. Validity: Foundational issues and statistical methodology. In C R Rao and S Sinharay, editors, *Handbook of Statistics, Vol. 26: Psychometrics*, pages 45–79. Elsevier Science B.V.: The Netherlands, 2007.

– Typeset with FoilT_EX (version 2)