Journal of Human Nature and Human Advancement
2025, Vol 1, Issue 1, 28-43
https://doi.org/10.5281/zenodo.18524687

Understanding Human Nature Through Psychology as a Quantitative Science

Paul De Boeck

The Ohio State University, Columbus

Abstract

Psychology is mainly a quantitative scientific discipline. We present and discuss three different quantitative approaches: psychometrics, experimental psychology, and mathematical psychology, from two perspectives. The first perspective is historical, to better understand current practices. The second perspective is a comparative and critical discussion pointing to possible extensions most of which have in fact already been implemented. The main story line of our contribution is that after a good start in the early stages of psychology, unfortunately experimental and correlation psychology grew apart but Cronbach's (1957) request for a federation of the two is now eventually being realized. Hopefully, the integration will further develop and improve our understanding of human nature. Where the integration and the future of a quantitative psychology will lead us more precisely is hard to predict.

Keywords: psychometrics, experimental psychology, mathematical psychology, human nature

Psychology as a science has originated from different sources and still is a rather heterogeneous scientific discipline. The dominant view today is that psychology as an empirical discipline is based on quantitative methods for the investigation, explanation and prediction of human behavior. James McKeen Cattell, the first professor of psychology in the United States after he had studied with Wilhelm Wundt in Germany and worked with Francis Galton in the United Kingdom, advocated that psychology had to be a quantitative science (Cattell, 1890). Cattell's work was influenced by the approaches of both, Wundt and Galton, for example, in his development of mental tests to measure mental power, a unitary concept for what was later called intelligence. Cattell's mental tests were popular experimental tasks at the time, and following Galton, he planned on using correlations for his investigations, based on individual differences in experimental tasks. His ideas testify to a combination of the two scientific disciplines of psychology: experimental psychology and correlational psychology as later discussed by Lee Cronbach in his presidential address for APA (Cronbach, 1957). In the same address, Cronbach argued that a "federation of the disciplines is required" (p. 671). In Cronbach's words, experimental psychology studies only variance among treatments and correlational psychology studies only variance among organisms (p. 671). In fact, James McKeen Cattell had already implemented the "federation." Although there have been interactions between the two disciplines, such as experimental approaches to construct validity (Cronbach & Meehl, 1955) and aptitude-treatment interaction approaches (Cronbach & Snow, 1977), the two disciplines largely grew apart. One discipline focused on individual differences and the other on experimental (i.e., treatment) effects. As far as there is a happy end to the story told in this contribution, there are now clear signs of a (re)integration. Clearly, in Cronbach's view, an integration would help to understand human behavior.

Corresponding Author: Dr. Paul De Boeck ✉ deboeck.2@osu.edu
Published by Advanced Psychological Resources, Inc.
https://johnaha.org/ | ISSN: 3065-4882

The correlational approach became strongly connected with *psychometrics* as the discipline of measuring psychological variables used in correlational methods. The experimental approach has evolved in two directions. The majority direction is one of *testing experimental effects*, using inferential statistics with a known distribution such as the t-statistic and the F-statistic to appraise experimental effects (e.g., t-tests and analysis of variance), often using null hypothesis significance testing (NHST). The minority and more quantitative direction of the experimental approach is called *mathematical psychology* in which equations are used to relate human behavior (mostly responses) to experimental conditions and stimulus features that can be linked to mostly cognitive process activities in preparation for response behavior. Early examples are from psychophysics in which physical features of a stimulus are related to measures of sensation and perception (Fechner, 1860) and from work by Ebbinghaus (1913) on forgetting curves in which time since encounter with stimuli is related to the proportion of forgetting. In the mathematical psychology tradition, NHST is less important; the aim is more comprehensive, to develop and test equations to represent processes.

Although the three categories of approaches (1. psychometrics, 2a. testing experimental effects, and 2b. mathematical psychology) provide the structure of our contribution, with one section per category, the categories are neither clearly delineated nor exhaustive. An example of overlap is that experimental effects can and are being investigated using psychometric and mathematical psychology models. An example of non-exhaustiveness is that psychological scaling of object perception as a topic is not included in our discussion except very briefly in the section on mathematical psychology, although it also has links with psychometrics. Where relevant, we will point to these unavoidable classification complexities. They are inherent to any kind of classification of research activities.

Our discussion of the three quantitative approaches is inspired by two perspectives. See Table 1 for a schematic presentation of the historical narrative. The second perspective is a comparative and critical discussion pointing to possible extensions some of which have in fact already been implemented.

Table 1. Schematic presentation of the development of quantitative psychology (psychometrics, experimental psychology and its mathematical branch): from integration to relative separation between the two disciplines of psychology, into a developing reintegration

<table>
<tr><td colspan="4">James McKeen Cattell Psychometrics & Experimental Psychology integration</td></tr>
<tr><td>Development of Psychometrics
(Spearman)
classical test theory (CTT)
reliability and validity, factor analysis, latent variable (LV) models

Focus on individual differences and correlations</td><td>Relative separation

“the two disciplines”
Lee
Cronbach</td><td colspan="2">Development of Experimental psychology
(Wundt)
experimental designs, ANOVA (Fisher)
null hypothesis testing of effects

Focus on differences between conditions, defined as effects</td></tr>
<tr><td rowspan="2">LV models in question
e.g., network psychometrics
(Borsboom)

Adding intra-individual differences</td><td rowspan="2"></td><td>Continuing experimental</td><td>Mathematical process models (Estes)</td></tr>
<tr><td>developing reintegration</td><td>Adding individual differences in effects</td><td>Adding individual differences in parameters</td></tr>
<tr><td colspan="4">Fulfillment of Cronbach’s (1957) request for a “true federation” of the two disciplines (p. 671)</td></tr>
</table>

From a subjective point of view, we have a winner among the three approaches in mind, but we also believe there is some value in Feyerabend's (1975) words that "anything goes" on the condition that it works and helps us make progress. From all three approaches, psychological scientists keep making contributions to their discipline. We close this paper with a final comparison of the three approaches and with considerations on the discipline of psychology and the potential of quantitative approaches to make progress in that discipline.

The Psychometric Approach

Early History

Based on the unitary mental power concept, Cattell was expecting high correlations between the mental tests he was interested in. When Clark Wissler, a student of Cattell, used Cattell's mental tests for his doctoral dissertation, he was disappointed by the poor correlations. However, when Charles Spearman saw Wissler's empirical results, he was not convinced. In his view, the poor correlations were due to reliability issues of the tests. His interpretation led Spearman to develop a correction for attenuation of correlations in the presence of imperfect reliability (Spearman, 1904), and after the correction in question, Wissler's correlations were found to be much higher.

Spearman's work was the origin of what is now called classical test theory (CTT) (e.g., Traub, 1997), the earliest approach to psychometrics. CTT specifies that an observed test score stems from two sources: a true score and measurement error. A natural extension of CTT is that multiple test scores all rely on the same true score plus a test-specific term with a variance that depends on the test. This extension is Spearman's two-factor theory stipulating that test scores stem from two factors: (a) a general factor and (b) a specific factor, referring to sources specific to the test and its application, a combination of measurement error and a systematic factor specific to the test.

Further Developments

The next step in the development of factor analysis is that more than one true score (e.g., called factor scores) plays a role and that tests can differ in how important the true scores are for each of the factor-analyzed tests. To make tests comparable, their variance is standardized and correlations are used for a factor analysis, so that factor loadings can be interpreted as standardized regression weights. Instead of test scores, item scores were also factor-analyzed. To interpret the factors, orthogonal or oblique rotation to simple structure is used (e.g., varimax) so that factors could be interpreted as quantitative reflections and measures of constructs. A simple structure is a structure with items or tests loading mostly on just one factor, so that subsets of items or tests can be combined to measure the same construct. The link of factors to psychological constructs explains the huge popularity of factor analysis. Most constructs are multidimensional and they all need to be measured. Factor analysis became the vehicle to investigate the dimensionality of constructs and to measure the dimensions.

From a measurement point of view, factor analysis was an important development. The standard view of measurement used to be (and often still is) *operational measurement*, defined by Stevens (1945) as the "assignment of numbers to objects or events according to rules". Although the operational strategy was still being used to design the items and the tests that were factor-analyzed, the factors are a step beyond operational measurement. Factors became an analysis-based intermediary between constructs and their measures.

Finally, factor analysis became a model based method, with factors as latent variables (LVs), as unobserved causes of the manifest (observed) variables such as item responses or test scores. The simple structure ideal was retained and built into the model, so that rotations would no longer be needed to interpret the factors. The simple structure model is one and perhaps the most common form of confirmatory factor analysis (CFA). At this stage in the development, using factor models as measurement models implies that measurement becomes *representational measurement* instead of

operational measurement. The representational aspect of measurement means that measurement represents the empirical relations between variables (Luce et al., 1990).

Given that factor analysis has become so popular, it is remarkable that sum scores are still the favorite way to measure, whereas factor scores would be the more logical scores to rely on for measurement (McNeish & Wolf, 2020). In practice, the estimation of a factor model only functions to verify whether the model applies or would need modifications, with consequences for scoring. Possible reasons to stay with sum scores are that they are simple to calculate, robust (Liu & Pek, 2024; Widaman & Revelle, 2023), and do not suffer from the problem of factor score indeterminacy (e.g., Steiger, 1979), meaning that different sets of factor scores would be equally consistent with the same factor model.

Intra-individual Differences and Ergodicity

An important feature of factor analysis applications is that the correlations and covariances as input for the analysis are relations across individuals, so that the corresponding measurements are *measurements of individual differences*. The differences do not need to be stable and can refer to changing differences.

Unfortunately, the structure of inter-individual differences does not necessarily generalize to intra-individual differences. For example, negative emotions tend to show high correlations across individuals, whereas an individual person would not easily feel depressed and angry at the same time. Constructs that are approached through factor analysis may not apply to intra-individual variation. The necessary condition to generalize the structure from interindividual to intra-individual, is called the ergodicity condition (Molenaar, 2008) which is related to Simpson's paradox and the ecological fallacy. It is important for the development of psychological theories to investigate the two levels of variation (interindividual and intraindividual) (Borsboom & Haslbeck, 2024; Kuppens et al., 2022).

Note on Reliability

Like factor analysis, the whole field of psychometrics is still largely based on individual differences. A prominent example is the reliability coefficient. The reliability coefficient of a score can be defined as the ratio of true variance versus observed variance of the score across individuals. The consequence is that the reliability is sample-dependent, larger in heterogeneous samples (with a larger true variance) and smaller in homogeneous samples (with a smaller true variance); see e.g. Wang and De Boeck (2022) for implications. Therefore, the reliability coefficient should not be considered a property of a test score but of a score on a test presented to a sample of respondents, without knowing what the (sub)population is. The same is true for validity coefficients.

It is remarkable that reliability coefficients are not used for measurement instruments in important disciplines other than psychology and related disciplines. In using weight balances for example, it does not make sense that measurement quality of balances would depend on the weight differences between people whose weight is measured. The more important notion is precision (standard error of measurement). The same degree of precision may co-occur with very different values of the reliability coefficient. It is a well-known but often ignored qualification of reliability coefficients.

Alternatives

The core aspects of factor analysis are still roughly the same as developed in the seventies and eighties of the previous century. There have been developments such as factor analysis of ordinal data, mixture factor models, multilevel approaches, dynamic factor models, growth curve models, the bifactor model, and various types of structural equation models (SEMs). Bayesian approaches and even AI approaches have also been introduced for estimation. However, the factors are still LVs and the factor loadings are regression weights to predict indicator variables (e.g., item responses).

An LV alternative to factor analysis is item response theory (IRT), an approach that has not made it to the mainstream in psychology though it is the most important LV approach for educational

measurement. Ignoring similarities and differences between models and estimation methods for the purpose of this contribution, there are two main differences between IRT and factor analysis. First, IRT is an approach for categorical data (binary, nominal and ordered-category data), whereas factor analysis is an approach for interval-scale data but has been extended for ordered-category and binary data. Fortunately, under certain conditions, treating rating scale data with five or more points on the scale as interval-scale has no dramatic consequences. Second, IRT has a richer toolbox than factor analysis to investigate, appraise, and improve test-based measurement (e.g., score dependent standard error, test information, differential item functioning). A possible reason for staying with factor analysis is the common practice to use sum scores so that CTT and Cronbach alpha can be relied on. It illustrates how sociological effects play a role in research practices.

A promising new development is a trend away from LVs to more specific notions. The trend is related and to some extent inspired by empirical findings regarding the specific variance of items. Mõttus et al. (2017) found that specific traits below the facets of the big five do have predictive value, and Revelle (2024) discussed findings that with an increasing number of items there is a trade-off between the reliability and validity of the test score as a function of inter-item correlations. Reliability increases whereas validity (primarily predictive validity) does decrease if the inter-item correlations are higher. What this means is that specific aspects of the items (e.g., based on trait wording) detract from reliability but do contribute to validity. These findings fit with a view that constructs are heterogeneous and complex, and that not for all purposes constructs or subconstructs should be narrowed down to just one vector in the space of psychological variables (De Boeck et al., 2023).

In the new development away from LVs, an alternative for LVs is formulated, called network psychometrics (Borsboom et al., 2021). Specific observable variables (e.g., items, symptoms) are considered to form a network with directed or undirected edges to explain covariation between the manifest variables that were previously considered to constitute LVs through what they have in common. Psychometric networks are based on partial correlations between pairs of variables, controlling for all other variables in the same set of variables to measure a construct (items, symptoms, etc.) such as depression items or symptoms. Certain types of networks are formally equivalent with LVs but not all networks are. Under certain conditions the networks can also be interpreted as causal explanations of the covariance structure of the manifest variables. Network psychometrics recognizes better than LV models the more specific aspects of test items and symptoms and the validity of those aspects, and it also avoids the tautology that LVs have causal effects on the items and symptoms. The tautology is that LVs are derived from manifest variables and are assumed to have causal effects on the same manifest variables. Network psychometrics is also a way to explore processes in the development of syndromes such as depression (Borsboom, 2017) and of attitudes (Dalege et al., 2016), assuming, however, that the processes are the same across individuals. Time series approaches based on experience sampling techniques can be used to investigate intra-individual covariation and eventually to model interindividual differences in the intra-individual covariations (Borsboom & Haslbeck, 2024).

Summary and Discussion

The primary strength of psychometrics is to support the measurement of psychological variables and to evaluate the quality of the measurement. However, psychometrics is almost exclusively focused on individual differences; see the section on mathematical psychology for the measurement of object and stimulus perception instead, for which the term psychological scaling can be used.

Because psychometrics makes use of human behavior data (e.g., item responses), it is necessarily a confounded enterprise. Item responses are subjected to psychological processes, whereas psychometric models can hardly be considered psychological process models (De Boeck & Gore, 2023). Nevertheless, psychometric models still seem to work in that they are useful for psychological measurement.

Experimental Approach

Experimental psychology is focused on the effects of interventions as independent variables (IVs) defined by factors and conditions in an experimental design; the effects concern behaviors and aspects of behaviors as dependent variables (DVs). The experimental method is the method par excellence to make causal inferences. The quantitative aspects of experimental psychology come from (a) quantitative DVs whereas the IVs are commonly nominal variables (although they can be quantitative), (b) the size of effects of IVs on the DVs, and (c) the quantified uncertainty of inferences about the effect of the IVs. The three quantitative aspects are further elaborated in this section.

Early History

Edwin Boring published a book on the early history of experimental psychology (Boring, 1929). He focuses on experimental psychology approaches as far as " what they meant to Wundt and what they meant to nearly all psychologists for fifty or sixty years—that is to say, the psychology of the generalized, human, normal, adult mind as revealed in the psychological laboratory" (p. viii). The quote reveals among other things two features of early experimental psychology. First, experimental psychology is empirical psychology in a broad sense of observations in controlled situations and contrasted with philosophical psychology. Second, rather simple perceptual and response time tasks and learning and memory tasks are being used as contrasted with the study of more complex concepts such as personality and culture, for which Wundt used the term “Völkerpsychologie” which also includes the study of individual differences (i.e., they are not part of “the generalized mind”). An interest in the generalized mind also explains why Fechner and Ebbinghaus are often mentioned as early experimental psychologists.

The experimental method is defined by Ronald Fisher (1935) as investigating and testing the effects of different treatments that are randomly assigned to the experimental units (i.e., subjects in psychological experiments). Apart from randomization, two other principles are (a) blocking into rather homogenous sets of experimental units, to counter heterogeneity that would reduce the precision of the effects, and (b) replication. Fisher has also developed and proposed the ANOVA method to test experimental effects using H0s and p-values. These methods are still popular in present day experimental psychology.

Dependent Variable (DV), Analysis Methods, Psychometric Qualities, and Independent Variables (IVs)

The DV

For most studies, the DV is quantitative, binary and nominal variations also being possible. Also for most studies, an operational type of measurement definition is used for the DV, not a representational measurement (i.e., without LV modelling). If just one observation is made per subject (e.g., one item response such as a likability rating or a sum score across items), the DV is a univariate DV and the design is necessarily a between-group design. If more than one observation is made (e.g., responses to a set of stimuli), the observations can either be considered as repeated observations (also called repeated measures) or they are considered as a multivariate DV.

Analysis methods

For univariate DVs and repeated measures, ANOVA or t-tests are commonly used. The underlying model for ANOVA and t-tests (and for multiple regression) is the general linear model (GLM). Other models of interest for an analysis of the data are:

- the *generalized linear model* (GLiM) for the case a link function is needed (e.g., for logistic models) and/or other distributions than the normal distribution apply for residuals;

- the *linear mixed model* (LMM) as an extension of the GLM for within-group factors with random effects;
- *the generalized linear mixed model* (GLMM) combines features of GLiM and LMM.
- MANOVA for multivariate DVs.

The LMM and GLMM require repeated measures, although the term "repeated measures" is not commonly used for LMM and GLMM. The LMM and GLMM can also be used for multilevel models.

Like the GLM model that underlies ANOVA, t-tests, and multiple regression, GLiM, LMM, and GLMM are data analysis models in the first place in that they are mostly used to estimate and test the effects of interest, rather than that they are meant to represent hypothesized processes at the basis of the data. Goodness of fit (GOF) of the models is less of an issue than for LV models, although GOF can of course be used to evaluate model extensions with additional effects.

Psychometric qualities of DVs

The psychometric qualities of DVs are often not assessed and reported. Flake et al. (2015) investigated practices in publications because they believe that issues of reliability and validity are possible explanations for the replication crisis. The authors concluded that the reporting practices should be improved. We agree but also believe that psychometric qualities are not necessarily qualities in an experiment. See our earlier discussion of the reliability coefficient. The reliability coefficient is a coefficient for the consistency of individual differences, whereas experiments are conducted to investigate differences between conditions. In fact, the reliability of a DV increases with the true variance of the DV in a study whereas in a between-group design the precision of estimated effects (e.g., the differences in the means of conditions) decreases with the observed variance (of which the true variance is one part). In a within-group design, the reliability of the observed effect increases with the individual differences of the effect but the precision of the general (i.e., mean) effect decreases. Another concern is that an experimental condition can affect the true variance of the DV within conditions (and not just the mean) with consequences for the reliability coefficients within conditions. Therefore, these reliability coefficient should perhaps not be interpreted in any other way than as stemming from a difference in true variance. Condition-specific reliability coefficients may stem from treatment effects on the variance of the DV. In sum, it is not clear how relevant reliability coefficients are for experiments.

IVs

Typically in an experimental study, the IVs (the factors and conditions) are nominal variables, and quantitative IVs would be categorized to conduct an ANOVA. The historical reason for the nominal nature of the IV is that Fisher worked as a statistician for an experimental agriculture research center and the treatments of fields were primarily qualitative and not quantitative .

Until today, ANOVA is the simplest way of analyzing data with nominal IVs. For example, regression analysis with nominal IVs requires coding (e.g., dummy coding, effect coding) and the results of the analysis are not easy to interpret, especially if interaction effects are included (Cohen et al., 2013, 3rd ed.).

Size of Effects

Because in most experimental studies the DV is quantitative, the size of an effect is an important result of the data analysis. Still, the effect size is not always the primary point of interest of researchers. They rather want to know whether a treatment (an experimental condition) has an effect at all. If one is primarily interested in whether something has a causal effect, it may seem less important how large the effect is as it may be influenced by enhancing the manipulation.

The presence of an effect is inferred from a rejection of H0 in line with Popperian thinking (Popper, 1959) and the hypothetico-deductive approach of testing falsifiable hypotheses. One reason for a

possible neglect of reporting the effect size is that the research hypothesis of interest is not quantitative. The research hypothesis is the alternative hypothesis in NHST, denoted as H1, the negation of H0. Statistical significance can be considered as a falsification of H0 and therefore as evidence in favor of H1. The size of the effect is only secondary information.

Even though the effect size would be only secondary information, reporting it is nevertheless recommended by organizations such as the American Psychological Association. The effect size is either expressed on the scale of the DV, or on the Cohen d scale (the effect on the DV scale divided by an estimate of the within-condition standard deviation), or the effect size is reported as a percentage explained variance, η^2, of the DV by factors of the design. The latter two, Cohen's d and η^2, are standardized ways to express effect size. They are introduced to deal with the problems that (a) the DV scale is arbitrary in most psychological studies and (b) that studies with different scales may not be comparable. Unfortunately, the solutions create new problems. Cohen's d depends on the within-condition variance and therefore, it also expresses sample heterogeneity independent of the effect. Standardization may lead to a false meta-analysis conclusion of effect heterogeneity not based on the raw effect size but on the heterogeneity of the sample instead. Standardized effect size can also be confusing as a basis to differentiate between small (0.2), medium (0.5), and large (0.8) effect sizes (Cohen, 1992). Also η^2 depends on the heterogeneity of the sample. An interesting feature of many experimental designs is that the factors are orthogonal, so that the effect size of a treatment does not depend on other possible factors in the design. We believe it is a good recommendation to also consistently report the unstandardized effect size (Pek & Flora, 2018).

Quantification of Uncertainty

The two common indications of inferential uncertainty about effects are the p-value and the confidence interval (CI). More recently, Bayesian approaches have gained popularity as an alternative to quantify inferential uncertainty, as will be explained after discussing the p-value and the CI.

The p-value links an observed effect through a statistic (e.g., a t-value) to the probability of observing an equally large or more extreme observed effect assuming H0 is true. (For simplicity's sake, we stay here with two-tailed testing). A common misinterpretation of the p-value is that p is the probability of the null hypothesis being true, as pointed out by Cohen (1994), so that the probability of H1 would be $1 - p$, which is an easy but incorrect interpretation. Unfortunately, there is no p-value for H1, so that it is not possible to compare H0 with H1 in terms of p-values.

Due to a proposal by Neyman and Pearson (1933), the p-value is used for a decision to either reject H0 or not to not reject H0, based on a preset cutoff level for the p-value called the α-level (e.g., .05). Although common practices exist regarding the value of α (e.g., .05, .01), an explicit rationale for these values is often absent. A possible rationale could be provided by a valuation of decision outcomes: a correct rejection of H0 (given that H0 is false), an incorrect rejection of H0 (given that H0 is true), a correct non-rejection of H0 (given that H0 is true), and an incorrect non-rejection of H0 (given that H0 is false). Suppose that H0 is true, then the probability of an incorrect rejection of H0 is .05 if $\alpha = .05$ and suppose instead that H0 is not true, then the probability of an incorrect non-rejection of H0 would be .20 following the recommended power rate of .80 with β=.20 (Cohen, 1992). What this means is that an incorrect rejection of H0 is four times as negative as an incorrect non-rejection. Cohen was well aware of the implicit valuation of the two incorrect decisions, whereas in practice, one seems to accept the valuation ratio of the two incorrect decisions (-4 versus -1) without further consideration, blindly following Cohen, independent of the topic and what is at stake. In a more complete rationale one can also take the two correct decision outcomes into account if the prior probability of H0 being true (versus not true) would be known, so that the expected value of the two decisions (reject or not reject H0) can be determined from which an optimal p-value (and power level) can be determined.

For the CI, the question one would want to see answered is what the interval is within which the true effect size is most likely located, for example, with a probability of .95. However, the CI does not answer that question. The CI would apply if it is delineated around the true effect size to locate the observed effect size whereas in fact it is delineated around the observed effect size to locate the true effect size. Still, Cumming (2014), based on his criticism of the p-value and his new statistics proposal puts an emphasis on the observed effect size and CIs. A problem with the proposal is that the CI easily leads to misinterpretations. For example, a 95% CI of an observed effect size does not mean that the probability of the true effect size being comprised in the CI is .95. In fact, a 95% CI only tells us that if an experiment is repeated, the expected percentage of times the true effect size being comprised in the corresponding CI is 95%. It may not help to know the CI for a specific study because the CI would be different each time depending on the specific study.

Bayesian statistics can help us with answers one expects from the p-value and the CI. As an alternative for the p-value, the Bayes factor (BF) tells us how likely the H0 is in comparison with a H1. The BF answers the question that many believe the p-value answers: how likely it is that the true effect is zero compared with the true effect being different from zero. Note that in fact the common way to express the BF puts the likelihood of H1 in the numerator and the likelihood of H0 in the denominator. For an application of the BF as an alternative for the t-test, see Rouder et al. (2009).
In a similar way, the Bayesian credible interval answers the question one expects to be answered by the CI: What is the interval within which the true effect size is most likely located, for example, with a probability of .95. As explained, the CI does not answer that question.

Unfortunately, the method to obtain the Bayesian answers to the inferential uncertainty questions (BF instead of p-value, credible interval instead of CI) is rather difficult. Fortunately, the software package JASP (JASP Team, 2025) can help to make Bayesian answers easier. To compare the two types of answers, JASP runs the Bayesian analysis in parallel with the more traditional frequentist analyses such as t-tests and ANOVA. For an indication of how p-values compare with BF-values, see Jeon and De Boeck (2017).

Summary and Discussion

The primary strength of the experimental approach is its potential to make causal inferences on effects of IVs on DVs. Whereas psychometrics is focused on the measurement of psychological variables with individual differences and on the quality of the measurements, the experimental approach is focused on differences between experimental conditions in terms of quantitative DVs. The size of the effects is not always of primary interest; instead, the question is whether the effects “exist” meaning whether H0 can be rejected.

Remarkably, the uncertainty is higher the larger the individual differences are, which implies that psychometric qualities of the DV measurement may not increase the precision of the effect size. Psychometrics and the experimental approach are two different worlds as they focus on differences of a different type (between individuals versus between conditions). An integration of both is desirable to understand human behavior, but it is best not to extrapolate from one to the other.

A better way psychometrics can assist experimental psychology is through SEM models in which the DV is a latent variable measured through a set of items (the measurement model of the SEM), with the IV as an exogenous variable that has an effect on the latent DV (the structural part of the SEM). A practical problem is that measurement models with LVs require very large sample sizes.

Commonly, experimental effects are interpreted as general effects in the sense of effects that apply to people in general. The H0 does not say anything about possible differences between individuals with respect to the effect. It is quite possible that the effect is statistically significant and that (at the same time) a substantial subset of the participants shows the opposite effect. It is unfortunate that individual differences in the effect often go unnoticed.

In principle, one could use individual differences in effects as an approach to measure important psychological variables. However, studies by Haaf and Rouder (2019) and Hedge et al. (2018) have not led to encouraging results. The experimental effects in the studies these authors refer to are robust but rather poorly correlated across trials and tasks. The results may be different for other experimental effects. For a tutorial on measuring individual differences in experimental effects, see Brysbaert (2024). For a psychometric approach to individual differences in the effect of within-subject factors, explanatory measurement approaches (De Boeck & Wilson, 2004) and Embretson's (1976, 1983) work on the potential of test design can be useful. The underlying idea is that the item responses on a test are repeated measures. Therefore, item features of test items can be conceptualized as within-group factors (covariates of repeated measures) and these (intra-individual) effects may show interindividual differences. In ANOVA terms, these interindividual differences in intra-individual effects show as interactions between subjects and within-group factors.

Mathematical Psychology

Mathematical psychology can be defined as using equations for how behavior is generated depending on conditions and stimuli that are encountered. Its approach used to share with experimental psychology a focus on general principles of response behavior. Compared with experimental psychology, there also are two differences. First, mathematical psychology is less focused than experimental psychology on effects of specific interventions although the mathematical models do include the effects of conditions and stimuli presented to respondents. Second, mathematical psychology makes use of models (with sets of equations) that are intended to reflect the processes of behavior generation, whereas experimental psychology mostly relies on atheoretical data analysis models to relate IVs with one or more DVs.

Early History

Historically, mathematical psychology started to take off in the 1950's with Estes' statistical learning theory (Estes, 1950) and Bush and Mosteller's (1951) linear operator theory. There have been predecessors of mathematical approaches of psychological phenomena, such as work by Fechner and Ebbinghaus, and by behaviorists who have developed equations to predict learning based on reinforcement through drive reduction (Hull, 1943; Spence, 1936). Just to give one example from Hull's work, ambivalence relies on the avoidance gradient being steeper than the approach gradient, so that a rat would approach an ambivalent object from a distance up to a certain point where avoidance takes over from approach and the rat runs away from the ambivalent object, then restarts to approach, and so on.

Like Spence and Hull, Estes and Bush and Mosteller were inspired by the behaviorism of their time. Estes himself was a student of Skinner. Estes' theory entails equations for how the latency of a response decreases with reinforcement from learning trials and how response rates increase. The linear operator approach by Bush and Mosteller is similar except that the form of the equations is somewhat different. It is rather typical for mathematical psychology that alternative (and competing) models are formulated. Apparently, different models can explain the data roughly to the same extent.

The primary motivation of developers of mathematical theories is clearly formulated by Estes "It seems likely that progress toward a common frame of reference will be slow so long as most theories are built around verbally defined hypothetical constructs which are not susceptible to unequivocal verification" (Estes, 1950, p. 94). The psychological phenomena of interest at the time were related to learning, in line with a behaviorist inspiration. Later, models for other phenomena were also developed, mostly for cognition related behavior, such as for choice and decision, which is a natural extension given that learning became interpreted as a cognitive phenomenon. All these models are probabilistic and meant to represent processes leading the behaviors of interest. The interest in processes is not

surprising given that behaviorism is a process theory. The ambition to develop equations to represent processes sets mathematical psychology apart from other quantitative approaches in psychology.

An Example

An example of a popular and successful model is the drift diffusion model for choice and decision (e.g., Ratcliff & McKoon, 2008). It is a model for responses and response times when a subject is presented with a question that requires a choice between two options. For example, the question is whether the number of scattered dots shown on a screen is larger than 100 or not. To simplify, the model has two main parameters: drift rate and boundary separation. The drift rate is the efficiency of evidence accumulation from random information sampling. The accumulation process moves up and down between two opposite boundaries, one for each of both options (e.g., larger than or not) until one of both boundaries is hit and a decision is made in favor of the corresponding option. Boundary separation is a parameter that takes care of the speed-accuracy trade-off. A larger separation requires more evidence for boundaries to be reached, with consequences for the accuracy and response time of the choice. Time pressure is a way to reduce the boundary separation. The clarity versus complexity of information is a way to manipulate the drift rate. The drift diffusion model is meant to explain which choice is made (i.e., one of the two options) and the time it takes to decide (i.e., the response time). As an alternative type of models, the race models (e.g., Heathcote & Matzke, 2022) are also based on evidence accumulation but with the choice options each running an evidence accumulation race, the winner of which determines the response.

Cognitive Psychometrics

For a model to be really comprehensive, it needs to account for individual differences in the parameters. The extension to individual differences is a rather recent phenomenon. It does not only reflect a necessity but also offers an opportunity to assess individual differences and to use LV modeling. The new trend is called *cognitive psychometrics* (e.g., Batchelder, 2010; Lee, 2011; Rouder et al., 2015) with measurement as a spin-off of a process model. For example, Vandekerckhove et al. (2011) have proposed and applied an approach to extend the drift diffusion model with random person effects functioning as LVs for the drift rate and boundary separation.

An advantage of cognitive psychometrics is that the LVs can directly be interpreted in terms of the processes implied in the model, which is a strong and inherent validity argument compared with purely psychometric approaches, with validity research following measurement (first measure, then validate the measurement). Embretson's (1983) view on validity based on her cognitive design system for psychometric models implies a similar view.

Present-day mathematical psychology and cognitive psychometric models make frequent use of Bayesian estimation methods because of their flexibility and broad potential to estimate models. Process-based models tend to be complex and they are also hierarchical with between- and within-person random effects. Such models require methods that are more flexible and have a broader range of potential than the traditional maximum likelihood estimation methods.

A practical limitation of mathematical psychology, including cognitive psychometrics, is that it works with rather simple tasks somewhat in line with early experimental psychology such as used in psychophysics by Fecher, in Wundt's laboratory, and in Cattell's mental tests. A possible reason is that to build a comprehensive model, more complex tasks would be a serious complication.

Related Fields

Some quantitative domains within psychology are traditionally associated with mathematical psychology although they may not fit with how we have presented mathematical psychology. Already early in the history of psychology, scaling methods had been developed, for example based on paired comparisons (Thurstone, 1927). A prominent example of psychological scaling is the scaling of object

perception (perception of stimuli, items, statements, values, etc.) (e.g., Coombs, 1964) which became known as multidimensional scaling (MDS) (Young & Hamer, 1987). Psychological scaling could also be categorized under psychometrics if psychometrics were not mainly focused on measuring people but would extend to measuring perceived objects. MDS can also be applied to perceptions of people and to any kind of pairwise (dis)similarities and associations of objects, and has also been used for scaling the interactions between people in social networks. Interesting to note is that MDS-like network models are now also used in the context of IRT (e.g., Jeon et al., 2021) and for the association of social networks with LVs based on attitude questionnaires (Sweet & Wang, 2025), which is another illustration of how the three quantitative approaches overlap.

A somewhat related topic is signal detection theory (SDT) (Green & Swets, 1966) with models for the differentiation between signal and noise and between stimuli more generally. SDT is also used to assess binary predictions using receiver operator curves (ROCs) and area under the curve (AUC) indices that are in vogue to evaluate models for binary and categorical data.

More generally speaking, mathematical psychology now largely overlaps with computational models from cognitive science and cognitive neuroscience. The mathematical modelling approach is also further developed in those neighboring disciplines of psychology. It is a general phenomenon that the discipline of psychology fans out to related disciplines.

Summary and Discussion

The primary strength of mathematical psychology is the use of process models to explain behavior in experimental tasks. In the more traditional experimental psychology, the models function to analyze the data without the ambition to represent the processes that generate the data. Instead, experimental psychology focuses on effects of interest, to answer the question whether interventions have an effect, based on NHST. Effects of conditions are part of a broader mathematical model and embedded in that broader model.

One might consider mathematical psychology also a form of psychometrics if the models account for individual differences in the parameters of interest as in cognitive psychometrics. The difference with psychometrics is that mathematical psychology models carry the ambition to represent the processes that generate responses, whereas for measurement models the first aim is measurement.

General Discussion

All three approaches used to focus on only one type of differences-- either individual differences as in psychometrics, or differences between conditions and stimuli as in experimental and mathematical psychology commonly interpreted as general intra-individual differences. However, for all three approaches, an interest has been and still is growing in the other type of differences: intra-individual differences in psychometrics, and individual differences in experimental and mathematical psychology. In other words, Lee Cronbach's wish that the two disciplines of psychology become integrated is being fulfilled. The integration and the growing overlap is, in theory, a welcome trend to better understand human behavior. Where the trend will lead us is an open question. Progress in science is often not predictable.

Even after the integration in line with Cronbach's wish, there are remaining differences, two of which we will discuss here. First, clearly, the experimental method is the most direct strategy to establish causality. Psychometrics is foreign to causality, and in mathematical psychology, the specific model that is used functions as possibly colored glasses through which the effect is seen. Causality is a complex issue. A discussion of causal approaches would lead us too far from our current discussion. Second, perhaps the experimental method is the more realistic approach for a reality that is highly complex with many extraneous factors that make it difficult to build models, such as LV measurement models and especially mathematical psychology models.

Apart from its focus on rather simple cognitive tasks, mathematical psychology models, in our subjective evaluation, seem to be the most desirable of quantitative approaches. They are not just instrumental, they are process models with substantive relevance. Therefore, they are well-suited as approaches to understand human behavior. In addition, they allow for investigating and testing experimental effects, and they have the potential to measure easily interpretable relevant individual differences. Whereas the models are commonly models for simple cognitive tasks, they can in fact (and should) also be applied more to a broader type of data, for example to affective dynamics data (e.g., Ryan, et al., 2025; Vanhasbroeck et al., 2024). A possible weakness is that different models may work about equally well.

A major question is whether quantitative methods, their integration and possible improvements, some of which are discussed in this contribution, are sufficiently equipped to deal with a complex behavioral reality (e.g., De Boeck, et al., 2023). That complexity is a possible reason to stay with the simpler approaches of experimental psychology possibly at the cost of some replication issues. Another uncertainty is whether not artificial intelligence methods have a better potential to deal with a complex psychological reality.

Finally, although psychology is mainly a quantitatively leaning discipline, also helpful are qualitative methods, for example to build theory and because of their flexibility to understand complex psychological phenomena. Our bet is that the future of psychological research will still be mainly quantitative but in unplanned ways (as it is the case for most innovations) and also inspired by substantive research questions. However, qualitative input and qualitative touches and a mixed quantitative and qualitative approach will remain important to solve clinical and other applied problems.

References

Batchelder, W. H. (2010). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 71–93). American Psychological Association. https://doi.org/10.1037/12074-004

Boring, E.G. (1929). *A history of experimental psychology.* Appleton-Century.

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5-13.

Borsboom, D., & Haslbeck, J. (2024). Integrating intra- and interindividual phenomena in psychological theories. *Multivariate Behavioral Research*, *59*(6), 1290-1309. https://doi.org/10.1080/00273171.2024.2336178

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., et al. (2021). Network analysis of multivariate data in psychological science**.** *Nature Reviews Methods Primers*, *1*, 58. https://doi.org/10.1038/s43586-021-00055-w

Brysbaert, M. (2024). Designing and evaluating tasks to measure individual differences in experimental psychology: A tutorial. *Cognitive Research: Principles and Implications, 9*(1), Article 11. https://doi.org/10.1186/s41235-024-00540-2

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review, 58*(5), 313–323. https://doi.org/10.1037/h0054388

Cattell, J. M. (1948). Mental tests and measurements, 1890. In W. Dennis (Ed.), *Readings in the history of psychology* (pp. 347–354). Appleton-Century-Crofts. https://doi.org/10.1037/11304-040

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates Publishers.

Coombs, C. H. (1964). *A theory of data*. Wiley.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671–684. https://doi.org/10.1037/h0043943

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* Irvington.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29.

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The Causal Attitude Network (CAN) model. *Psychological Review, 123*(1), 2–22. https://doi.org/10.1037/a0039802

De Boeck, P., Gore, L. R. (2023). The Janus face of psychometrics. In: van der Ark, L. A., Emons, W. H. M., Meijer, R. R. (Eds.). *Essays on contemporary psychometrics. Methodology of educational measurement and assessment*. Springer, Cham. https://doi.org/10.1007/978-3-031-10370-4_2

De Boeck, P., Pek, J., Walton, K., Wegener, D. T., Turner, B. M., Andersen, B. L., Beauchaine, T. P., Lecavalier, L., Myung, J. I., & Petty, R. E. (2023). Questioning psychological constructs: Current issues and proposed changes. *Psychological Inquiry, 34*(4), 239-257. https://doi.org/10.1080/1047840X.2023.2274429

De Boeck, P. & Wilson, M. (2004). (Eds*.) Explanatory item response models: A generalized linear and nonlinear approach*. Springer. https://doi.org/10.1007/978-1-4757-3990-9

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology.* (H. A. Ruger & C. E. Bussenius, Trans.). Teachers College Press. https://doi.org/10.1037/10011-000

Embretson, S. (1976). Solving verbal analogies: Some cognitive components of intelligence test items. *Journal of Educational Psychology*, *68*(2), 234-242. https://doi.org/10.1037/0022-0663.68.2.234

Embretson (Whitely), S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, 57(2), 94–107. https://doi.org/10.1037/h0058559

Fechner, G. T. (1948). Elements of psychophysics, 1860. In W. Dennis (Ed.), *Readings in the history of psychology* (pp. 206–213). Appleton-Century-Crofts. https://doi.org/10.1037/11304-026

Feyerabend, P. (1975). *Against method*. Verso Books.

Fisher, R. A. (1935). *The design of experiments.* Oliver & Boyd.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370-378.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley.

Heathcote, A., & Matzke, D. (2022). Winner takes all! What are race models, and why and how should psychologists use them? *Current Directions in Psychological Science, 31*(5), 383–394. https://doi.org/10.1177/09637214221095852

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*(3), 1166-1186. https://doi.org/10.3758/s13428-017-0935-1

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. Appleton-Century. JASP Team. (2025). *JASP (Version 0.19.3)[Computer software]*. https://jasp-stats.org/

Jeon, M., & De Boeck, P. (2017). Decision qualities of Bayes factor and *p* value-based hypothesis testing. *Psychological Methods, 22*(2), 340–360. https://doi.org/10.1037/met0000140

Jeon, M., Jin, I. H., Schweinberger, M. et al. Mapping unobserved item–respondent interactions: A latent space item response model with interaction map. *Psychometrika*, *86*(2), 378–403. https://doi.org/10.1007/s11336-021-09762-5

Kuppens, P., Dejonckheere, E., Kalokerinos, E.K. et al. (2022). Some recommendations on the use of daily life methods in affective science. *Affective Science, 3*, 505–515. https://doi.org/10.1007/s42761-022-00101-0

Lee, M.D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, 55(1), 1–7. https://doi.org/10.1016/j.jmp.2010.08.013

Liu, Y., & Pek, J. (2024). Summed versus estimated factor scores: Considering uncertainties when using observed scores. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000644

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement, Vol. 3. Representation, axiomatization, and invariance.* Academic Press.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods, 52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Molenaar, P. C. M. (2008). On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Developmental Psychobiology, 50*(1), 60–69. https://doi.org/10.1002/dev.20262

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology, 112*(3), 474–490. https://doi.org/10.1037/pspp0000100

Neyman, J., & Pearson, E. S. (1933). On the problems of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, 231A,* 289338. https://doi.org/10.1098/rsta.1933.0009

Pek, J.**,** & Flora, D. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *25*(5), 590-605. https://doi.org/10.1037/met0000126

Popper, K. R. (1959). *The logic of scientific discovery*. University Press.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Revelle, W. (2024). The seductive beauty of latent variable models: Or why I don't believe in the Easter Bunny. *Personality and Individual Differences*, 221, 112552. https://doi.org/10.1016/j.paid.2024.112552

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review, 26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heath-cote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*(2), 491–513.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Ryan, O., Dablander, F., & Haslbeck, J. M. B. (2025). Toward a generative model for emotion dynamics. *Psychological Review, 132*(2), 416–441. https://doi.org/10.1037/rev0000513

Spence, K. W. (1936). The nature of discrimination learning in animals. *Psychological Review, 43*(5), 427–449. https://doi.org/10.1037/h0056975

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 100*(3-4), 441–471. https://doi.org/10.2307/1422689

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's: Some interesting parallels. *Psychometrika, 44*(2), 157–167. https://doi.org/10.1007/BF02293967

Sweet, T., & Wang, S. (2025). Network science in psychology. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000745

Traub, R.E. (1997) Classical test theory in historical perspective. *Educational Measurement: Issues and Practices*, *16*(4), 8-14.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. https://doi.org/10.1037/h0070288

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. https://doi.org/10.1037/a0021765

Vanhasbroeck, N., Loossens, T., & Tuerlinckx, F. (2024). Two peas in a pod: Discounting models as a special case of the VARMAX. *Journal of Mathematical Psychology*, *120*, 102856, 1-8. https://doi.org/10.1016/j.jmp.2024.102856

Wang, S., & De Boeck, P. (2022). Understanding the role of subpopulations and reliability in between-group studies. *Behavioral Research Methods*, *54*(5), 2162-2177. https://doi.org/10.3758/s13428-021-01700-8

Widaman, K.F., Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavioral Research Methods, 55*(2), 788–806. https://doi.org/10.3758/s13428-022-01849-w

Young, F. W., & Hamer, R. M. (Eds.). (1987). *Multidimensional scaling: History, theory, and applications*. Lawrence Erlbaum Associates, Inc.