The Designometric Perspective and the Psychometric Fallacy in Developing User Experience Rating Scales

Martin Schmettow Simone Borsci Stéphanie M. van den Berg

Rating scales are widely used in user experience (UX) research to compare and rank design alternatives, yet their development typically relies on psychometric procedures originally created for assessing stable individual traits. We argue that this practice commits the psychometric fallacy: evaluating designometric instruments—tools intended to discriminate between designs—using person × item response matrices that exclude the essential third dimension of design. Because design evaluation inherently forms a design × person × item data structure, valid scale development requires samples of designs large enough to assess how well items rank design alternatives. We show that collapsing the data cube along persons yields a design × item matrix that allows psychometric tools to be applied meaningfully, whereas collapsing along designs yields misleading psychometric results about person sensitivity rather than design discriminability. A simulation study demonstrates that scales can appear highly reliable psychometrically while being effectively unusable for ranking designs. Secondary analyses of eight commonly used UX scales across large design samples reveal systematic distortions in reliability estimates, item performance, and dimensionality when instruments are evaluated under the psychometric fallacy. We establish the designometric measurement perspective, provide principles for proper designometric scale development, and illustrate how misuse of psychometric validation leads to faulty instruments and incorrect inferences in UX research.

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 critical psychometric properties of the scale, such as item consistency, reliability and factor structures.

The development of effective rating scales is a methodologically rigorous endeavor. Psychometrics, 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 & Hornbæk (2011) 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 sub scale identification through factor analysis. Additionally, practitioners implementing designometric instruments may conduct preliminary evaluations to ensure data quality and integrity.

The central theoretical argument 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 $\times$ person $\times$ item), which is an extension of the two-dimensional response matrices (person $\times$ item) used in traditional psychometric analyses.

The phenomenon of the "psychometric fallacy" occurs when designometric rating scales are evaluated using conventional psychometric response matrices (person $\times$ 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.

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

Given that psychometric analytic tools typically require two-dimensional input matrices, and designometric data cubes can be collapsed along either dimension to yield either designometric (design $\times$ item) or psychometric (person $\times$ 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.

In the following section we present a short recap on psychometric principles and workflows, after which we derive principles of proper designometric scale development and discuss the psychometric fallacy. While we believe that the theoretical argument can stand on its own, we next present a small simulation study to show the mechanics, followed by an illustration of the real-world consequences of the psychometric fallacy on empirical data, obtained from several commonly used UX rating scales. Last, we discuss more sophisticated methods of dealing with design evaluation data.

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 schools of thought provide the theoretical framework 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.

Item Selection and Scale Development

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. In practice, this is achieved by using sets of items.

Multi-item 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).

Cronbach's 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).

Factor Structures and Latent Variable Models

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

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. As an example, the Five-Factor Model proposes 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 often 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 channels, thereby supporting a separate subscale for each channel (Wickens, 2002).

When theoretical structures exist a prior, 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.

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. This is inconvenient because an unknown factor structure implies an unknown number of factors.

Finding the number of factors is possible by successively increasing the number of factors as long as a factor retention criterion is met. Common criteria include the Kaiser-Guttman rule (eigenvalues > 1) and scree plot inspection, but these have been criticized for being to lenient and subjective (Hayton et al., 2004). A more accurate alternative is parallel analysis, which

compares observed eigenvalues to those obtained from data of identical size but randomized through resampling (Lim & Jahng, 2019).

Another decision to make is factor rotation which depends 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 $\times$ 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 analytic 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.

To begin with, samples must sufficiently represent the population in question, with the general rule that heterogeneous populations require larger sample sizes. During statistical analysis, sample size must at least match the number of free parameters in the analytic model to ensure identifiability, but in practice this is usually not sufficient (Bollen, 1990). For example, in Confirmatory Factor Analysis, 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 can be used to identify suitable subscales. However, these procedures are sample-dependent, with solutions potentially capitalizing on chance or researchers' degrees of freedom (Simmons et al., 2011). 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*: 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 psychometric tools take flat response matrices as input and are unfit to process higher-dimensional data. While "deep" designometric models can be constructed using multi-level models (Schmettow, 2021b, pp. 307–323), a practical solution exists to put psychometric tools to use. By averaging over Person, a two-dimensional response matrix can be constructed from a designometric box. 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 psychometric scale evaluation, 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 measurement can take from several minutes to hours and a complete experimental design with every possible encounter becomes impractical. A way to mitigate this problem is to use *planned incomplete* experimental designs. 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 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 standard terminology in psychometric tools may still 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 aesthetics. 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! When swapping roles, validation studies that included multiple, but only a few, designs, may not be repairable. One

example is the study by Ho & MacDorman (2010) validating 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 exploratory factor analysis results, suggests that they used a psychometric response matrix, as the designometric matrix would have been too small to produce stable estimates. 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 existing 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 often require to report test reliability, it may be tempting for the researcher to run a psychometric test. 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.

Simulation study

The following example demonstrates the difference by simulating a situation, where a fictional three-item scale for Coolness is highly reliable for persons, but seemingly has no reliability at all for discerning the tested designs. Such a pattern can occur when the design sample bears little variance with respect to the property in question. In the following simulation, we assume that the Coolness scale has been tested on a sample of 50 premium law firm home pages and 50 participants of various ages and social background.

The simulation uses zero-centered Normal distributions to draw the parameters for 20 design, 20 participant and 4 items. Subsequently these are combined into responses $R = D + P - I$ with some extra noise ($\sigma_{\text{Part}} = .5$). The key here is that items and participants vary strongly in their appreciation of Coolness ($\sigma_{\text{Part}} = \sigma_{\text{Item}} = .2$), whereas the sample of designs varies much less in Coolness ($\sigma_{\text{Design}} = .05$).

Table 1: Reliability under both perspectives using simulated data (fictional Coolness scale)

Scale	Perspective	center	lower	upper
Coolness	designometric	0.63	0.37	0.79
Coolness	psychometric	0.93	0.85	0.97

This simple example demonstrates that psychometric reliability (person sensitivity) can be excellent (.93, see Table 1), whereas designometric reliability is poor (.60). In the following study we examine how severe the psychometric fallacy is in real practice.

Empirical demonstration

In order to assess the biases introduced by the psychometric fallacy, a secondary data analysis was conducted using data from seven prior experiments, which were originally testing original hypotheses on User Experience and Human-Robot Interaction (Table 2). What was common is that data was obtained in complete designometric encounters, with large samples of designs.

Methods

In QB, JK, SP and DN participants saw screen shots of home pages and responded to several user experience scales, whereas in AH, DK and PS the stimuli were robot faces (see Table 2 for abbreviations). All experiments used manipulation of presentation time to collect data on subconscious cognitive processing. For the analysis here, presentation times lower than 500ms were discarded. All experiments used a (mildly) incomplete design in that participants encountered all designs and items several times, but not in every possible combination.

DN used OpenSesame for stimulus presentation and implemented the collection as graded responses (Mathôt et al., 2012). All other experiments used the same PsychoPy program and collected continuous responses using a visual analog scale (Peirce, 2008).

Table 2: Summary of data sets used for analysis

ID	Scale	Designs	N_{Design}	N_{Item}	N_{Part}	N_{Obs}	Reference
SP	Attractiveness	Homepages	66	6	40	1440	Polst & Schmettow (2014)
DN	Beauty	Homepages	48	4	42	2688	Nazareth & Schmettow (2014)
JK	Credib	Homepages	76	5	25	500	Kuurstra & Schmettow (2013)
QB	HQI	Homepages	76	7	25	700	Boom & Schmettow (2013)
QB	HQS	Homepages	76	7	25	700	Boom & Schmettow (2013)
DN	Hedonism	Homepages	48	4	42	2688	Nazareth & Schmettow (2014)
DN	Usability	Homepages	48	4	42	2688	Nazareth & Schmettow (2014)
AH	nEeriness	Robot faces	20	8	45	10800	Haeske & Schmettow (2016)
DK	nEeriness	Robot faces	80	8	35	2800	Keeris & Schmettow (2016)
PS	nEeriness	Robot faces	87	8	39	2808	Slijkhuis & Schmettow (2017)

In total, eight rating scales were applied to four design samples. The bipolar *Eeriness* scale is a primary research tool on the Uncanny Valley phenomenon and measures negative emotional responses towards artificial faces (see Section). Factor structure and reliability were originally established under psychometric perspective (Ho & MacDorman, 2017). AH used morphing levels between human and robot faces as stimuli, whereas the other two experiments used a subset of Mathur & Reichling (2016), with a few new designs added.

All other scales were applied a sample of commercial home pages collected by Tuch et al. (2012) (a subset in DN):

- The *Attractiveness* scale is unipolar subscale of the User Experience Questionnaire (UEQ) inventory and has undergone basic psychometric evaluation in six studies with a single design each (Theo et al., 2008).
- The two 7-item bipolar scales *Hedonic Quality - Identity (HQI)* and *Hedonic Quality - Stimulation (HQS)* are from the AttrakDiff2 inventory, which underwent primary psychometric validation on three designs (Hassenzahl et al., 2003). The scales *Hedonism* and *Usability* in Nazareth & Schmettow (2014) were taken from the short version of AttrakDiff2, but will be considered separately in this analysis.
- DN composed the Beauty scale from the item used in Hassenzahl & Monk (2010) and three items representing classic aesthetics from Tractinsky et al. (2006).
- The *Credibility* scale was originally designed to compare people's attitude towards media (newspapers, TV, radio). The validation study used exploratory factor analysis for psychometric item selection with 1468 participants (McGrath & Gaziano, 1986).

Participants were sampled by convenience with sizes between 25 and 45 and a strong over-representation of university-level Social Sciences students and associated circles.

Goal of the analysis was to examine how the psychometric fallacy compromises the evaluation of rating scales. For this purpose, each of the data sets was collapsed into a psychometric and a designometric response matrix. Subsequently, three basic psychometric techniques were applied to both perspectives and compared. Scale reliability was computed using Cronbach's alpha. Item consistency was studied by inspecting corrected item-total correlations (William Revelle, 2025).

The number of factors (dimensionality) was determined based on a parallel analysis based on minimized residuals, using the fa.parallel function from the psych R package. *Number of factors* was identified using parallel analysis with `psych::fa.parallel`. This produces an eigenvalue obtained on real data and compares it to eigenvalues obtained from simulated data obtained from randomized data. The number of factors is determined as the point before the the real eigenvalue drops below the simulated level.

Results

Scale reliability

Overall scale reliabilities cover a broad range from excellent to unusable Figure 1. All scales look better under the designometric perspective, albeit, the difference ranges from barely noticable (HQS, HQI) to very strong (Hedonism, Usability, Beauty and Attractiveness). The

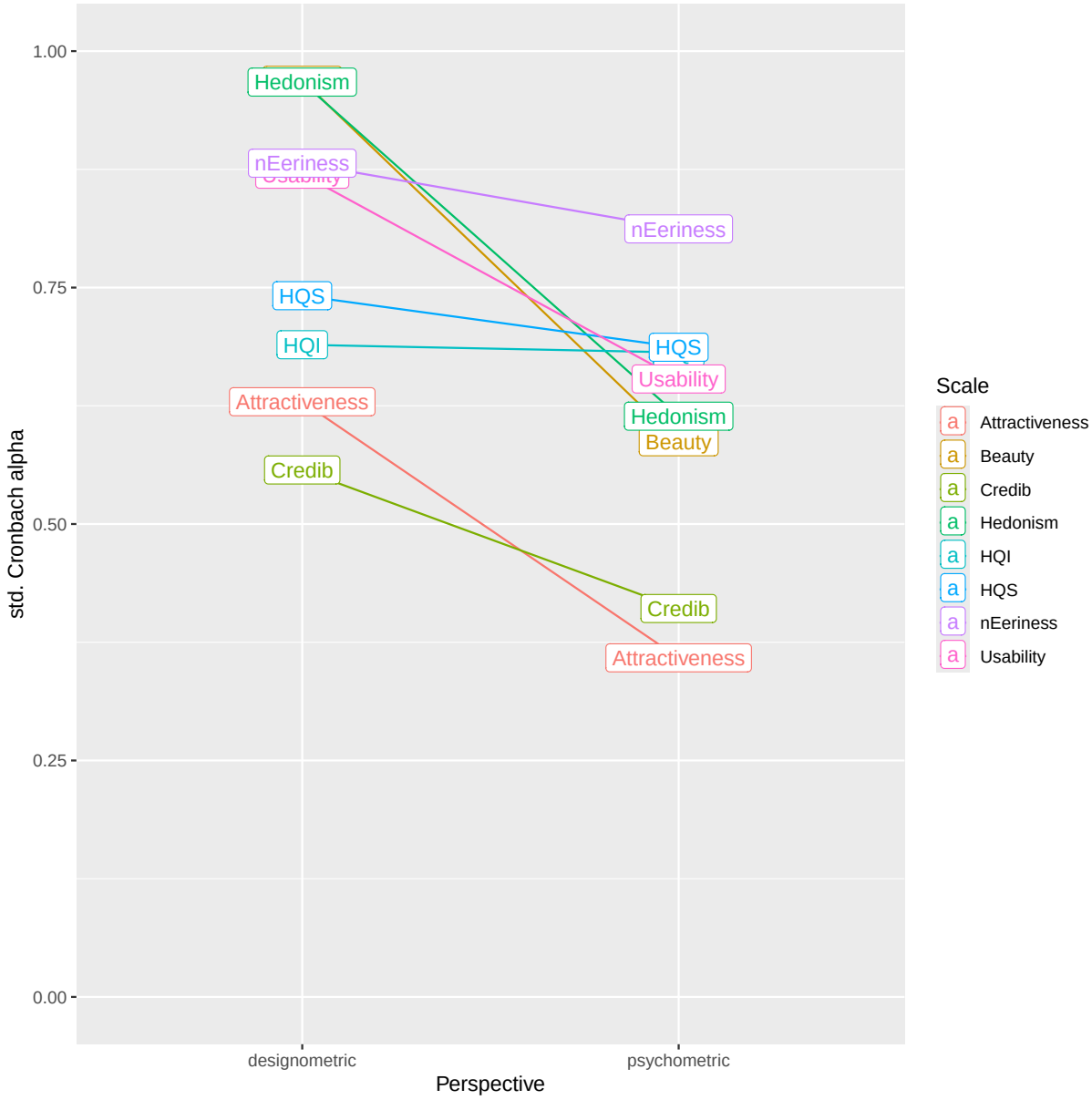

Figure 1: Cronbach alpha item-level reliability estimates compared by perspective and scale

most dramatic difference can be seen in Hedonism and Beauty, which both have excellent designometric reliability, but poor reliability from a psycvhometric perspective.

Item consistency

Figure 2 shows corrected item-total correlations reflecting item consistency. Beauty and Hedonism stand out, because all items show higher designometric item reliabilities. To some extent this also seems to hold for Usability and Eeriness. For Credibility, HQ-I, HQ-S and Attractiveness some items perform poorer under the psychometric perspective, whereas others improve, with one extreme cases: Item Att6 is already on a very low level on designometric performance, showing even a negatively correlation under the psychometric perspective. Items HQI5 and HQI6 show poor designometric performance, but are among the overall best performing psychometric items.

Number of factors

Given a response matrix, the number of factors were estimated using parallel analysis. Ideally, this procedure returns exactly as many factors as there are separate scales in every data set.

The Eeriness scale is part of a larger Godspeed Index inventory and is supposed to represent a single latent variable. However, Ho & MacDorman (2017) found that the scale decomposes into two slightly different aspects, summarized as “eerie” and “spine-tingling”. This was established using principal component analysis, whereas a dedicated identification of the number of factors has not been reported. Since this was tested with only 12 designs, most likely it was under psychometric perspective. The results in Figure 3 show that for both perspectives the eigenvalue drops below the simulated eigenvalue with two factors under both perspectives.

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. As Figure 4 shows the two perspectives deviate in opposite directions: for the psychometric perspective, the eigenvalues drop below their simulated counterparts at two factors, whereas for the designometric perspective stays above this line with five factors.

Finally, in study DN three independent scales, Hedonism, Usability and Beauty, were used und we expect three factors. In contrast, the eigenvalues drop below even the simulated eigenvalues with two factors, suggesting that the same latent variable is captured by all three scales under both perspectives (Figure 5).

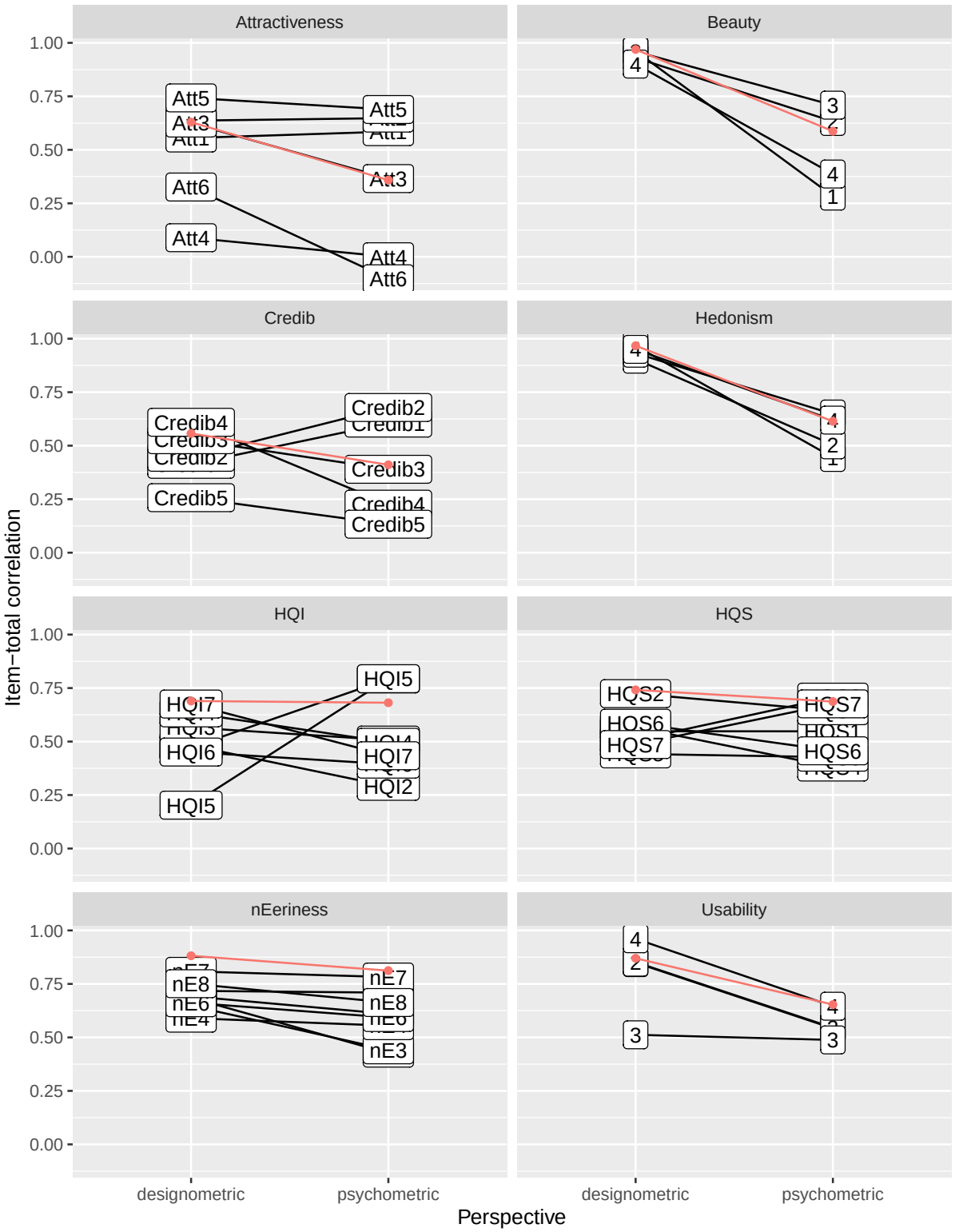

Figure 2: Cronbach alpha item-level reliability estimates compared by perspective and scale

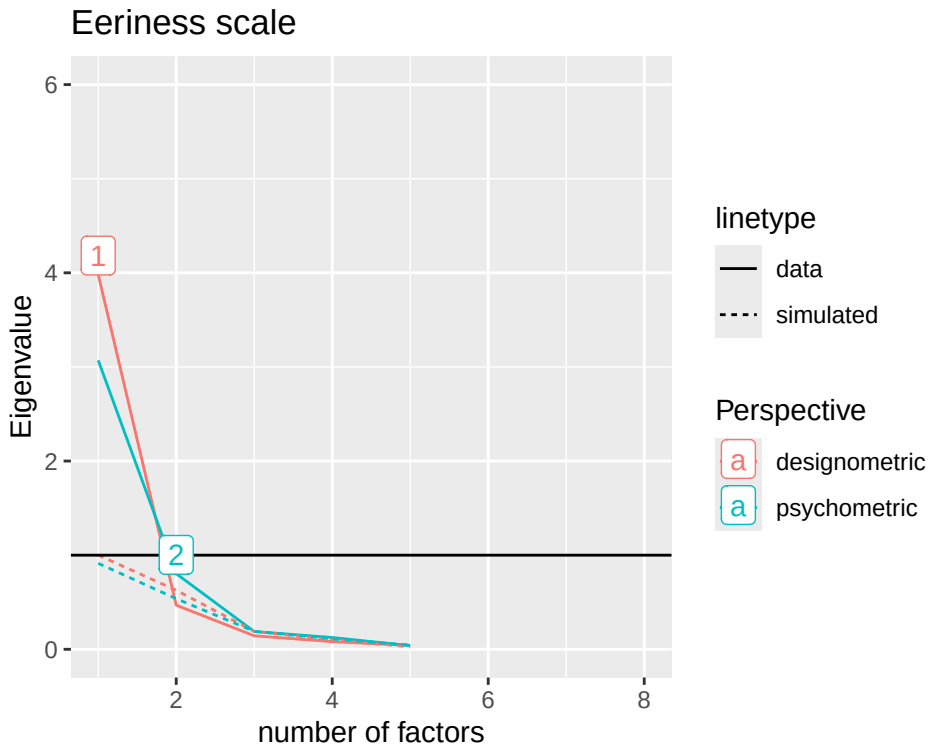

Figure 3: Number of factors under designometric and psychometric perspectives for the Eeriness scale using parallel analysis

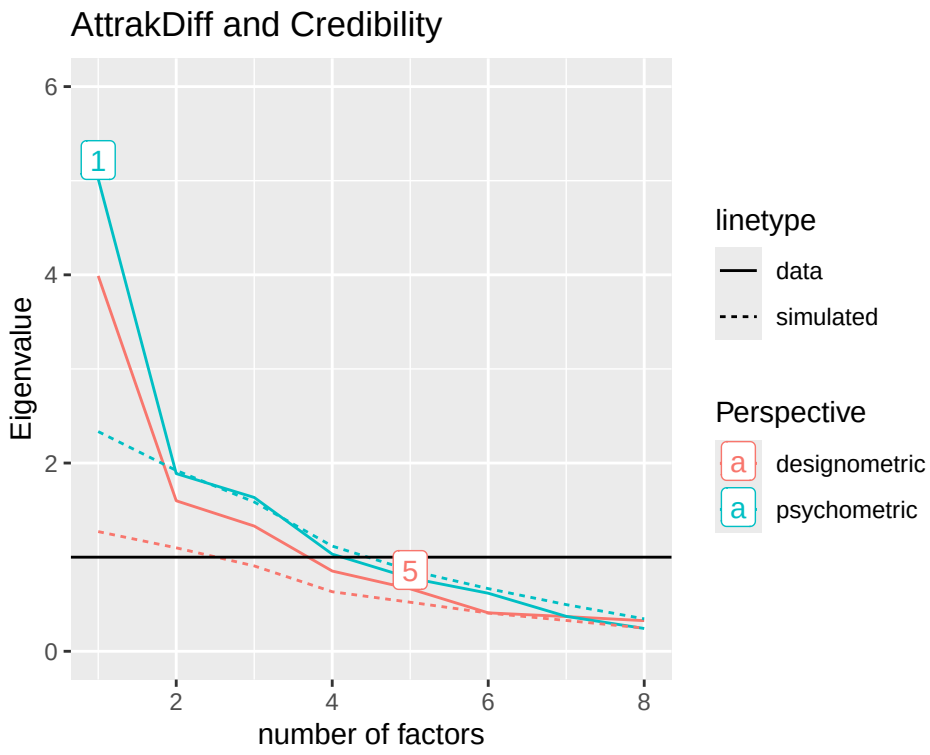

Figure 4: Number of factors under designometric and psychometric perspectives for the AttrakDiff inventory using parallel analysis

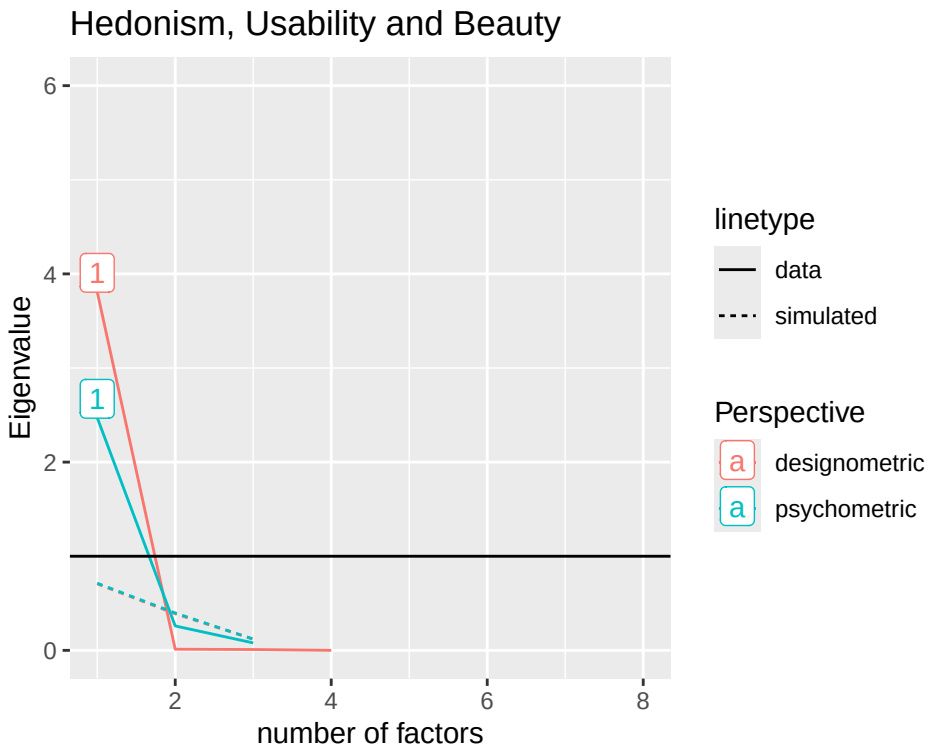

Figure 5: Number of factors for Hedonic value, Usability and Beauty using parallel analysis

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 poor.

With empirical data from five experiments we showed that the psychometric fallacy has real-world implicitations and produces wrong interpretations across the board, sometimes dramatic. In many cases, scale reliability is very different between the two perspectives, with designometric reliability being generally higher. While the differences in reliability were large across scales, on designometric level they were all in a useful range, ranging from just useful (Credibility) to excellent (Hedonism, Beauty). But, many differences exist on item level, going in both directions. Accordingly, the factor cardinality differs from theoretical expectations in all but one cases (Eeriness).

In the following we discuss the details and implications of our findings for scale developers and users, before we outline an agenda for more advanced (deep) designometric methods.

Implications for scale development

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 HQ-Stimulation did not meet this criterion, even under the designometric perspective.

The HQ-Identity scale (and to some extent also Credibility) shows a concerning pattern, where some items perform well psychometrically, but are designometrically extremely weak. This shows that developing a designometric scale under psychometric perspective can lead to *falsely favored* items that are well-behaved in ranking persons, but are inefficient for designs. To make the case, scale reliability when removing item HQI5 improves to 0.73 compared to 0.69. By further removing HQI1, reliability is 0.68. While this is not a major increase in reliability,

the same level is effectively reached with fewer items - the psychometric fallacy can lead to inefficient scales.

While we cannot show that directly, it is likely that fallacy also leads to *falsely rejected* 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 rejecting good items and selecting inefficient ones. A recent example is the development of the BUS-11 scale, where face validity demands (and factor analysis has confirmed) that *Privacy* is a separate construct, but only one item was left after item selection under psychometric perspective (Borsci et al., 2022).

Implications for users

For practitioners, the good news is 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 psychometric perspective being the lowest ranked designometrically.

Applications

A key idea in usability engineering is that interaction designers learn 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, receiving a lot of attention for intelligent agent design. Humanoid and animalistic robot design is coming out of its niche, and virtual reality applications are already 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. This space is huge and effective designometric instruments are needed to guide critical design decisions in hyper-excited times (Gartner, 2023). Our results show, that the psychometric fallacy is

harmful during scale development, leading to inefficient item sets and factor structures. It must not be purported.

Towards Deep Designometrics

By comparing the two perspectives, we illustrated that designometrics can be accomplished with standard psychometric tools by flattening the response box across participants. In principle averaging across users is legit, except in situations where users did not evaluate the same set of designs. But even when all users were evaluating the same set of designs, this averaging across users results in information loss. More specifically we lose information about the users, which would be interesting in its own right. For instance, it would be possible to evaluate a designometric model on the basis of responses of a single user. Formally, this would be a valid designometric measure, reflecting a single person's sensitivity to differences in a design attribute.

Designometric scales are commonly used to measure a populations reaction to a design, which implies that on some level the psychometric matrix is useful, for example to study the distribution of user sensitivity to a feature. Imaginable cases exist where one could use a designometric scale for psychometric purposes. For example, an instrument to measure trustworthiness of designs could be used to estimate faithfulness 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 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 (2021a) 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.

Such a multi-level approach is well in line with the established field of generalizability theory (Brennan, 2001). In that approach, the variance in responses is partitioned into multiple sources. Thus, suppose we have a data collection design where multiple users judge several designs, on multiple items, the variance in responses can be partitioned into variance due to individual differences in users, differences in designs, and differences in items. This can be done for the response box, but can straightforwardly be extended to hyperboxes. 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 (Schmettow & Havinga, 2013).

Such variance decomposition can be implemented in multi-level models, where the estimated variances can be used to compute reliability, whether it relates to measuring differences in users or differences in designs, while controlling for all other relevant sources of variation. By

extending the modelling to multi-level IRT models Berg et al. (2007) the discrete nature of the item response data can be taken into account.

Multi-dimensional exploratory methods have been well developed in chemometrics and sensory science Harshman (1970), but have seen little integration into mainstream psychometric or UX-scale validation workflows. Bridging this gap—by extending factor-analytic methods to multi-dimensional designometric data—constitutes a critical next step in establishing a rigorous quantitative foundation for multi-factor measurements.

Finally, the Eeriness scale is proof that *universal rating scales* are possible, which robustly perform in more than one perspective. Understanding how to develop such scales is an important next step in developing a deeper designometric methodology.

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.

Bargas-Avila, J. A., & Hornbæk, K. (2011). Old wine in new bottles or novel challenges. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2689–2698. https://doi.org/10.1145/1978942.1979336

Berg, S. M. van den, Glas, C. A. W., & Boomsma, D. I. (2007). Variance decomposition using an IRT measurement model. *Behavior Genetics*, *37*, 604–616. https://doi.org/10.1007/s10519-007-9156-1

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin*, *107*, 256–259.

Boom, Q., & Schmettow, M. (2013). *Hedonic quality: The inference and perspective processing approach* [B.S. thesis]. University of Twente.

Borsci, S., Malizia, A., Schmettow, M., Velde, F. van der, Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The chatbot usability scale: The design and pilot of a usability scale for interaction with AI-based conversational agents. *Personal and Ubiquitous Computing*, *26*, 95–119. https://doi.org/10.1007/S00779-021-01582-9/TABLES/9

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

Brennan, R. L. (2001). *Generalizability theory*. Springer New York. https://doi.org/10.1007/978-1-4757-3456-0

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.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika*, *66*, 271–288. https://doi.org/10.1007/BF02294839

Gartner, Inc. (2023). *Hype cycle for emerging technologies, 2023*. Gartner, Inc.

Haeske, A. B., & Schmettow, M. (2016). *The uncanny valley : Involvement of fast and slow evaluation systems* (pp. 1–49) [B.S. thesis]. University of Twente.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 1–84.

Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität. *Mensch & Computer 2003*, 187–196. http://link.springer.com/chapter/10.1007/978-3-322-80058-9_19

Hassenzahl, M., & Monk, A. (2010). The inference of perceived usability from beauty. *Human–Computer Interaction*, *25*, 235–260. https://doi.org/10.1080/07370024201050013 9

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*, 191–205. https://doi.org/10.1177/1094428104263675

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

Ho, C.-C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the godspeed indices. *Computers in Human Behavior*, *26*, 1508–1518. https://doi.org/10.1016/j.chb.2010.05.015

Keeris, D., & Schmettow, M. (2016). *Replicating the uncanny valley across conditions using morphed and robotic faces* (pp. 1–57) [M.S. thesis]. University of Twente.

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

Kuurstra, J., & Schmettow, M. (2013). *The influence of visual complexity and prototypicality on credibility judgments of websites* [B.S. thesis]. University of Twente.

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

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*, 314–324. https:

//doi.org/10.3758/s13428-011-0168-7
Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition*, *146*, 22–32. https://doi.org/10.1016/j.cognition.2015.09.008
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
McGrath, C., & Gaziano, K. (1986). Measuring the concept of credibility. *Journalism Quarterly*, *63*, 451–462.
Nazareth, D., & Schmettow, M. (2014). *The fluency effect as the underlying variable for judging beauty and usability* [M.S. thesis]. University of Twente.
Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10. https://doi.org/10.3389/neuro.11.010.2008
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
Polst, S., & Schmettow, M. (2014). *The user experience questionnaire and the impact of early information processing* [B.S. thesis]. University of Twente. https://purl.utwente.nl/essays/65768
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.
Schmettow, M. (2021a). Multilevel models. In *New statistics for design researchers* (pp. 267–322). Springer. https://doi.org/10.1007/978-3-030-46380-9_6
Schmettow, M. (2021b). *New statistics for design researchers*. Springer International Publishing. https://doi.org/10.1007/978-3-030-46380-9
Schmettow, M., & Havinga, J. (2013). Are users more diverse than designs? Testing and extending a 25 years old claim . In S. Love, K. Hone, & T. McEwan (Eds.), *Proceedings of BCS HCI 2013- the internet of things XXVII*. BCS Learning; Development Ltd.
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.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632
Slijkhuis, P. J. H., & Schmettow, M. (2017). *The uncanny valley phenomenon a replication with short exposure times* [M.S. thesis]. University of Twente. http://essay.utwente.nl/72507/
Theo, Bettina, S. M. L., & Held. (2008). Construction and evaluation of a user experience questionnaire. In A. Holzinger (Ed.), *HCI and usability for education and work* (pp. 63–76). Springer Berlin Heidelberg.
Tractinsky, N., Cokhavi, A., Kirschenbaum, M., & Sharfi, T. (2006). Evaluating the consistency of immediate aesthetic perceptions of web pages. *International Journal of Human-Computer Studies*, *64*, 1071–1083. https://doi.org/10.1016/j.ijhcs.2006.06.009

Tuch, A. N., Presslaber, E. E., Stöcklin, M., Opwis, K., & Bargas-Avila, J. a. (2012). The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *International Journal of Human-Computer Studies*, *70*, 794–811. https://doi.org/10.1016/j.ijhcs.2012.06.003

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, *31*, 279–311. https://doi.org/10.1007/BF02289464

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

William Revelle. (2025). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

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