library(tidyverse)  
library(psych)  
library(mascutils)  
library(printr)  
#library(lavaan)  
options(mc.cores = 8)  
  
purp.analysis <- T

#' Psychometric/designometric response matrix from long designometric data  
#'   
#' @param x long designometric data  
#' @returns psychometric response matrix.  
#' @examples  
#' ldmx <- expand.grid(Design = 1:7, Part = 1:5, Item = 1:3)  
#' ldmx$response <- rbeta(105, 2, 2)  
#' rm\_pmx(ldmx)  
#' rm\_dmx(ldmx)  
  
rm\_pmx <- function(x)  
 x |>   
 group\_by(Part, Item) |>   
 summarize(mean\_resp = mean(response)) |>   
 ungroup() |>   
 arrange(Item) |>   
 spread(Item, value = mean\_resp) |>   
 select(-Part)  
  
#' @rdname rm\_pmx  
rm\_dmx <- function(x)   
 x |>   
 group\_by(Design, Item) |>  
 summarize(mean\_resp = mean(response)) |>   
 ungroup() |>   
 spread(Item, value = mean\_resp) |>   
 select(-Design)  
  
  
#' Psychometric function wrappers  
#'   
#' @param Data long designometric data  
#' @returns Results per Perspective and Scale  
  
alpha\_ci <- function(Data){  
 Scale <- str\_c(distinct(Data, Scale)$Scale)   
 model\_psych <-   
 psych::alpha(rm\_pmx(Data), check.keys = FALSE, n.iter = 100)$boot |>   
 as\_tibble() |>   
 mutate(Perspective = "psychometric")  
 model\_design <-   
 psych::alpha(rm\_dmx(Data), check.keys = FALSE, n.iter = 100)$boot |>   
 as\_tibble() |>   
 mutate(Perspective = "designometric")  
 out <-   
 bind\_rows(model\_psych,  
 model\_design) |>   
 select(Perspective, std.alpha) |>   
 group\_by(Perspective) |>   
 summarize(center = mean(std.alpha),  
 lower = quantile(std.alpha, .025),  
 upper = quantile(std.alpha, .975)) |>   
 mutate(Scale = Scale) |>   
 mascutils::go\_first(Scale, Perspective)  
 out  
}  
  
item\_rel <- function(Data){  
 #Data <- D\_1 |> filter(Scale == "HQI")  
 Scale <- str\_c(distinct(Data, Scale)$Scale)  
 model\_psych <-   
 psych::alpha(rm\_pmx(Data), check.keys = FALSE)$item.stats |>   
 as\_tibble(rownames = "Item") |>   
 mutate(Perspective = "psychometric")  
 model\_design <-   
 psych::alpha(rm\_dmx(Data), check.keys = FALSE)$item.stats |>   
 as\_tibble(rownames = "Item") |>   
 mutate(Perspective = "designometric")  
   
# model <- M\_1\_design  
 out <-   
 bind\_rows(model\_psych,  
 model\_design) |>   
 mutate(Scale = Scale) |>   
 go\_first(Scale, Item, Perspective) |>   
 arrange(Scale, Item, Perspective)  
 out  
}  
  
parallel\_analysis <- function(Data, n, persp, scales){  
 if (persp == "D") {  
 data <- rm\_dmx(Data)  
 main <- str\_c("Designometric Parallel Analysis of ", scales)  
 }  
 if (persp == "P") {  
 data <- rm\_pmx(Data)  
 main <- str\_c("Psychometric Parallel Analysis of ", scales)  
 }  
 psych::fa.parallel(data,  
 fa = "fa",  
 fm = "minres",  
 nfactors=n,  
 main=main)  
   
}

# The designometric perspective and the psychometric fallacy

## Introduction

In modern industrial practice, rating scales have their place as an always available and cheap method for comparing or benchmarking designs. In the decision process, the everyday value of a rating scales stands and falls with two properties: validity and reliability.

*Developing* a valid and reliable rating scale is quite an undertaking. *Psychometrics* is the science of assigning numbers to persons so that they can be compared by psychological functioning. Traditionally, this served to measure skills, such as mathematical intelligence or comprehension of language. With the time, researchers also became more interested in elusive properties of persons, such as psycho-social tendencies (known as the Big Five).

After the landrush phase of the UX revolution, Bargas-Avila & Hoernbaek ((Old wine in new bottles)) counted hundreds of new rating scale instruments. Barely any of these instrument underwent the same scrutiny as, for example, a rating scale for psycho-pathological diagnosis would have to. But, frequently some psychometric tools were used at one point during development time, for example by reporting reliability estimates. Also for users of such instruments, it is common to perform basic psychometric sanity checks on their data. The central point of this paper is a certain catch that can occur when translating between psychological research and design research, which we call the *psychometric fallacy*.

In psychometric situations, the atomic observation is an encounter of a person with a test item. This is repeated with more items to improve the precision of the measurement. If many persons are assessed this way, the result is a Person X Item *response matrix*, from which *person scores* can be extracted for ranking. *Item scores*, are mostly used during scale development.

The logical argument developed in this paper is that instruments in design research exist to rank designs and the atomic observation is an encounter of a design with a person and an item. This forms a box of data, which can be collapsed by three perspectives. We demonstrate how standard psychometric tools can be used correctly using the designometric response matrix. We call it the psychometric fallacy, when a psychometric matrix is used instead.

…

Aim of this study is also to seek empirical evidence that the psychometric fallacy is not just sophistry, but can result in real biases when developing or using rating scales. For this purpose, data from five experiments was subjected to typical rating scale validation techniques under both perspectives, psychometric (pretending the fallacy) and designometric (using the proper response matrix).

### Psychometrics

When it only takes a single probe to accurately estimate the body temperature of a person, why are psychological tests and math exams composed of several items? The primarily reason is that psychological instruments are highly noisy, be it self-report scales, reaction times or physiological measures. The idea of repeated measures goes back to *Classic Test Theory*, which solves this problem by decomposing every measure into a relevant *systematic component* and a nuisance *error component*. By definition, the systematic component re-occurs and is strengend with every measurement, whereas the error is fully random, cancelling itself out in the long run. When certain conditions are met, it is possible to create an arbitrary precise estimate by adding more repetitions.

The primary practical uses of psychometric instruments are either about making cost effective *prediction* about a person. A good score in a written drivers test predicts that a person will make fewer errors on the road which decreases the chance of collisions. Psycho-diagnostic rating scales are often used in screening tests, e.g. for depression, whereas performance tests are used in competitive situations, such as personnel selection or education.

#### Item selection

The first challenge when designing a psychometric inventory is to understand the domain, ideally in terms of involved areas of psychological functioning. Based on the domain analysis, the researcher creates a candidate *item pool*, which is usually much larger than the targeted set of items. Up to this point, the process is mostly qualitative, divergent and creative.

Several psychometric methods can be used to successively reduce the initial set of items to reliable scales and to find effective multi-scale strucures.

According to CTT, the errors of multiple items cancel each other out, *if they fully agree* on the systematic component. This can be the case with extremely repetitive items, such as trials in the Stroop task. But in practice, item scores only agree to some degree on what they measure. Scale consistency by Cronbach is a statistic to measure the overall level of agreement. From this derives a basic procedure for item selection, where consistency of the full item set is compared to the data set *excluding* item . If scale consistency improves by removing the item, the item is marked for removal.

The procedure of step-wise removal is sometimes adequate, but can also fail, when the alleged property truly has more than one component. In this case, *factor-analytic* methods can be used to separate components into multiple sub scales with good consistency.

#### Factor Structures

In modern psychometrics, the relation between a properties of persons and multiple measure is known as the distinction between *latent variables* (not observable, but true) and *indicator variables* (observable, but imperfect).

Modern instruments often carry a complex domain structure, where multiple latent variables are put in relation to each other. A prominent example is the Big 5 inventory, which claims that social behaviour can be predicted by five psycho-social traits, each of which is assessed by a main scale divided into multi-item subscales. For the five main traits a high level of independence can be expected by the way the Big5 was constructed, but the subscales should correlate more strongly.

While the primary aim of domain analysis is to gain sufficient coverage of the domain, it often at least indicates a possible structure for a multi-item instrument. In some cases, a domain analysis is driven by a theoretical structure in the first place. For example, a rating scale for mental workload could very well be based on the multiple-resource theory, which predicts that different sensory modes are processed independently, translating into one scale per sensory mode.

1. Auditory load: *It is difficult to understand the signal tones.*
2. Visual load: *The amount of visual information can be overwhelming*.

Prior structures often also emerge from qualitative results in mixed-method psychometric processes. …

When prior structures exist, Confirmatory Factor Analysis (CFA) is the most recommended technique to critically assess prior assumptions about the latent variable structure. In the above example of the Big5 inventory, a hierarchical CFA model can be used to verify that main scales correlate weakly and subscales correlate moderately. Once a good model is established, CFA also supercedes Cronbach , as parameter loadings can be used to diagnose items.

When a robust prior structure is not available, a common technique for finding new structures is Exploratory Factor Analysis, which requires the researcher to name the number of factors and how these factors are correlated. Several methods have been proposed to identify the *optimal number of factors* before-hand. Early tools used the decline in Kaiser eigenvalues and the visual *elbow* criterion, whereas modern tools use resampling techniques to determinje the optimal number of factors. The second choice to make is *scale rotation.* Orthogonal rotation applies, when the components are largely independent, such as arithmetic skills and text comprehension. For true sub scales a significant correlations between underlying factors is very likely, and oblique rotation should be used, instead.

#### The Response Matrix

All psychometric methods mentioned so far have in common that measures are provided in form of response matrices, where rows identify with participants and columns with items. A common operation during item selection is to estimate consistency on the full response matrix, remove a dubious item from the matrix to see if this improves consistency. EFA takes a response matrix as input and produces item loadings for it. For multi-scale data CFA tools take response matrix with named columns as input, together with a formula to group items to factors. …

The idea of response matrices is most pronounced in item response theory (IRT), a psychometric approach rivaling classic test theory and factors analysis. In IRT, each cell in the response matrix is seen as an encounter between an item and a participant, from which a measure arises. Items, like participants, are seen as individuals, with different response characteristics during the encounter. In the most basic Rasch model, the probability of success in a person-task encounter depends only on the difference between task difficulty and the person ability. More complex IRT models allow more subtle item characteristics, and sophisticated tools have been developed to detect and prevent *differential item functioning*

in world-spanning studies like TIMMS and PISA. IRT models take the person x item matrix as input and produce estimates on both margins, item parameters and person parameters. During the development of an IRT model, sophisticated tools are used to select only those items that strictly behave as the model states.

For all levels of sophistication or schools of thought, what counts during scale development is that item of a test are well-behaved. In the same way teachers use multiple tests to estimate a student’s skill with adequate accuracy, many responses are needed to precisely assess the characteristics of an item during development time. This is what creates the demand for the enormous sample sizes, for which psychometric development is most feared for.

#### Sample sizes

The enormous sample sizes consumed during psychometric scale validation have several causes:

* Psychometric techniques tend to have many parameters and by the rule that the number of independent observations must at least match the number of parameters, a lower boundary can be established. A 24 item three-way EFA already produces 72 parameters. But, during scale validation, the number of items is often three times larger, which adds up.
* With noisy data certain estimation procedures need a lot of data to converge properly, for example in conditional maximum likelihood estimation.
* When a structure is desired, but no prior structure exists, EFA can be used to find factors. But the consequence is that the data set is consumed. To confirm the found structure by CFA, a new sample is required.
* The main areas of psychometric research are education and clinical diagnosis, where a rating scale or test can have a significant impact on a person’s life. These instruments ought to be extra-hardened against biases and require the greatest scrunity. For item in a PISA test, proof is required that it has the same characteristics all around the globe.

### Designometrics

Everything said above is essentially rooted in mathematical theory, and therefore independent of semantics. Formally, constructing rating scales to rank humans should be not much different to constructing a usability metric for websites. When a UX scale is used to decide between designs A and B in industrial practice, it is legit to ask how precise the measurements are. Indeed, designometric scales can efficiently be constructed and evaluated using existing psychometric tools, such as Cronbach , factor analysis techniques ans IRT models for maximum rigor.

There are just two differences: First, in psychometrics the entity to be ranked is persons, whereas designometric applications one or more designs get numbers attached. Second, a psychometric measure is an encounter of a person with an item, whereas a designometric measure is an encounter of a design with a person and an item. The three-way encounter and designs being subject to measurement is what defines *designometric situations*.

#### The designometric perspective

A complete designometric data set is a cross-product of three samples, designs, users and items. It forms a “response box”, rather than a matrix, and can not directly be processed with psychometric tools.

A practical solution is to average across one factor, which produces a flat response matrix to fit into psychometric tools. For designometric instrument validation, averaging across Person creates a design-by-item *designometric response matrix*. This matrix takes the place of the psychometric response matrix and researchers can use it with standard psychometric tools to select proper items and find good structures.

#### Designometric sampling

When Design takes the place of Person, it is implied that items must now be well-behaved in ranking designs, and substantial samples of designs are now required to prove that. This can be a minor or huge problem, depending on the class of designs. For e-government websites a large design sample will be easier to come by than a sample of human-like robots or self-driving cars.

A connected problem is that the designometric perspective participant-level information is reduced, but they are still involved in the designometric encounter. In the special case that scales are used for first impression ratings, it is possible to rund cross-product complete designometric encounters. For example, when we used the Eeriness scale on artificial faces, a single observation only took around 4 seconds, making it relatively easy to fill the designometric box with data, even repeatedly ((BT Robbin Koopmans)).

When a real interactive experience is subject of the measure, a measurent can take from several minutes to hours and a complete designometric encounter becomes impractical. A way to mitigate this problem is to use an experimental design that is *planned incomplete*. Essentially, a planned incomplete validation study has all participants encounter only a partition of the design sample. For example a study consisting of a sample of 100 designs let every participant encounter a different set of ten designs. As long as all designs are covered by at least one participant, this will result in a complete design-by-item matrix after collapsing along participants.

A variation of planned incomplete studies is to *successively* build the sample of designs. This is especially useful, when dealing with emerging classes of designs. This happened in the BUS-11 studies, where initially it was difficult to build a substantial sample, before it became more common.

#### The psychometric fallacy

Designometric scales can be developed and validated with established psychometric tools, when using proper design x item response matrices and a sufficiently large sample of designs to prove that items are well behaved. In contrast, many designometric instruments have not been validated using a large sample of designs, but rather on a psychometric matrix. This we call the *psychometric fallacy*.

Formally, a designometric box can produce a psychometric response matrix by averaging over Design. When a scale validation study in design research is under the psychometric fallacy, validation metrics such as item reliability may be meaningless for the purpose of ranking designs. Rather, the metric will refer to the capability of the item to discriminate persons by their sensitivity to the design feature in question. For example, a scale for comparing designs by beauty would become a scale to rank persons by how critical they are with respect to interface beauty. This is not the same and in the next section we show by simulation that the differences between designometric and psychometric perspectives can be dramatic.

During the scale development process, the psychometric fallacy appears in two forms, one is repairable, whereas the other is fatal. We speak of a repairable psychometric fallacy, when the validation study properly collected designometric data, but used psychometric response matrices for validation. This can be repaired by redoing the analysis now using a designometric matrix.

A study that did not collect a sample of designs, but used only a single or very few designs fell for the *fatal psychometric fallacy*. In these cases, researchers have failed to recognize a simple truth: The capacity to discriminate between designs can impossibly be validated on a single design. Every alleged designometric instrument, where this has happened during the validation phase, cannot be trusted.

Recall that psychometric validations require large samples of participants! Similar sample sizes will be required for designs when validating designometric instruments. So, even validation studies that included multiple designs, may in practice not be repairable, because the sample of designs is too small for that type of analysis (e.g. factor analysis). To give an example MacDorman (XY) validated the Godspeed Index, a common inventory to evaluate robotic designs. While scale reliability was assessed under a psychometric perspective, validity was correctly assessed using a sample of four designs. This allowed the authors to do a rough test on the ranking capabilities of their scales, but would not be suitable assess more detailed metric properties of the scale.

As a milder form *run-time psychometric fallacy* appears when an instrument is used in practice to take measures on a single design. The result will inevitable look like a psychometric response matrix and, given that publication rules (e.g. APA guidelines) often require to report some psychometric properties, it may be tempting for the researcher to run a psychometric test on reliability. While the run-time fallacy does not have the same impact as development-time fallacies, it may have cause some head aches when a validated instrument seems to have poor reliability.

In the following section, we construct a case by simulation, showing that psychometric and designometric perspectives can result in dramatically different results. In the remainder of the study, we will use data from past experiments to evaluate how strong the deviations are with several existing and commonly used rating scales.

The formal argumeńt for the psychometric fallacy is that using psychometric response matrices would allow to construct an instrument for ranking designs using a single design.

The aim of the present study is to assess the possible risks and consequences of falsely using the psychometric perspective. In the following section, we construct a case by data simulation, where psychometric and designometric perspectives result are dramatically different. In the remainder of the study, we will use data from past experiments to compare psychometric versus designometric perspective on several widely used UX rating scales.

## Simulation study

The following example demonstrates the difference by simulating an extreme situation, where a fictional three-item scale for Coolness is highly reliable for persons, but has no reliability at all for discerning the tested designs. Such a pattern can occur for the trivial reason that the sample have little or no variance with respect to Coolness. In the following simulation, we assume that the Coolness scale be tested on a sample of 50 designs and 50 participants. The key here is that participants vary strongly in their appreciation of Coolness (), whereas the sample of designs varies little in Coolness (), perhaps

set.seed(42)  
  
n\_Design = 20  
n\_Part = 20  
n\_Item = 4  
n\_Obs = n\_Design \* n\_Part \* n\_Item  
  
Designs <- tibble(Design = as.factor(1:n\_Design),  
 cool\_Design = rnorm(n\_Design, -0.5, .02)) ## low level, little variation  
  
Parts <- tibble(Part = as.factor(1:n\_Part),  
 cool\_Part = rnorm(n\_Part, 0, .2)) ## strong variance in tendency to judge sth. cool  
  
Items <- tibble(Scale = "Coolness",  
 Item = as.factor(1:4),  
 cool\_Item = rnorm(n\_Item, 0, .2)) ## item strength: understating items get lower values  
  
Coolness <- expand\_grid(Design = Designs$Design,  
 Part = Parts$Part,  
 Item = Items$Item) |>   
 left\_join(Designs) |>   
 left\_join(Parts) |>   
 left\_join(Items) |>   
 mutate(response = mascutils::rescale\_zero\_one(cool\_Design + cool\_Part - cool\_Item + rnorm(n\_Obs, 0, .5)))

## Joining with `by = join\_by(Design)`  
## Joining with `by = join\_by(Part)`  
## Joining with `by = join\_by(Item)`

alpha\_ci(Coolness) |> knitr::kable()

## `summarise()` has grouped output by 'Part'. You can override using the `.groups` argument.  
## Number of categories should be increased in order to count frequencies.  
## `summarise()` has grouped output by 'Design'. You can override using the `.groups` argument.  
## Number of categories should be increased in order to count frequencies.

## Warning in psych::alpha(rm\_dmx(Data), check.keys = FALSE, n.iter = 100): Some items were negatively correlated with the first principal component and probably   
## should be reversed.   
## To do this, run the function again with the 'check.keys=TRUE' option

## Some items ( 1 ) were negatively correlated with the first principal component and   
## probably should be reversed.   
## To do this, run the function again with the 'check.keys=TRUE' option

| Scale | Perspective | center | lower | upper |
| --- | --- | --- | --- | --- |
| Coolness | designometric | 0.1108806 | -0.5513231 | 0.5186778 |
| Coolness | psychometric | 0.9291677 | 0.8392027 | 0.9642918 |

This simple example demonstrate that a scale can produce excellent reliability when measuring person sensitivity, but poor and uncertain reliability on designs. Under the psychometric fallacy it can happen that excellent reliability is reported, while it is actually unknown, or very poor.

In the following study we use data from several previous experiments that had produced designometric data sets.

## Methods

From a theoretical perspective the psychometric fallacy is obvious and we have demonstrated by simulation that the worst case is possible. Here, we explore the biases that can occur when a psychometric response matrix is used, in place of a designometric. Designometric data was collected from several experiments to compare three commonly used psychometric statistics under the correct designometric perspective and under the psychometric fallacy.

### Data sets

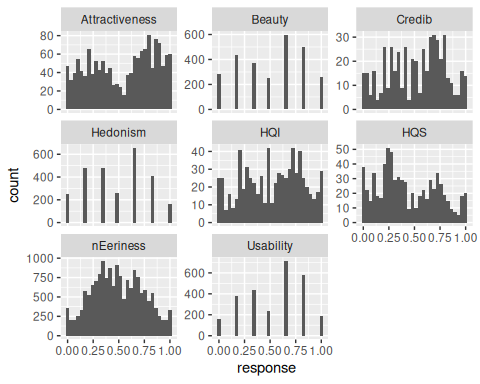
The data used for analysis originates from five experiments (DK, PS, AH, QB, DN). While these experiments were carried out to test their own hypotheses, they have in common that participants saw pictures of many designs and were asked to respond to items taken from one or more scales. In QB and DN participants saw pictures of home pages and responded to several user experience scales, whereas in AH, DK and PS the stimuli were robot faces. Some of the original experiments used manipualtion of presentation time to collect data on subconscious cognitive processing. For the analysis here, only responses at presentation times of 500ms and 2000m were used.

Per trial participants saw a single design followed by a random single item, resulting in a sparse designometric box. However, when collapsing the box to either psychometric RM or designometric RM, the result is completely filled response matrices.

load("DMX\_data.Rda")

D\_1 |>   
 ggplot(aes(x = response)) +  
 geom\_histogram() +  
 facet\_wrap(~Scale, scale = "free\_y")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



D\_1 |>   
 distinct(Study, Design) |>   
 group\_by(Study) |>   
 summarize(n\_Design = n()) |>   
 ungroup() |>   
 left\_join(D\_1 |>   
 distinct(Study, Part) |>   
 group\_by(Study) |>   
 summarize(n\_Part = n())|>   
 ungroup()) |>   
 left\_join(D\_1 |>   
 group\_by(Study) |>   
 summarize(n\_Obs = n())|>   
 ungroup()  
 ) |> knitr::kable()

## Joining with `by = join\_by(Study)`  
## Joining with `by = join\_by(Study)`

| Study | n\_Design | n\_Part | n\_Obs |
| --- | --- | --- | --- |
| AH | 20 | 45 | 10800 |
| DK | 80 | 35 | 2800 |
| DN | 48 | 42 | 8064 |
| PS | 87 | 39 | 2808 |
| QB | 76 | 25 | 1900 |
| SP | 66 | 40 | 1440 |

D\_1 |>   
 distinct(Scale, Design) |>   
 group\_by(Scale) |>   
 summarize(n\_Design = n()) |>   
 ungroup() |>   
 left\_join(D\_1 |>   
 distinct(Scale, Part) |>   
 group\_by(Scale) |>   
 summarize(n\_Part = n())|>   
 ungroup()) |>   
 left\_join(D\_1 |>   
 group\_by(Scale) |>   
 summarize(n\_Obs = n())|>   
 ungroup()  
 ) |> knitr::kable()

## Joining with `by = join\_by(Scale)`  
## Joining with `by = join\_by(Scale)`

| Scale | n\_Design | n\_Part | n\_Obs |
| --- | --- | --- | --- |
| Attractiveness | 66 | 40 | 1440 |
| Beauty | 48 | 42 | 2688 |
| Credib | 76 | 25 | 500 |
| HQI | 76 | 25 | 700 |
| HQS | 76 | 25 | 700 |
| Hedonism | 48 | 42 | 2688 |
| Usability | 48 | 42 | 2688 |
| nEeriness | 127 | 119 | 16408 |

### Scales

For the following rating scales responses have been extracted from the original experimental data:

The *Eeriness* scale has been developed for measuring negative emotional responses towards robot faces and is a primary research tool on the Uncanny Valley phenomenon. Ho & MacDorman(2017) present an advanced psychometric validation of the scale. The study made use of 12 animated characters (Designs), avoiding the fatal fallacy to some degree, but the data analysis is under psychometric perspective.

The *Attractiveness* scale is part of the User Experience Questionnaire (UEQ) inventory. Is has been vaidated by ((Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and Evaluation of a User Experience Questionnaire. . 63–76. <https://doi.org/10.1007/978-3-540-89350-9_6>)) The UEQ has undergone basic psychometric evaluation in six studies with a single design each.

The two scales *Hedonic Quality - Identity (HQI)* and *Hedonic Quality - Stimulation (HQS)* are from the AttrakDiff2 inventory. AttrakDiff2 underwent basic evaluation using only three Designs under psychometric perspective (level 1 fallacy) ((Hassenzahl, M., Burmester, M., Koller, F., AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität)).

The Credibility scale … #### ((HERE))

D\_1 |>   
 group\_by(Study, Scale) |>   
 summarize(n\_Items = n\_distinct(Item),  
 n\_Part = n\_distinct(Part),  
 n\_Design = n\_distinct(Design),  
 n\_Obs = n()) |>   
 ungroup() |> knitr::kable()

## `summarise()` has grouped output by 'Study'. You can override using the `.groups` argument.

| Study | Scale | n\_Items | n\_Part | n\_Design | n\_Obs |
| --- | --- | --- | --- | --- | --- |
| AH | nEeriness | 8 | 45 | 20 | 10800 |
| DK | nEeriness | 8 | 35 | 80 | 2800 |
| DN | Beauty | 4 | 42 | 48 | 2688 |
| DN | Hedonism | 4 | 42 | 48 | 2688 |
| DN | Usability | 4 | 42 | 48 | 2688 |
| PS | nEeriness | 8 | 39 | 87 | 2808 |
| QB | Credib | 5 | 25 | 76 | 500 |
| QB | HQI | 7 | 25 | 76 | 700 |
| QB | HQS | 7 | 25 | 76 | 700 |
| SP | Attractiveness | 6 | 40 | 66 | 1440 |

### Statistics

Goal of the analysis is to examine in how much the psychometric fallacy creates real biases. For this purpose, three basic psychometric techniques were applied to several data sets. Scale consistency and item reliability, were measured Cronbach was used. For the number of factors parallel analysis was used, which compares the eigenvalues to a baseline established by bootstrapping ((REF)). For all three statistics, functions from R package Psych were used ((REF)).

## Results

### Scale consistency

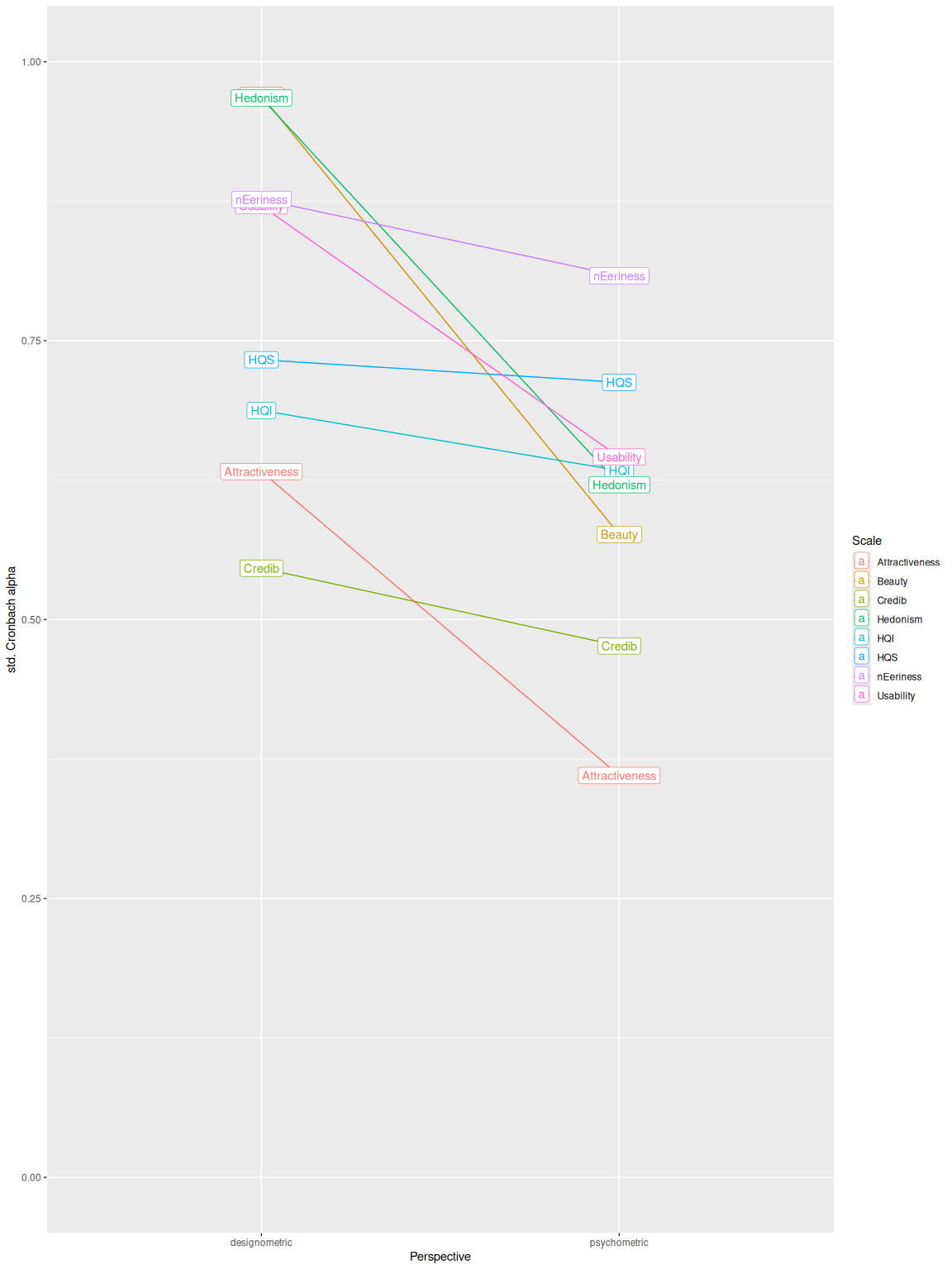
Scale\_rel <-  
 D\_1 |>   
 group\_by(Scale) |>   
 group\_split() |>   
 map\_df(alpha\_ci)

## Some items ( Att4 Att6 ) were negatively correlated with the first principal component and   
## probably should be reversed.   
## To do this, run the function again with the 'check.keys=TRUE' option

Scale\_rel |> knitr::kable()

| Scale | Perspective | center | lower | upper |
| --- | --- | --- | --- | --- |
| Attractiveness | designometric | 0.6327322 | 0.4558714 | 0.7786169 |
| Attractiveness | psychometric | 0.3602675 | -0.0474338 | 0.5841987 |
| Beauty | designometric | 0.9697228 | 0.9551937 | 0.9809968 |
| Beauty | psychometric | 0.5763924 | 0.3488843 | 0.7478395 |
| Credib | designometric | 0.5460676 | 0.2793386 | 0.7198503 |
| Credib | psychometric | 0.4764126 | 0.0080803 | 0.7018859 |
| HQI | designometric | 0.6874288 | 0.5464933 | 0.7657372 |
| HQI | psychometric | 0.6336471 | 0.1239533 | 0.8607239 |
| HQS | designometric | 0.7329270 | 0.6028342 | 0.8140758 |
| HQS | psychometric | 0.7126751 | 0.4094632 | 0.8659158 |
| Hedonism | designometric | 0.9676909 | 0.9518783 | 0.9785622 |
| Hedonism | psychometric | 0.6209545 | 0.4178271 | 0.7460030 |
| Usability | designometric | 0.8709788 | 0.8027483 | 0.9238719 |
| Usability | psychometric | 0.6457525 | 0.4301707 | 0.7738356 |
| nEeriness | designometric | 0.8766474 | 0.8365439 | 0.9065569 |
| nEeriness | psychometric | 0.8081287 | 0.7054118 | 0.8727069 |

Scale\_rel |>   
 ggplot(aes(color = Scale,  
 label = Scale,  
 x = Perspective,  
 y = center,  
 ymin = lower,  
 ymax = upper)) +  
 geom\_point() +  
 geom\_line(aes(group = Scale)) +  
 ylab("std. Cronbach alpha") +  
 geom\_label() +  
 ylim(0,1)

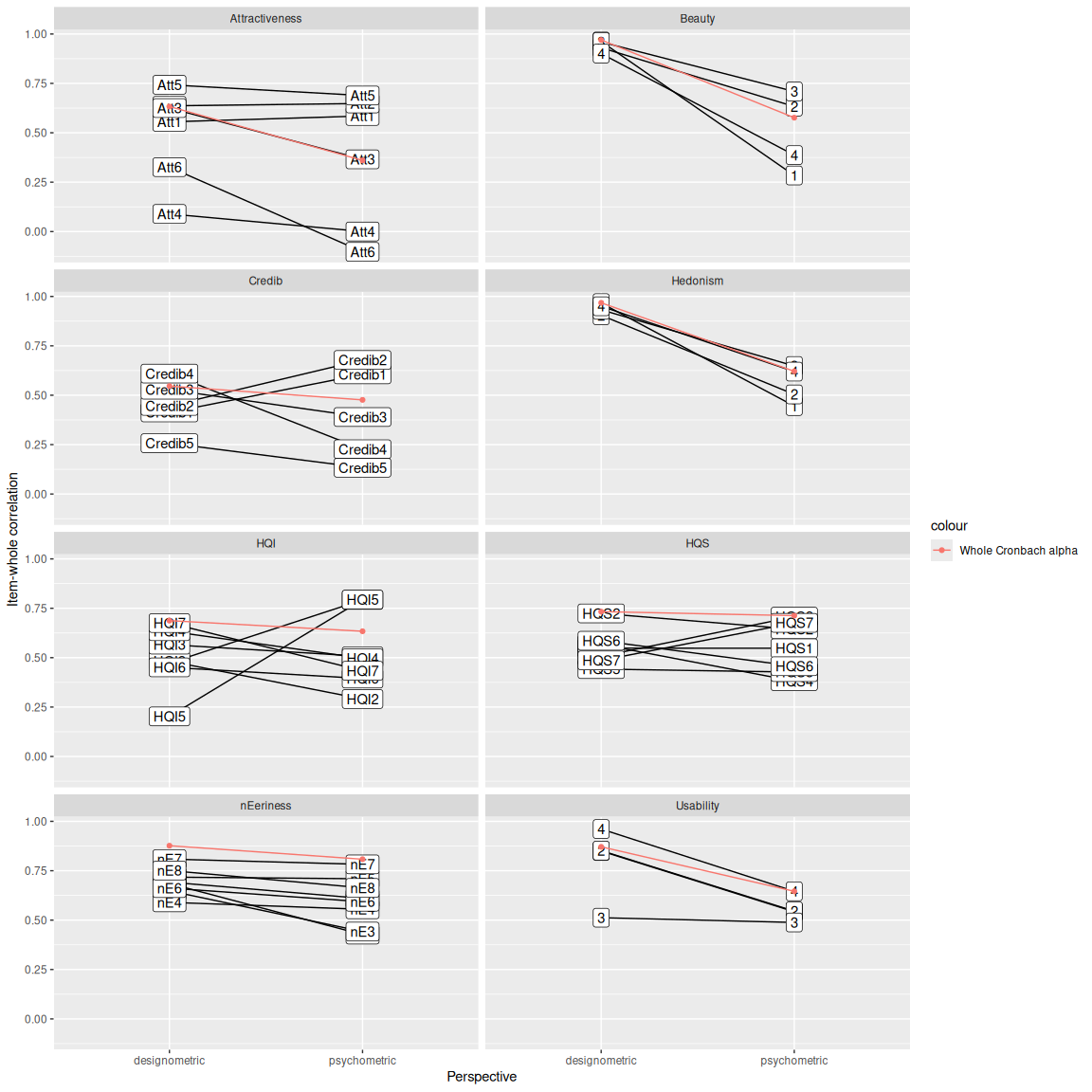


### Item reliability

Item\_rel <-  
 D\_1 |>   
 group\_by(Scale) |>   
 group\_split(Scale) |>   
 map\_df(item\_rel)

## Some items ( Att4 Att6 ) were negatively correlated with the first principal component and   
## probably should be reversed.   
## To do this, run the function again with the 'check.keys=TRUE' option

Item\_rel |>   
 ggplot(aes(x = Perspective,  
 y = r.cor)) +  
 geom\_line(aes(group = Item)) +  
 ylab("Item-whole correlation") +  
 geom\_label(aes( label = Item)) +  
 facet\_wrap(~Scale, ncol = 2) +  
 geom\_point(data = rename(Scale\_rel, alpha = center),  
 aes(x = Perspective,   
 y = alpha,  
 col = "Whole Cronbach alpha")) +  
 geom\_line(data = rename(Scale\_rel, alpha = center),  
 aes(x = Perspective,   
 y = alpha,  
 group = Scale,  
 col = "Whole Cronbach alpha"))



### Number of factors

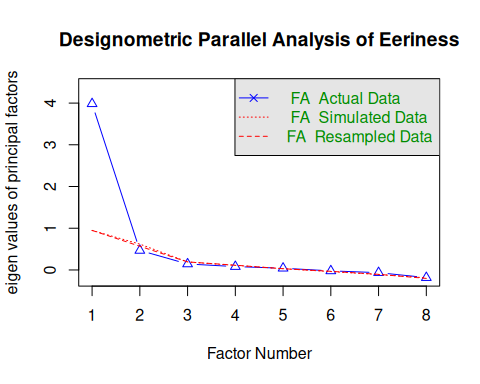
Often, different scales are used in combination to create a more complete picture. It is usually the aim that a scale measures exactly one construct (or latent variable) and that different scales measure different constructs. As a counter-example, MacDorman found that the Eeriness scale decomposes into two slightly different aspects, summarized as “eery” and “spine-tingling”. In contrast, the AttrakDiff2 questionnaire comprises two scales to capture supposedly different aspects.

Given a response matrix, the number of factors can be estimated using parallel analysis. Ideally, this procedure returns exactly as many factors as there are separate scales. Here, we use parallel analysis to assess whether the two perspectives produce the expected number of factors, or at least agree on a number.

*Eeriness* was initially considered a one-dimensional construct. Nevertheless, it has been suggested that it comprises two slightly different factors ((REF)).

parallel\_analysis(D\_Eer, 2, "D", "Eeriness")

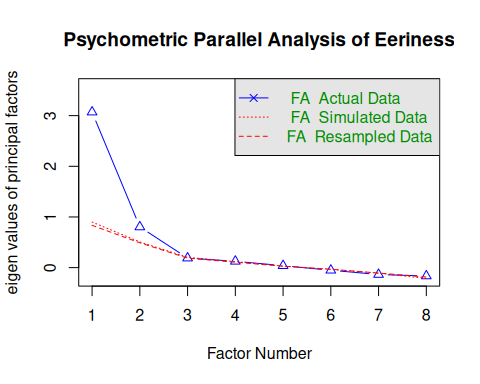
## `summarise()` has grouped output by 'Design'. You can override using the `.groups` argument.



## Parallel analysis suggests that the number of factors = 1 and the number of components = NA

parallel\_analysis(D\_Eer, 2, "P", "Eeriness")

## `summarise()` has grouped output by 'Part'. You can override using the `.groups` argument.



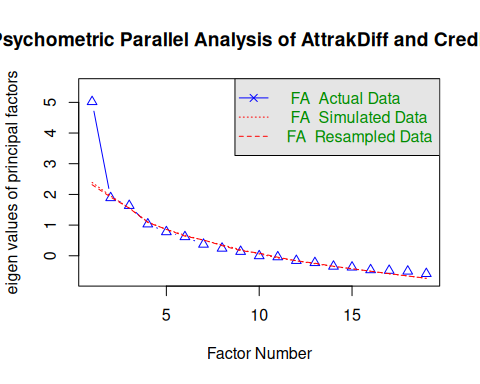
## Parallel analysis suggests that the number of factors = 2 and the number of components = NA

The results suggest that under psychometric perspective there is only one latent variables, whereas the designometric perspective produces two, as expected.

On theoretical grounds, the AttrakDiff2 inventory splits hedonistic quality into two components, Identity and Stimulation, while the credibility scale is a complete separate construct . Depending on whether Hedonism splits up like expected by theory, we would expect a two or three-factor solution.

parallel\_analysis(D\_Att, 3, "P", "AttrakDiff and Credibility")

## `summarise()` has grouped output by 'Part'. You can override using the `.groups` argument.

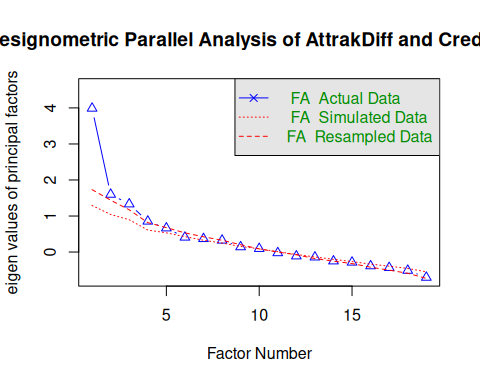


## Parallel analysis suggests that the number of factors = 1 and the number of components = NA

parallel\_analysis(D\_Att, 3, "D", "AttrakDiff and Credibility")

## `summarise()` has grouped output by 'Design'. You can override using the `.groups` argument.

## Warning in fa.stats(r = r, f = f, phi = phi, n.obs = n.obs, np.obs = np.obs, : The estimated weights for the factor scores are probably  
## incorrect. Try a different factor score estimation method.



## Parallel analysis suggests that the number of factors = 5 and the number of components = NA

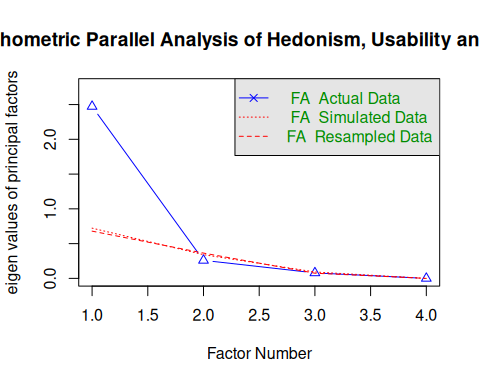
Under a psychometric perspective, all items represent a single latent construct. In contrast, the designometric analysis yielded five factors.

### Hedonism, Usability and Beauty

In DN three separate scales were used, but parallel analysis suggests that these capture the same latent variable under both perspectives.

parallel\_analysis(D\_HUB, 3, "P", "Hedonism, Usability and Beauty")

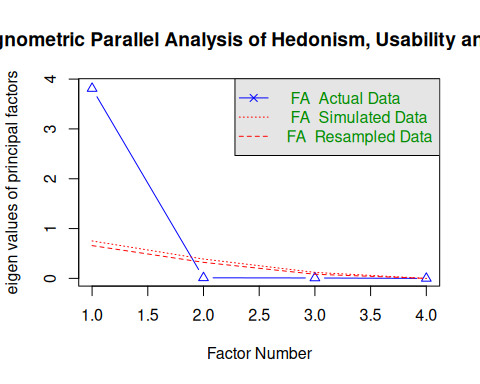
## `summarise()` has grouped output by 'Part'. You can override using the `.groups` argument.



## Parallel analysis suggests that the number of factors = 1 and the number of components = NA

parallel\_analysis(D\_HUB, 3, "D", "Hedonism, Usability and Beauty")

## `summarise()` has grouped output by 'Design'. You can override using the `.groups` argument.



## Parallel analysis suggests that the number of factors = 1 and the number of components = NA

## Discussion

Rating scales in Human Factors research are commonly used to discriminate between poor and good design options, rank designs, choose designs, or perform UX regression tests in continuous design cycles. Our logical argument is that the capability of a scale to rank designs can only be seen on multiple designs and using design-by-item response matrices. We called it the psychometric fallacy to use person-by-item response matrices in place.

A simulation showed, that the worst case can happen under the psychometric fallacy: excellent reliability is reported, when it actually is very poor.

### Dev time implications

### Run time implications

When looking at real data from commonly used rating scales, a clear bias appears, but this time the Designometric perspective significantly improves reliability. This is good news for future users of these scales, as they can reduce their sample sizes.

Past users may have scratched their heads more often than we know. Techniques such as Cronbach are commonly applied as quick sanity checks, whenever multi-item scales are used in run-time studies. Everyone in the past who used designometric scales and followed APA guidelines, was likely to the psychometric fallacy and experience some unnecessary headaches.

With better item-level reliability, experiments can also be streamlined by using a reduced set of items. Reducing a scale is basically item selection and several strategies are possible, as long as it is carried out on a design-by-item matrix.

A basic strategy is to use item-level reliability a ranking criterion. In many cases this may work well, but takes the risk of inadvertently reducing the precision in certain ranges of the scale. A well-constructed scale produces similarly precise measures in the lower, middle and high range on the whole scale. During development time this can be achieved by selecting items that are varying in strength, like “The interface is beautiful.” is stronger than “… pleasant”. By taking the average score

### Criticism of individual scales

One can rightfully argue that we used the rating scales outside their specification, as the encounters were brief and without interaction. This depends on the scale and on the purpose. Face processing is one of the fastest complex mechanisms in the human mind, which makes it legit to test it in quick succession on a screen. Similarly, the beauty judgement of a website is known to stabilize within 500ms, and the same can be expected for attractiveness. For more longitudinal feelings about designs, like hedonism, usability and credibility, the encounters were not valid. We will therefore not criticize these instruments, the point still standing that the two perspectives produce different results.

* Eeriness
* Beauty
* Attractiveness

### Future Applications

A founding idea in usability engineering is that a good designers has learned to bridge the gap between the system model and the users mental model, cognitive skills and feelings. Emerging domains are often characterized by an innovation phase, where multiple design paths are explored, before this diversity converges into an oligarchy of mature sibling frameworks with established best practices and quality principles.  
Two emerging domains of human-technology interaction are currently gaining significant momentum: interactive deep learning models, such as chatbots and social robots. Both fields have in common that they tap much stronger into the social mind of users, which, next to being a cave of snakes, is mostly uncharted terrain. We can expect a wild growth of theoretical concepts and instruments to measure the respective latent variables.

((Simone))

### Towards Deep Designometrics

By comparing the two perspectives, we illustrated that designometric analysis can fully be done with standard psychometric tools, as long as one uses the correct response matrix. However, by reducing the designometric box to a flat matrix, we loose all information on users. Formally, it would even be possible to evaluate a designometric model on the responses of a *single user*, while the situation

If the cube is collapsed to a psychometric matrix, which can be used to estimate *user sensitivity*. Legitimate cases exist to use a designometric scale for psychometric purposes. For example, an instrument to measure trustworthiness of designs could be used to estimate faithfulness levels of participants in a study (or a training) on cyber security.

By flattening the designometric box one way, then the other, we still loose information that is needed to secure that items are truly well-behaved. In educational psychometrics *differential item functioning* is the idea that items must be fair and function the same for every tested person. This also is a desirable property for a designometric scale, but a statistical model for verification would need individual parameters for participants, designs and items, simultaneously. Schmettow(2021) proposed multi-level models for capturing designometric situations in their full dimension, which could be well-suited for run-time use or basic scale development.

Another consideration is that the designometric encounter may not be end of story. For example, for comparing multi-purpose designs a researcher may want to add tasks as fourth population of interest. With the mentioned limitations, multi-level models extend to such a case (Schmettow, 2016, Egan’s assumption). For development-time purposes, Generalizability Theory may provide …

An unsolved issue is to identify exploratory methods that can operate on deep designometric data. EFA is often used with CFA to find and confirm candidate structures.