The designometric perspective and the psychometric fallacies

Martin Schmettow, Simone Borsci, Stephanie van den Berg

# Introduction

Rating scales constitute a fundamental measurement instrument in contemporary industrial applications, providing cost-effective and readily accessible methods for comparative evaluation and benchmarking of design artifacts. The utility of rating scales in decision-making processes is contingent upon two critical psychometric properties: validity and reliability.

The development of psychometrically sound rating scales represents a methodologically rigorous endeavor. Psychometrics, defined as the scientific discipline concerned with the quantitative assessment of psychological attributes to enable comparative evaluation of individual performance and functioning, has traditionally focused on cognitive abilities such as mathematical reasoning and linguistic comprehension. Subsequently, the field expanded to encompass measurement of latent psychological constructs, including personality dimensions (e.g., the Five-Factor Model).

Following the rapid expansion of user experience (UX) research, Bargas-Avila and Hoernbaek documented the proliferation of hundreds of novel rating scale instruments. However, the majority of these instruments have not undergone the rigorous psychometric validation procedures typically required for clinical assessment tools. Nevertheless, psychometric methodologies are occasionally employed during instrument development phases, particularly for reliability estimation and subscale identification through factor analysis. Additionally, practitioners implementing designometric instruments may conduct preliminary evaluations to ensure data quality and integrity.

The central theoretical proposition advanced in this investigation posits that design research instruments, specifically rating scales, function primarily to establish rank orderings among design alternatives. Consequently, the development of designometric instruments necessitates evaluation across extensive design samples to assess their ranking capabilities. This requirement results in a three-dimensional data structure (design × person × item), which is an extension of the two-dimensional response matrices (person × item) utilized in traditional psychometric analyses.

The practical contribution of this work demonstrates that psychometric analytical tools remain applicable to designometric data through dimensional reduction of the three-dimensional array into a design × item response matrix. This transformation constitutes solely a semantic reinterpretation, whereby statistical measures originally developed for ranking individuals are reapplied to ranking design artifacts.

We introduce the phenomenon of the “psychometric fallacy,” which occurs when designometric rating scales are evaluated using conventional psychometric response matrices (person × item). This approach contradicts the fundamental requirement that design samples must be incorporated into the validation process. Numerous purported designometric instruments have been developed under this fallacy, potentially compromising their capacity to effectively discriminate among design alternatives.

Given that psychometric analytical tools typically require two-dimensional input matrices, and designometric data cubes can be collapsed along either dimension to yield either designometric (design × item) or psychometric (person × item) matrices, the implications of the psychometric fallacy can be systematically investigated through comparative analyses using both matrix configurations.

This investigation establishes the designometric measurement perspective and demonstrates the appropriate application of fundamental psychometric tools for both the development and practical implementation of designometric scales. The consequences of the psychometric fallacy are further examined through secondary analysis of designometric data collected using eight widely employed UX rating scales.

## Principles of Psychometric Test Development

Psychometric instruments serve primarily as cost-effective predictive tools for human performance and psychological attributes. Psychodiagnostic rating scales function as screening instruments for conditions such as depression (Kroenke et al., 2001), while performance assessments facilitate decision-making in personnel selection and educational contexts (Schmidt & Hunter, 1998).

Three psychometric frameworks provides the theoretical framework specifying conditions under which multi-item instruments yield reliable individual rankings: Classical Test Theory (CTT) focusses mainly on the aspect of error reduction through repetition, whereas Item Response Theory (IRT) adds more rigor to item selection and Factor Analysis (FA) provides support for multi-dimensional constructs.

Unlike physiological measurements that may require only single observations (e.g., body temperature), psychological assessments necessitate multiple items due to inherent measurement error across self-report scales, reaction time measures, and physiological indicators. The central theoretical insight of CTT is that taking repeated measures improves measurement precision through error reduction, as formalized by the Law of Large Numbers.

### Item Selection and Scale Development

Scale construction begins with comprehensive domain analysis to identify relevant psychological processes and behavioral dimensions. Following domain specification, researchers develop extensive item pools that typically exceed the target scale length. This initial phase employs qualitative, divergent methodologies emphasizing content validity and theoretical coverage (Hinkin, 1998).

Subsequent item selection procedures employ quantitative criteria including item-total correlations, factor loadings, and reliability coefficients. The iterative selection process aims to optimize measurement accuracy while maintaining domain representativeness. According to CTT principles, measurement errors across multiple items cancel–provided systematic variance components demonstrate strong intercorrelation (Cronbach, 1951). While extremely repetitive items (e.g., simple reaction time trials) may achieve near-perfect agreement, practical applications typically exhibit moderate item intercorrelations.

Cronbach’s alpha (α) quantifies internal consistency by measuring inter-item agreement within the response matrix (Cronbach, 1951). Item selection procedures compare full-scale reliability against reliability estimates with individual items removed. Items whose removal improves overall reliability are identified for elimination. Similar procedures examine item-total correlations to identify poorly performing indicators (Clark & Watson, 1995).

Stepwise item removal procedures prove adequate for unidimensional constructs but may produce unstable results when multiple components comprise the measured domain. Factor-analytic methods therefore serve to identify and separate distinct components into psychometrically sound subscales (Fabrigar et al., 1999).

### Factor Structures and Latent Variable Models

Contemporary psychometric theory distinguishes between latent variables (unobservable true scores) and indicator variables (observable but imperfect measurements). Complex instruments often incorporate multiple latent variables representing distinct domain aspects. The Five-Factor Model exemplifies this approach, proposing that personality can be assessed through five primary traits, each measured via multi-item subscales (McAdams, 1992). While primary factors demonstrate relative independence by design, subscales within factors exhibit stronger intercorrelations.

Domain analysis typically suggests potential factor structures, particularly when theoretical frameworks guide instrument development. For example, mental workload assessment scales may derive from Multiple Resource Theory, which predicts independent processing of sensory modalities, thereby supporting separate subscales for each sensory channel (Wickens, 2002).

When theoretical structures exist a priori, Confirmatory Factor Analysis (CFA) provides the optimal approach for testing structural assumptions about latent variable relationships (Brown, 2015). Hierarchical CFA models can verify independence among primary scales while confirming stronger correlations among subscales. Well-fitting CFA models supersede Cronbach’s alpha by providing item loadings that inform selection decisions.

Exploratory Factor Analysis (EFA) serves to identify novel factor structures when robust theoretical frameworks are unavailable (Costello & Osborne (2005)). EFA requires researchers to specify the number of factors and their correlation structure. However, with modern resampling techniques the optimal number of factors can be identified (Lim & Jahng, 2019). Factor rotation decisions depend on theoretical expectations: orthogonal rotation applies when components are independent (e.g., mathematical and verbal abilities), while oblique rotation accommodates correlated factors typical of subscales (Fabrigar et al., 1999).

### Response Matrix Structure and Item Response Theory

Traditional psychometric methods operate on response matrices with participants as rows and items as columns. Standard procedures include computing reliability estimates on complete matrices, evaluating consistency improvements following item removal, and conducting factor analyses on participant × item data structures.

Item Response Theory (IRT) represents an alternative framework that treats response matrices as collections of person-item encounters (Embretson & Reise, 2013). Unlike CTT, IRT models both person and item parameters simultaneously, allowing formal specification and empirical testing of item characteristics. The Rasch model, representing the simplest case of unidimensional and dichotomous responses, specifies that response probability depends solely on the difference between person ability and item difficulty (Rasch, 1960).  
Advanced IRT applications include differential item functioning detection to identify and prevent measurement bias across demographic groups (Penfield & Lam, 2000).

### Sample Size Requirements in Psychometric Development

Irrespective of theoretical orientation or analytical sophistication, all psychometric instrument development activities center on establishing psychometrically sound item sets. The substantial participant samples required throughout this developmental process represent one of the most challenging aspects of psychometric research. These requirements stem from several converging methodological imperatives.

The fundamental principle governing minimum sample size requirements stipulates that observations must exceed the number of free parameters in the analytical model (Bollen, 1989). In standard Confirmatory Factor Analysis applications, each item contributes two free parameters (intercept and factor loading), requiring participant-to-parameter ratios of 5:1 to 20:1 (Brown, 2015). Contemporary simulation studies suggest 200-500 participants typically suffice for well-specified models with strong factor loadings, while complex structures may require 1000 or more participants (Wolf et al., 2013).

When theoretical structures are absent, Exploratory Factor Analysis procedures are sample-dependent, with solutions potentially capitalizing on chance characteristics. Cross-validation procedures therefore necessitate data splitting: EFA on one subsample followed by CFA on another, effectively doubling sample requirements (Anderson et al., 1988).

## Designometrics

Psychometrics as a formal theory describes how a set of differing instruments can produce a combined metric on which to compare the measured entities. Formally, it should not matter much to construct a web usability rating scale or a human skill test. Yet, there is an important difference that defines the *designometric situation* and has

First, psychometric measures form a two-way encounter, whereas comparative design studies has Design as an additional entity, forming a *response box*. Second, in psychometrics the entity to be ranked is Person, whereas designometric applications Designs are compared. Whatever role Person parameters play in psychometric processes, must now be assigned to the Design parameters.

### The designometric perspective

A practical implication of the designometric perspective is that a designometric “response box” cannot be processed with psychometric tools. While “deep” designometric models can be constructed by multi-level models [Schmettow, dmx], a practical solution exists. By averaging over one of the dimensions, a two-dimensional response matrix can be constructed. Crucially, the correct dimension to be collapsed over is Person, as this produces a *designometric response matrix* (design x item), which in turn is needed to assess the item properties with respect to ranking designs.

For a reliable designometric scale, its items must inter-correlate strongly, which can only happen when referred-to design features coincide. Take as an example a hypothetical scale to measure trustworthiness of robot faces, with two sub-scales, Realism and Likability. The impression of realism can be triggered by different features of the face, such as skull shape, details in the eyes region and skin texture. For a proper scale on realism, it would be required that these features correlate, and this essentially is a property of the robot face design process. It is a quite strong assumption that the effort a robot face designers puts into the eyes region must be strongly correlated with the effort put into skin texture, but by using psychometric models with designometric data, assumptions like these can be tested.

### Designometric scale development

Analog to psychometrics substantial samples of designs are required for item selection and factor identification. This can be a huge problem, depending on the class of designs. For e-government websites it will be easier compared to a scale dedicated to human-like robots or self-driving cars.

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 sample of 100 designs can be tested by letting every participant encounter overlapping subsets. 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 large language models broke through (Borsci & Schmettow, 2024).

### The psychometric fallacy

Designometric scales can be developed with psychometric tools, if using design x item matrices with sufficiently large sample of designs. 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*.

A purportedly designometric study instrument that has been validated on 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.

If a substantial sample of designs has been collected, a correct designometric response matrix can be created by averaging over Persons. However, the use of standard psychometric tools may mislead the researcher to believe that producing a psychometric matrix is correct.

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.

Recall that psychometric validations require large samples of participants! So, even validation studies that included multiple designs, may in practice not be repairable, because the sample of designs is too small for a proper analysis.

One example is the study by Ho & MacDorman (2017) validated the Godspeed Index, a common multi-scale inventory to evaluate robot designs. Their designometric study included 38 items and 30 participants, but only 12 designs. While they did not specify how the designometric box was collapsed, the fact that they were still able to report exporatory factor analysis results, suggests that they used a psychometric response matrix, as the designometric matrix would have been too small. To be fair, the study tested validity correctly comparing designs, although with simple ANOVA models.

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 test reliabilities, 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 cause confusion 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 investigation, data from past experiments was used to assess how strongly several existing scales are compromised.

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

# 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 because it were all homepages of premium law firms.

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 | 0.0538197 | -0.7678047 | 0.4908936 |
| Coolness | psychometric | 0.9284072 | 0.8760182 | 0.9620385 |

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.

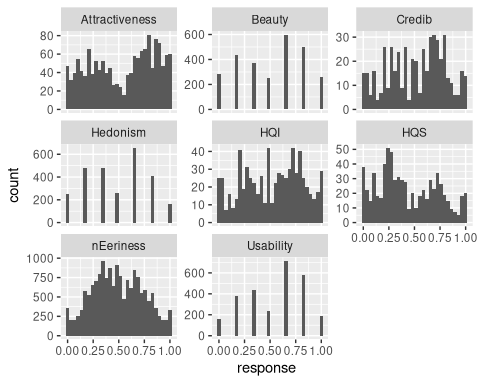
# Methods

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.



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()

| 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()

| 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()

| 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

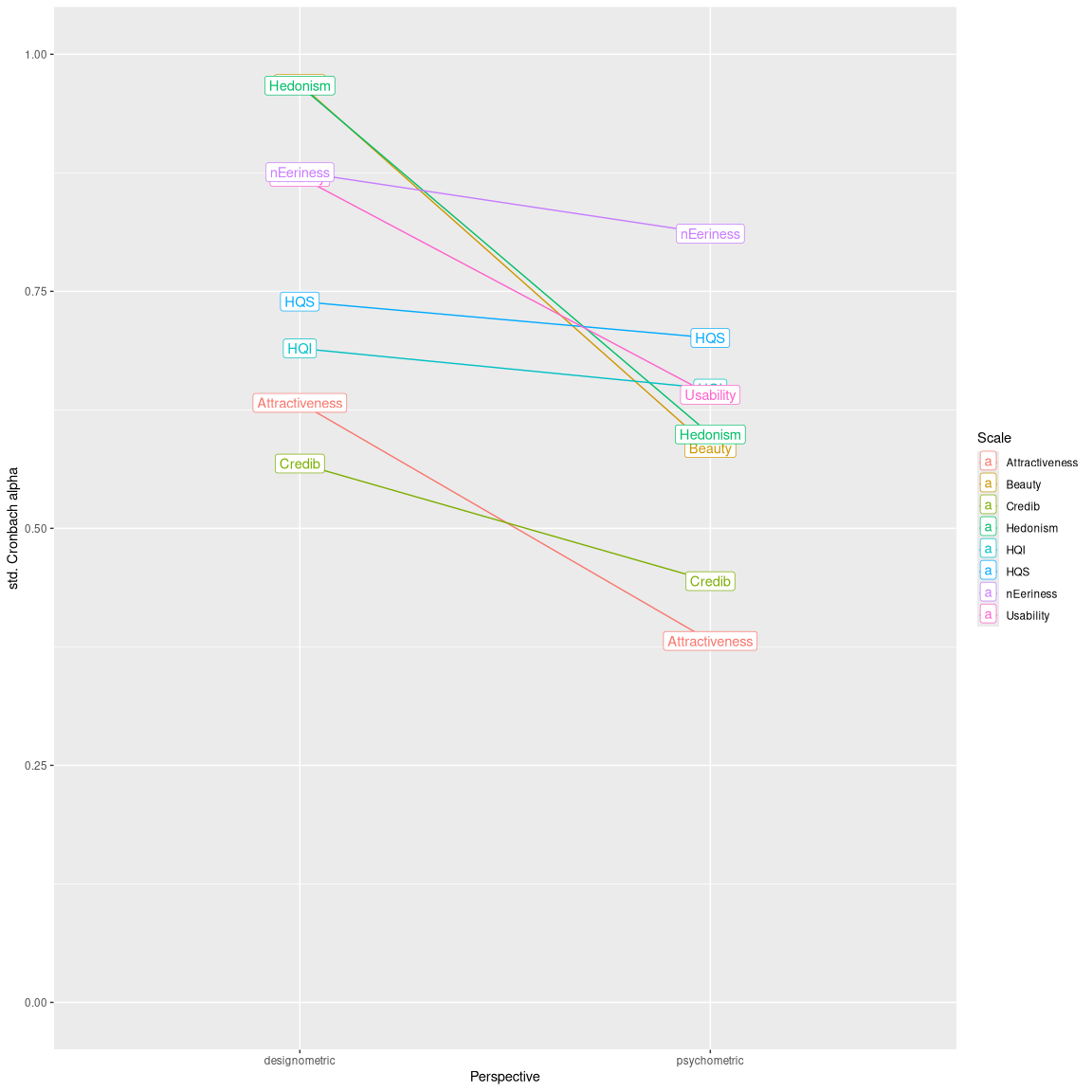
Goal of the analysis is to examine in how much the psychometric fallacy creates real biases. For this purpose, three basic psychometric techniques were applied to several data sets. Scale reliability was measured using Cronbach . For item consistency, the corrected item-total correlation was used and for the number of factors parallel analysis was applied, 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 reliability

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 | Perspective | center | lower | upper |
| --- | --- | --- | --- | --- |
| Attractiveness | designometric | 0.6323901 | 0.4402591 | 0.7707267 |
| Attractiveness | psychometric | 0.3812511 | 0.0401928 | 0.6494977 |
| Beauty | designometric | 0.9687384 | 0.9584027 | 0.9777105 |
| Beauty | psychometric | 0.5847188 | 0.4213651 | 0.7474795 |
| Credib | designometric | 0.5683567 | 0.2490173 | 0.7532719 |
| Credib | psychometric | 0.4443016 | -0.0448150 | 0.7233751 |
| HQI | designometric | 0.6897988 | 0.5024870 | 0.7938411 |
| HQI | psychometric | 0.6474037 | 0.2302728 | 0.8598254 |
| HQS | designometric | 0.7392031 | 0.6187896 | 0.8239894 |
| HQS | psychometric | 0.7009069 | 0.4392283 | 0.8396887 |
| Hedonism | designometric | 0.9670416 | 0.9506223 | 0.9785137 |
| Hedonism | psychometric | 0.5991254 | 0.3480033 | 0.7587976 |
| Usability | designometric | 0.8707024 | 0.7972996 | 0.9226968 |
| Usability | psychometric | 0.6409583 | 0.4211647 | 0.7845830 |
| nEeriness | designometric | 0.8756717 | 0.8411502 | 0.9078357 |
| nEeriness | psychometric | 0.8108672 | 0.7131149 | 0.8767375 |

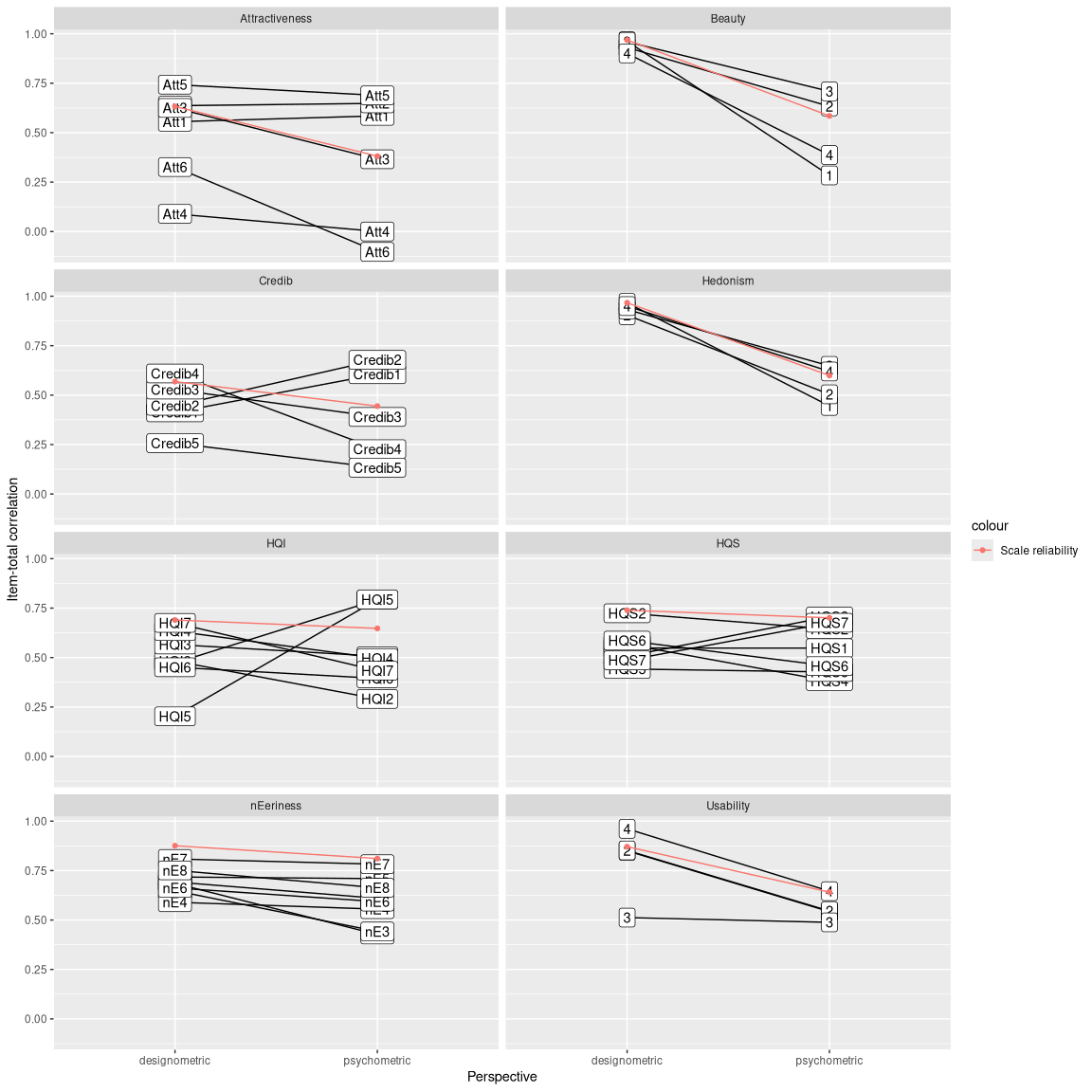


Cronbach alpha item-level reliability estimates compared by perspective and scale

**?@fig-scale-consistency** shows the Cronbach scale reliability estimates produced by designometric and psychometric response matrices. Overall scale reliabilities cover a broad range from excellent to unusable. All scale reliabilities improve under the designometric perspective, albeit, the difference ranges from barely noticable (HQS, HQI) to very strong (Hedonism, Usability, Beauty and Attractiveness). The most dramatic difference can be seen in Hedonism and Beauty, which both have excellent dmx reliability, which drops to an almost unusable level under pmx.

## Item consistency

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



Cronbach alpha item-level reliability estimates compared by perspective and scale

**?@fig-item-reliability** shows the corrected item-total correlations as a measure for item consistency. Beauty and Hedonism stand out, because all items take a similar sharp drop in reliability under pmx. To some extent this also seems to hold for Usability and Eeriness. For Credibility, HQ-I, HQ-S and Attractiveness some items drop under pmx, whereas others improve, with two extreme cases: Items Att4 andf Att6 are already on a very low level of reliability under dmx, but under pmx, they even become negatively correlated. Items HQI5 and HQI6 perform poorly under dmx, but are among the overall best performing items under pmx.

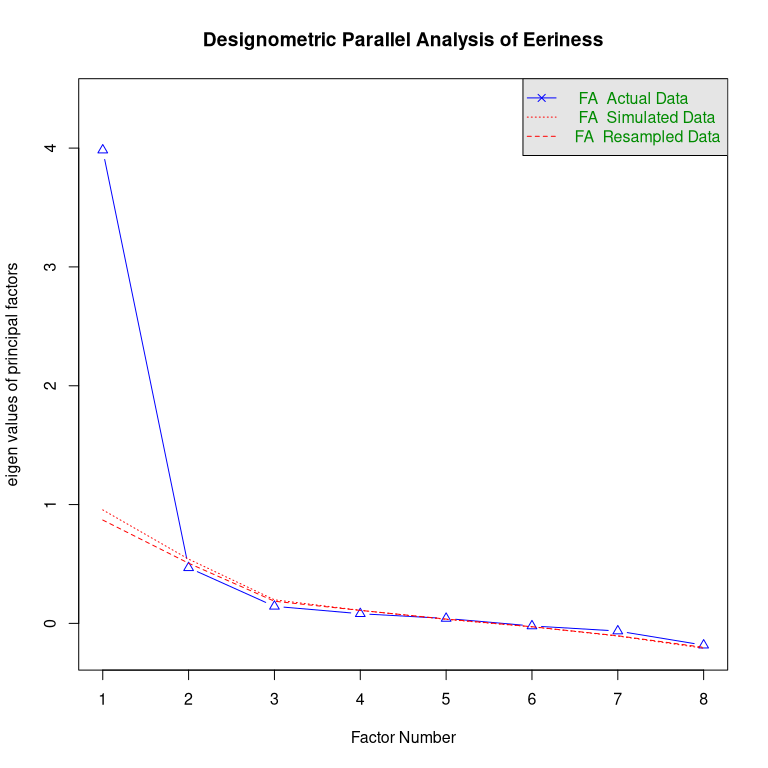
## Number of factors

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.

In contrast, the AttrakDiff2 questionnaire comprises two scales to capture supposedly different aspects.

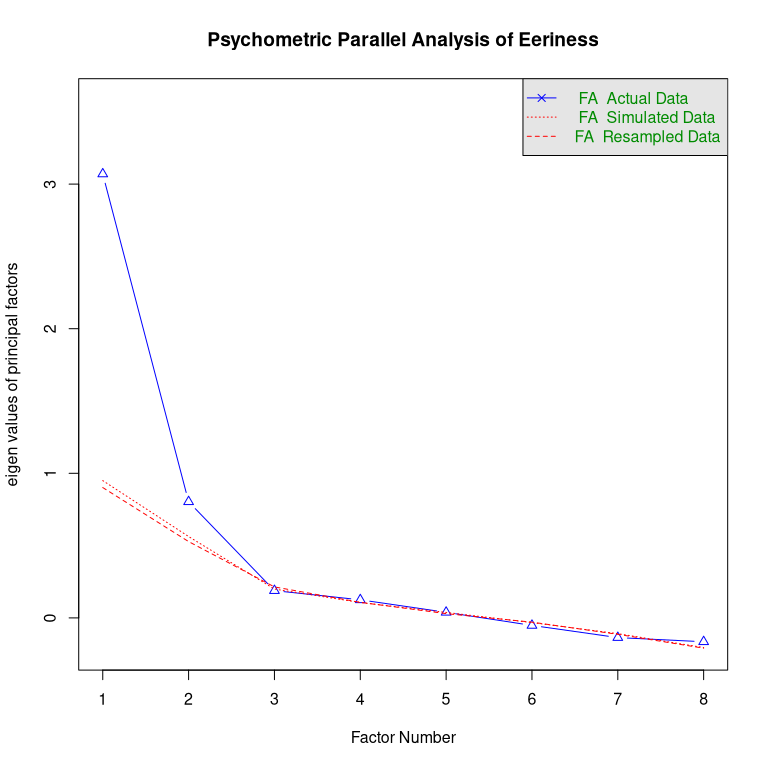
Given a response matrix, the number of factors can be estimated using parallel analysis. Ideally, this procedure returns exactly as many factors as there are separate scales. Here, we use parallel analysis to assess whether the two perspectives produce the expected number of factors, or at least agree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

((MacDorman)) found that the Eeriness scale decomposes into two slightly different aspects, summarized as “eery” and “spine-tingling”.



Suggested number of factors for the Eeriness scale compared by perspective

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

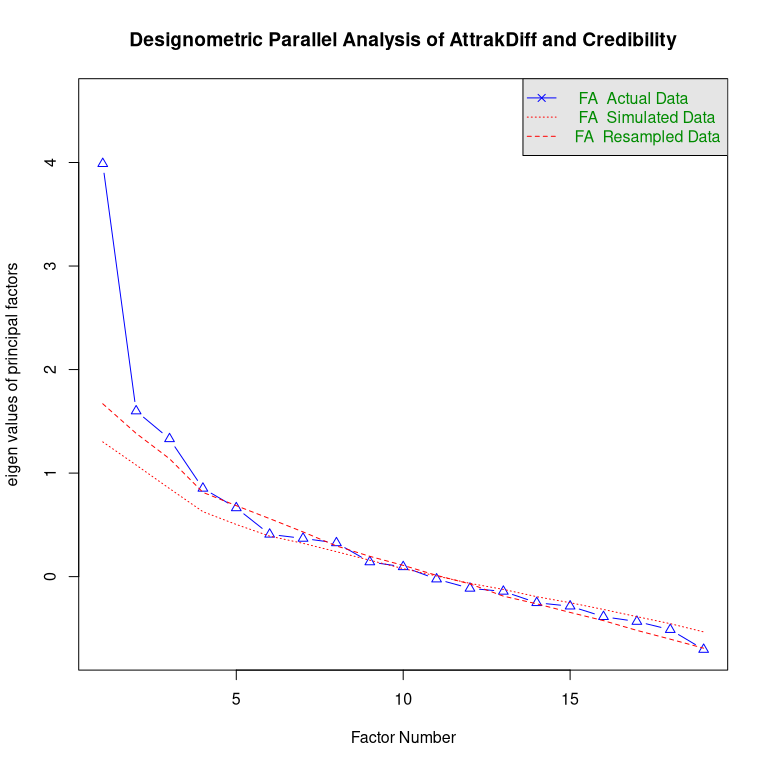


Suggested number of factors for the Eeriness scale compared by perspective

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

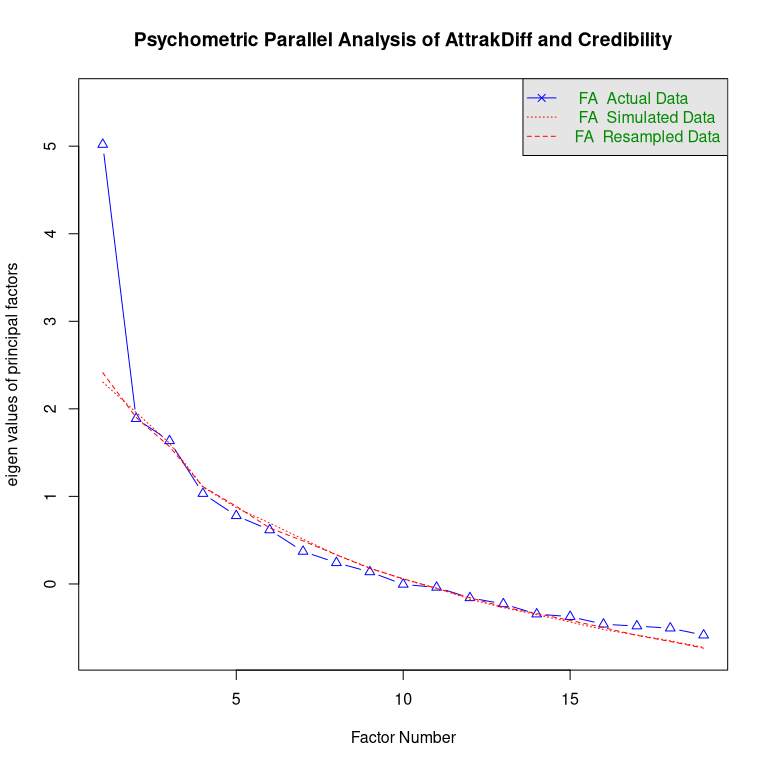
The results suggest that under dmx only one latent variables exists, whereas pmx produces two.

On theoretical grounds, the AttrakDiff2 inventory splits hedonistic quality into two components, Identity and Stimulation, while the credibility scale is a completely separate construct. We would expect three factors to emerge.



Suggested number of factors for AttrakDiff inventory plus Credibility compared by perspective

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

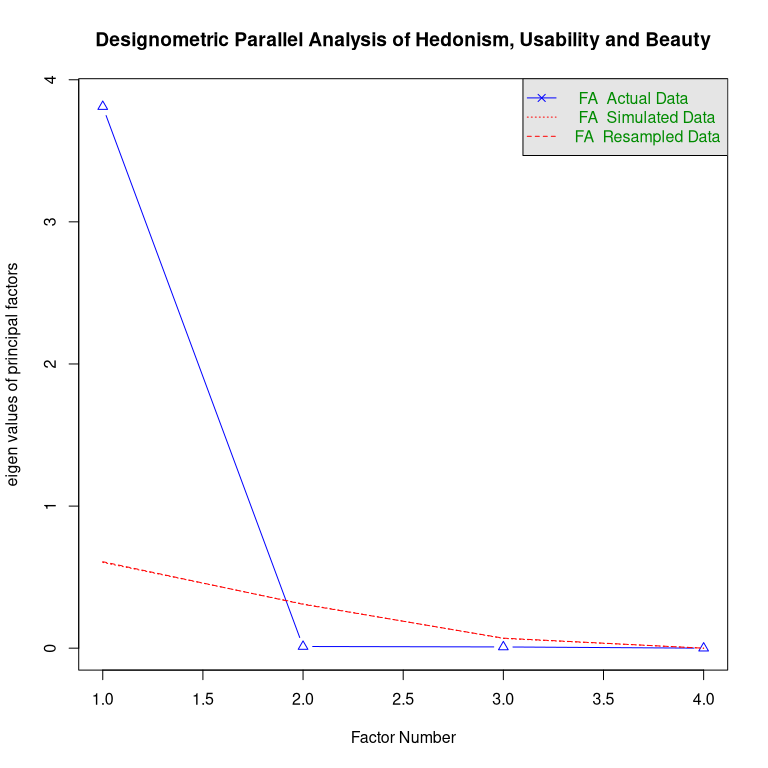


Suggested number of factors for AttrakDiff inventory plus Credibility compared by perspective

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

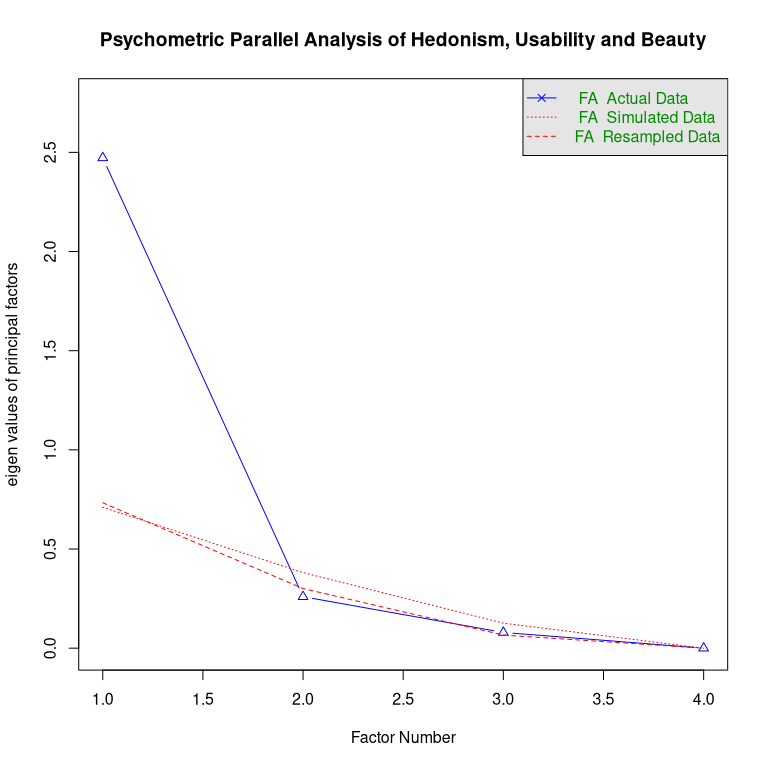
Under a desginometric perspective, the three scales have five underlying factors, but merging into one under pmx.

Finally, in study DN three independent scales, Hedonism, Usability and Beauty, were used. But, parallel analysis suggests that these capture the same latent variable under both perspectives.



Suggested number of factors for Hedonism, Usability and Beauty scales compared by perspective

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



Suggested number of factors for Hedonism, Usability and Beauty scales compared by perspective

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.

We used data from five experiments to assess the severity of the psychometric fallacy in real practice. Overall, the results show that the psychometric fallacy produces dramatic biases for all tested methods. For practicioners using these scales the good news is that all scales performed better under the correct desgnometric perspective, and most of them even fairly well.

In contrast, item consistency and factor analysis showed that the psychometric fallacy can lead to strong biases. Items can suddenly become negatively correlated, as in the case of Attractiveness. The two Hedonism scales from AttrakDiff and the Credibility scale showed an extreme pattern, where the majority of items remained relatively stable, whereas two items switched from poor reliability to excellent reliability under pmx. With these patterns in mind, it is almost not surprising that factor analysis can also produce quite different results. Most strikingly, under dmx not a single result matched the theoretical expectations.

## Dev time implications

In design research the target of all research is quickly changing and expanding target. A certain swiftness and pragmatism is required to keep up with the pace. Development of new scales is a common task, and often it is carried out by researchers with a basic understanding of psychometric principles, such as (item) reliability and exploratory factor analysis.

Basic psychometric tools produce vastly different results under the psychometric fallacy. While our study used mature scales, which had already undergone item selection and perhaps factor analysis, we can interpolate the consequences for future scale development.

The most severe consequence is that a scale may be developed that is not capable of ranking designs. According to an often cited rule-of-thumb, scale reliability should be at least .7. Three scales in our study, Attractiveness, Credibility and HQS did not meet this criterion, even under the designometric perspective.

Interestingly, on item level Credibility and HQ-I show the same pattern, where two items perform well under pmx, but extremely poor under dmx. This may be a co-incidence, or the result of developing a dmx scale under pmx is to *false favor* items that are well-behaved in ranking persons, but are inefficient for designs. To make the case, (**Tab-rel-after-removal?**) compares dmx scale reliability on these three scales with and without their two ill-behaved items.

[1] "TODO"

A pmx perspective may also *false reject* items that are actually well-behaved in ranking designs. Creating an item pool is by itself a time-consuming process, and the psychometric fallacy can make it even more difficult by unnecessarily rejecting items. A possible example is the development of the BUS-11 scale, where face validity demands (and factor analysis has confirmed) ((REF Simone)) that *Privacy* is a separate construct. Unfortunately, only one item was left after item selection.

## Run time implications

For practitioners, the good news are that if they were under the run-time psychometric fallacy by routinely reporting scale reliability, they were always better than they said. And when they continue to use these scales in the future, the improved precision will allow them to reduce sample sizes.

But, practitioners may not yet have the most efficient rating scales. Even if a false favored item is not directly harming reliability, it can make the scale inefficient. In practice, UX scales are often deployed during use, for example in usability tests. With a shorter scale measures can be taken in quicker succession, for example once per task, or everyday in a longitudinal study. It is therefore not uncommon for practitioners to create a reduced scale, for example, when many latent variables are involved. For some scales (Hedonism, Beauty) it is safe to just pick three items at random. Other scales are quite mixed bags, with the highest ranked item under pmx being the lowest ranked under dmx.

## Future Applications

A key idea in usability engineering is that interaction designers must learned to bridge the gap between the system model and the users mental model, cognitive skills and feelings. Emerging technologies are often characterized by an innovation phase, where multiple design paths are explored in a rush to the market and several domains of human-technology interaction are currently gaining momentum: large language model technology is, as of writing, causing much attention for intelligent agent design. At the same time, humoid and animalistic robot design is coming out of its niche, and virtual reality applications are on their way to mainstream. These three domains have in common that, compared to classic computer applications, they are tapping into new territories of the users mind, the social mind and the sensation of physical reality. When Social Experience (SX) or Virtual Experience (VX) become the new UX, it may start with the same abundance of new instruments trying to map the uncharted design space. From a practical perspective, our results suggest that the psychometric fallacy can be harmful in the item selection process. Design researchers

## 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.

# References

Anderson, J. C., Kellogg, J. L., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411–423.

Borsci, S., & Schmettow, M. (2024). Re-examining the chatBot usability scale (BUS-11) to assess user experience with customer relationship management chatbots. *Personal and Ubiquitous Computing*, *28*, 1033–1044. <https://doi.org/10.1007/s00779-024-01834-4>

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (p. 482). The Guilford Press.

Clark, L. A., & Watson, D. (1995). Scale-validity. *Psychological Assessment*, *7*.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*. https://doi.org/<https://doi.org/10.7275/jyj1-4868>

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. <https://doi.org/10.1007/BF02310555>

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press. <https://doi.org/10.4324/9781410605269>

Fabrigar, L. R., Wegener, D. T., Maccallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*, 104–121. <https://doi.org/10.1177/109442819800100106>

Ho, C. C., & MacDorman, K. F. (2017). Measuring the uncanny valley effect: Refinements to indices for perceived humanness, attractiveness, and eeriness. *International Journal of Social Robotics*, *9*, 129–139. <https://doi.org/10.1007/s12369-016-0380-9>

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*, 606–613. <https://doi.org/10.1046/j.1525-1497.2001.016009606.x>

Lim, S., & Jahng, S. (2019). Determining the number of factors using parallel analysis and its recent variants. *Psychological Methods*, *24*, 452–467. <https://doi.org/10.1037/met0000230>

McAdams, D. P. (1992). The five‐factor model personality: A critical appraisal. *Journal of Personality*, *60*, 329–361. <https://doi.org/10.1111/j.1467-6494.1992.tb00976.x>

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, *19*, 5–15. <https://doi.org/10.1111/J.1745-3992.2000.TB00033.X>

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. In *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* (pp. 184, xiii, 184–xiii). Nielsen & Lydiche.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*, 159–177. <https://doi.org/10.1080/14639220210123806>

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*, 913–934. <https://doi.org/10.1177/0013164413495237>