Chapter 2
The Janus Face of Psychometrics

Paul De Boeck and L. Robert Gore

Abstract Most psychometric data are behavioral data: responses to cognitive problems and to questionnaire items referring to behavior in a direct or indirect way. Therefore, measurement models are at the same time psychological models. The Janus face metaphor refers to these two sides of psychometrics. Measurement models can fail as psychological models. We discuss three examples, called vignettes in this chapter. The first refers to reflective measurement models not being in line with the psychology of what is measured. The second example concerns measurement invariance and the psychological meaningfulness of measurement invariance violations. The third example refers to the error variance (unexplained variance) in measurement models and models in general and how the error may be explained by individual-specific psychological phenomena.

Psychological measurement is the quantification of person variables of interest, such as cognition, skills, achievement levels, affect, and motivation, among many others. Psychological tests can quantify rather stable traits, variables subject to growth and change, and states depending on situations and occasions of measurement (Cronbach et al., 1972). Nearly all measurements quantify behavior of the person measured, including introspective self-reports (McFall & Townsend, 1998). Outside of physiology and group sociology, psychologists measure by observing participants and recording or rating their behaviors, using archival records or ratings of behavior, but in most cases, they ask participants to provide self-ratings or quantification of their own behaviors or experiences, and they present cognitive and other problems to work on in tests. Here we do not consider biological measures, such as

P. De Boeck (✉)
Department of Psychology, The Ohio State University, Columbus, OH, USA
e-mail: deboeck.2@osu.edu

L. R. Gore
Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL, USA
e-mail: Bob.Gore@moffitt.org

L. A. van der Ark et al. (eds.), *Essays on Contemporary Psychometrics*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-031-10370-4_2

cortisol, or neuroscience measures, such as fMRI, but focus instead on classically psychological variables such as attitudes, personality traits, moods, and intentions, as well as cognitive variables such as problem-solving, judgments, and response times.

It is noteworthy that often the measurement tool and the measurement object coincide. The tool consists of a person's behavior (e.g., test responses), and the object of measurement is a behavioral proclivity of the same person. This is not the case for the measurement of a person's weight or height. How a scale functions is independent of a person's weight, how a ruler functions is independent of a person's height, and how a thermometer functions is independent of the temperature to be measured. It is only in case that something is wrong with a measurement tool that the measurement tells us something about the tool. If all goes well, a thermometer tells us about the temperature of the room, of the body, etc., and a scale tells us about the weight of another object than itself. Indeed, psychology was born out of the difficulty human judges have in providing objective accounts of physical reality, such as the transit times of stars observed through telescopes (Traub, 1997), such that the observed times reflect something about the observer and not merely something about celestial mechanics and optics.

Because the tool and the objective of measurement are intrinsically inseparable, there always are two sides to psychological measurement, even when the researcher does not realize there are. The measurement model is at the same time a behavioral model, a model for how persons act while being measured. Classical test theory, factor models, and item response models (IRT) are at the same time measurement models and behavioral models. The researcher can focus on the first side and just consider the quantification of target behavioral proclivities or can focus on the implied behavioral model to understand people's test behavior, which in many cases is relevant as such, independent of the measurement outcome. Test items can require reports of knowledge, judgments, or decisions, so that the psychometric models are cognitive models, or test items can be self-descriptive and reflect a person's attitudes, feelings, and behavior, in which case the models are models of attitudes, feelings, and behaviors and models of how people describe themselves. A psychometric model is at the same time a measurement model (a model of the instrument) and a psychological model (a model of the person). The lack of separation and the two sides have inspired us to use the metaphor of a *Janus face.*

For example, in a cognitive test, information is collected to measure cognitive abilities, but the processes underlying the responses are of interest for substantive reasons, independent of the resulting measurement, and this has long been of interest to clinical psychologists (Lezak et al., 2012). The situation is different for the thermometer. We are not measuring the thermometer, and the temperature to be measured is not a feature of thermometer but of an object or space outside the thermometer. We do know how an analog thermometer works (the mechanism) to change heat into measured distance by causing expansion of the mercury in a linear tube, but we do not know with much certainty how a set of cognitive problems works on a person's mind to result in a response and

response time. We can learn from the cognitive test not only what the level of a person's cognitive ability is but, in principle, also to some extent how cognitive problem-solving works. Apart from the test responses themselves, measurement may require assumptions that are driven by domain knowledge, intuition, and potentially also self-reflection or introspection on the part of the researcher, given that the researcher is also an object of the same class (humans) as the object being measured. But some of these assumptions may be unjustified. For example, much response time research assumes that a set of response times is independently and identically distributed. However, this assumption was proven false decades ago (Luce, 1986).

Another aspect of two-sidedness of psychological measurement is that the tool can influence the measured object. Self-monitoring changes smoking behavior, for example (McFall, 1970). Taking a test can affect cognitive processes, such as when a person has a memorable insight that affects their future responses, perhaps years after taking the test (e.g., the Cognitive Reflection Test; Shane, 2005). Responding to questions on one's feelings can affect the feelings, a fact that has been exploited by political opinion pollsters (Gerstmann & Streb, 2004). Measuring a person's weight does not change the weight; measuring a person's temperature does not change their temperature appreciably.

A cognitive ability test yields a measure that is necessarily based on cognitive processes. The two sides to a cognitive test are the measurement and the underlying processes. A measurement model may not fit well with the resulting data, and that kind of failure is informative regarding the underlying focal processes. Ideally, the measurement model is at the same time a process model, which means that the two faces of psychometrics are consistent (McFall & Townsend, 1998). Outside of mathematical models in cognitive science (see, e.g., Ratcliff & McKoon, 2008), this condition is rarely met.

The Janus face of psychometrics implies that psychometric qualities are aspects of human behavior with relevance for psychology. We tend to isolate the measured quantity from its object (e.g., ignoring the response process) and to consider the measurement outcome as the objective output of an impartial instrument. A balance is a measurement instrument, but there is more to a psychological test than its role as a measurement instrument, of which the "psychometric" qualities are to be investigated and reported without psychology itself being at stake. Psychometrics is as much psychology as it is metrics.

In the following, we will discuss three possible cases, which we will call "vignettes," to illustrate the two faces. Vignette 1 concerns the internal consistency of test items, and vignette 2 concerns measurement invariance. Vignette 3 concerns error variance as unexplained variance and offers a clinical and idiographic perspective on that variance. They all three illustrate how a psychometric model is a psychological model and how psychometric qualities of an instrument can reflect important psychological principles.

2.1 Internal Consistency

Cronbach's alpha and alternative coefficients are popular quantifications of reliability. Cronbach's alpha is often interpreted as a measure of internal consistency. However, as Sijtsma (2009) explains, internal consistency is a vague notion. In this first vignette, we refer to internal consistency as an "average degree of 'interrelatedness'" between items (p. 114 in Sijtsma, 2009), which does contribute to coefficient alpha. Internal consistency arises from a reflective psychometric model with one dimension or multiple positively correlated dimensions, where each indicator reflects the latent variable in a direct way and with conditional independence. Not only are there psychologically meaningful other types of reflection than an independent direct reflection, but reflection is not the only way indicators can be linked to a latent variable. In *formative* latent variable modeling, for example, the link is from indicators to the latent variable and the link is cumulative. For both types of links (reflective, cumulative), at least three variants can exist: direct link, competitive link, and intermittent link.

2.1.1 Reflective Models

The three variants of reflective links are investigated by Tuerlinckx et al. (2002) for the Thematic Apperception Test (TAT). The TAT consists of a set of cards, and the respondent is requested to give a narrative interpretation of each card, a description which is believed to reflect underlying situationally specific psychological motivational tendencies.

The default type of reflection is *independent direct reflection* which means that the indicators do not affect one another (i.e., are independent conditional on the latent trait value) and that the reflection is not just intermittent (i.e., does not depend on the occasion). The common factor models and IRT models are in line with independent direct reflection. As a result, the common reliability coefficients for tau equivalent and congeneric models apply, such as the alpha coefficient and the omega coefficient, respectively.

A different type of reflection is *competitive reflection*, which means that reflection through one form of manifestation competes with reflection through other forms of manifestation. In psychometric models this would show through negative local dependencies and a reduction of internal consistency of indicators for the same trait. In the Tuerlinckx et al. (2002) study, the underlying principle is based on the Atkinson and Birch (1970) dynamics of action theory. The implication of the theory is that after an achievement motivation expression, the achievement action tendency is reduced, which shows as a negative effect of a response on the next response and thus as negative serial dependence. Competitive reflection is a more general phenomenon based on the dynamics of action theory. Any time there is restriction of resources related to the expression of a trait, competition follows. Time and

finances are examples of resources one needs when one follows interests related to leisure activities and social activities. One can have only so many interests, so many leisure activities, so many social activities, independent of the strength of one's needs, the breadth of one's interest, and the intensity of one's social motivation. Other principles at the basis of competition are habit formation and specialization. Habits may exclude other habits, as anxiety finds its expression in specialized fears, extreme political opinions fixate on certain topics and not on other topics associated with one's adversaries, and the set of fixations flowing from a particular ideological point is in flux. It is in theory possible that internal consistency of a test is very low and even zero or negative, although the indicators are all indicators of the same trait, albeit competing indicators.

Related to this but not discussed by Tuerlinckx et al. (2002), we might posit its opposite: accelerating reflection. This might occur when a behavior, once emitted, tends to raise the tendencies toward similar behavior. An example from social psychology is priming (Molden, 2014). A person who cooperates with an experimental confederate in one task may become more likely to cooperate on the next. On a measure of personality traits, as a person scans memory for examples of a particular trait (consider generosity for example), more such memories may come to mind, such that their proclivity to agree with similar trait descriptors increases as their progress through the test continues. On a multi-factorial test with shuffled items, this would suggest that the internal consistency of items grows from the first to the second half of the test.

The third type of reflection is *intermittent reflection*, which means that a trait is reflected only now and then but not on all occasions. Tuerlinckx et al. (2002) use the term "stochastic drop out" for this phenomenon. For example, a person can have a high need for achievement, but the need does not show at all possible achievement occasions. For the TAT, that would mean the need would not be reflected in the responses to all cards, which is a possibility suggested by Murray (1943, p. 15). When intermittent reflection is random, it is formally equivalent with an upper-asymptote model (as in the four-parameter model, but with a zero lower asymptote). What this means is that there always is a chance that the need for achievement is not expressed, which implies that the maximum probability (the upper asymptote) of an achievement response is smaller than 1.00. Dependent on the card, the upper asymptote is higher or lower (Tuerlinckx et al., 2002). When a response drops out of the normal response process, the response does not reflect the respondent's need for achievement, but instead some other need-induced phantasy is reflected in response to a TAT card. The assumption that intermittent reflection is random (conditional on the level of the upper asymptote) is an approximation for the fact that many different needs may take over to be expressed, conditionally independent of the achievement-related content of the card, induced by the varying strength of those other needs. Without the simplifying approximation with an upper asymptote, it would be too complex a model to be estimated, although it is possible to simulate the resulting intermittent reflection phenomena based on the dynamics of action theory (Atkinson & Birch, 1970). Intermittent reflection is the consequence of changing competitive strengths as postulated in the dynamics of action theory.

To give an example from another behavioral domain, suppose that a trait we are interested in is punctuality. Behaviors such as showing up on time for an appointment and making deadlines in time can be interpreted as reflections of punctuality. It is possible though (depending on the occasion) that another trait takes over to determine the behavior, which may lead to a violation of punctuality. For example, helpfulness may take over from punctuality if one needs more time than expected to solve someone else's problem, with consequences for the next appointment. Other traits taking over from a trait one wants to measure can explain intermittent reflection. Whereas competitive reflection refers to competition between indicators, intermittent reflection may refer to competition between traits for expression in a single behavior. Like competitive reflection, intermittent reflection also reduces the internal consistency. In physics, Brownian motion describes the process by which a dust particle is buffeted by random atomic collisions causing it to drift around in still air. Psychological tendencies may have a similar character, buffeting behavior in different directions depending partly on truly random factors. This cannot be explained by an error term when the whole response itself is captured by another tendency related to a trait one does not intend to measure, just as for an upper-asymptote model, the response cannot be captured by the common notion of an error term. An upper-asymptote IRT model is a mixture model for the pairs of persons and items, just as the three-parameter IRT model with a lower asymptote also is a mixture model.

Based on the empirical application in Tuerlinckx et al. (2002), the model with intermittent reflection was the best-fitting model for the TAT. The dropout probability in a constrained (but well- fitting) model with a common upper asymptote for all cards was 0.34, a close approximation of Murray's (1943) guess that 30% of responses are nondiagnostic responses.

2.1.2 Cumulation Models

As mentioned earlier, an alternative to reflection is *cumulation*, as in formative models. To explain the concept, let us use the example of happiness and assume that happiness has different sources (referring to different aspects of life). Let us further assume that the happiness from these different sources adds up (i.e., accumulates): relational happiness, happiness in one's job, and leisure time happiness. These sources do not need to be correlated, but they can as in the following examples. When people experience less happiness from one source, they may compensate by seeking and obtaining more happiness from another source, just as different sources of income add up and one source can compensate for another. Alternatively, when people reach a threshold of happiness, they may stop seeking happiness from untapped sources. In both these cases, the correlation would be negative. Independent of the relationships, if the sources of happiness add up, happiness is a cumulative trait, and internal consistency must not be expected. Control theory more generally describes a variety of phenomena where a person, motivated to

maintain homeostasis, experiences changes in appetite and behavior due to variation in goal satisfaction levels (Carver & Scheier, 1982). For accumulation, the same three types as for reflection can exist: independent direct accumulation, competition (and acceleration), and intermittence. The above examples of compensation and satisfaction are formally equivalent with competition as they lead to negative correlations. Possibly, the different sources of happiness do literally compete with one another or reinforce each other. For example, if happiness depends on the time invested in the sources of happiness (i.e., codetermines how much happiness is derived from the source), then investing in one's job may come at the cost of investing in relationships, which may lead to a negative correlation. It also is possible that happiness in one respect of life carries over to other respects of life. Intermittent cumulation would imply that the same source does not always contribute to one's happiness, depending on one's focus of the occasion. As a result, an inventory of pleasant activities a person has enjoyed (MacPhillamy & Lewinsohn, 1982), for example, does not necessarily lead to a high internal consistency.

The different kinds of reflection and accumulation illustrate how internal consistency is not just a measurement quality but a possible indicator of psychological processes. From a measurement point of view, a high internal consistency may seem desirable, while from a psychological point of view, a low internal consistency may be a meaningful result, even when it would lead to a low coefficient alpha value.

2.2 Measurement Invariance

2.2.1 Relevance of Measurement Invariance and Its Violation

Psychologists would like to quantify differences and changes to understand influences on human behavior. It is a well-known rule that measures cannot be compared if the condition of measurement invariance is not met (Millsap, 2011). A violation of the condition implies that using the same instrument results in measures of different variables, as if a scale does not always measure weight but sometimes quantifies volume or height instead. As a result, variations in the numeric output of instruments may be the quantification of dissimilar qualities, while the person doing the measuring believes they are comparing dissimilar individuals on the same quality.

Violations of measurement invariance are interpreted as an issue and may lead to adjustments of the measure. Rarely are the violations interpreted as interesting psychological phenomena, while a result that is undesirable from a *metric* perspective can be helpful from a *psychological* perspective. To illustrate, after a psychotherapeutic intervention, it can be expected that a trait is expressed in a different way and that the same behavior (the same response to an item) now has a different meaning. After a treatment for anxiety, perhaps not only the level of anxiety is reduced, but the threshold of some fearful behaviors has increased (a change

of the intercept, corresponding to a violation of scalar invariance), or previous behaviors driven by anxiety are now carried out for other reasons when they occur (a change of the loadings, corresponding to a violation of metric invariance). The lowering of a threshold for a fearful behavior, independent of an overall decrease of anxiety, is reflected in the intercept parameter of the behavior and is a violation of scalar invariance comparing pre- and post-intervention conditions. The lowering of a loading means that anxiety has less influence on the behavior and is a violation of metric invariance. These formal kinds of differences in thresholds and in the relationship with an underlying latent variable may exist between groups, across gender categories, between ethnic groups, between cohorts, between cultures, and between different points in time for the same set of persons, such as before an intervention and after an intervention. Fokkema et al. (2013) find evidence for such processes, called response shifts, with respect to depression.

The differences between groups and within groups across time that correspond to violations of measurement invariance will be called qualitative differences, and they can be of interest as psychological phenomena as such even though they interfere with the conditions of measurement invariance. One may have to give up on making inferences about quantitative differences such as differences between the means of a latent variable or between sum scores, but instead one may follow up on the specific violations and make inferences on qualitative differences instead.

For example, developmental psychologists hope to measure the growth in logical reasoning and vocabulary across childhood, and clinical psychologists hope to measure the reduction in anxiety, depression, or addictive cravings resulting from intervention. Social psychologists hope to measure geographical differences and cohort effects in implicit bias. To measure changes and differences requires that the measuring instrument preserve the conditional relationship between the behavioral proclivity as input and the output of the instrument, such as a sum score, across groups (such as geography) and time (in development or treatment outcome studies). This is considered a precondition for valid measurement, but its violations are interesting phenomena themselves, and violations may be a foreseeable result of the psychologist's theory of difference or change.

2.2.2 An Example

To give a clinical example, consider a group of spider-phobic undergraduates who have heretofore avoided spiders at all costs. Imagine that these individuals undergo a single-session arachnophobia treatment and that they make ratings of anxiety (0 = "not at all anxious" to 100 = "as anxious as you have ever been or could imagine being") before and after the session, in response to different items such as mentally imagining a spider, viewing a real spider, and touching a spider (which they may never have done), each time regarding a spider that sits still or crawls. In the sessions, the individuals approach and eventually touch spiders provided by the researcher. As a result of the treatment itself, anxiety ratings provoked by really

touching spiders may arise from different processes, perhaps with a clearly lower anxiety rating for touching spiders, while ratings of anxiety levels on seeing and thinking of spiders may have decreased less. It also is possible that the scale is recalibrated after the experience and not just by an additive constant for all items. In both cases there would be a violation of scalar measurement invariance, but the finding may be informative about the specific effects of the session, even though no inference can be made regarding an increased or decreased fear for spiders as measured by a latent variable or a sum score. The measurement invariance failure (viewed conventionally) invalidates the evidence base for the treatment.

In some cases, psychological theory predicts that a group effect or an intervention effect is different for a subset of items compared with other items. The Saltus model (Wilson, 1989) is a model for theory-based violations of measurement invariance in which a subset of items shares a common violation of scalar invariance, for example, a subset of Piagetian tasks becomes more difficult or easier with age.

2.2.3 Mathematical Models and Clinical Interpretations

Psychologists can formulate models simultaneously of the person being measured and the process of measuring the person. Such models could incorporate shifts in the judgments (such as in the arachnophobia example) and more complicated shifts in the meaning of measures. Although one way to jointly examine the state or trait being measured along with the response process would be with sophisticated, tailored mathematical models (McFall & Townsend, 1998), that would not be the only way. Psychologists who listen to the people who provide the measures, who give the measured the voice to speak about their experience of providing numbers or quantifiable behaviors, could gather a great deal of useful information in qualitative form (e.g., mixed methods research, Tashakkori & Teddie, 2003). Practicing clinical psychologists do this routinely and may find qualitative evidence that makes violations of measurement invariance interpretable. Clinical interpretations may also open the black box of error variance or unexplained variance and try to understand the particularity of individual human behavior and an individual person's life, as discussed in the following.

2.3 Error Variance and Unexplained Variance

Error variance is a common parameter in psychometric models, in classical test theory, in factor models, and in item response theories (if formulated in terms of latent responses). From a statistical and measurement point of view, error is specific yet unexplained variance: specific to the measurement indicator in question, unrelated to other measurement indicators (and unrelated to the latent variable or true score in CTT), and therefore unexplained.

2.3.1 Two Views

An interesting view on error variance is Kahneman's (2011) distinction between statistical thinking and causal thinking. Error variance is an example of what he calls statistical thinking, foregoing the meaningful effects of specific events and circumstances in a person's life. What Kahneman understands by causal thinking is thinking based on individual cases and individual events instead. Individual events and circumstance may affect a person's behavior and responses in a test, and such effects are globally summarized in error variance – (Kahneman et al. (2020) use the term "noise") – while they may refer to psychologically meaningful phenomena that cannot be captured because the events and circumstances are person-specific and not part of the design (van Bork, 2019). Common sense causal thinking is often necessary in situations where causal inference is statistically underdetermined, and this kind of common sense has a place in psychological science.

When psychologists shift from academic to applied roles, in some cases the importance of peer-reviewed, published analyses of reliability (and validity) increases, while they may benefit from qualitative information on possible sources of the error variance that leads to a lower reliability. Whereas the statistical way of thinking is important, a more individualized approach in the line of "causal thinking" can be a useful complementary perspective.

2.3.2 An Example

Consider a parent whose fitness has been questioned in a contentious divorce proceeding. The parents in such a case may be court ordered to undergo an extensive psychological evaluation, which in some areas of the United States may include psychological testing (such as with the MMPI-2-RF), extensive reviews of background records including criminal background checks and children's medical records, interviews with people who know the parent and their children well, observation of the children's behavior with and without each parent present, and diagnostic interviewing of each parent (see vignette in Emery et al., 2005). Each of the resulting scores is an evaluation component, but these scores also contain error variance. However, it would be impractical and too ambitious to quantify fully the amount of this error in real-world, high-stakes cases, and adding individual information for a clinical judgment may be problematic. Clinical judgment research suggests that clinicians should be modest in their claims because complex constellations of additional individual information have been shown to be rife with judgment error (Garb, 2003; Dawes et al., 1989).

The interpretation of the joint collection of information components is subjective: different clinicians confronting the same collection of aggregated data could reach different conclusions (Garb, 1989), and the attorneys and judges in the case may select, block, amplify, and downplay different aspects of the record. The kind of

extensive evaluation performed in child custody cases may run to 100 pages, and each person who reads the record will no doubt face problems of how to consider the mass of information provided. Issues of selective attention and recall and individual bias will enter the process. In principle the content in the child custody record is individual information that may explain psychometric error. Unfortunately, relying on human judgment may not be a good way to interpret the information.

2.3.3 Two Issues

This third vignette highlights two issues. The complexity of error variance is not captured in an estimate of its size, whereas the multiple sources are psychologically meaningful and can be noticed in individual cases. This could help inoculate non-psychologists and psychologists alike against any tendencies to ascribe exclusive meaning to the global psychometric information while being blind to the qualitative information.

However, as a second point, there may be objective quantitative indicators of severe problems with reliability that ought to be highlighted for any users of the data. If, for example, an observed MMPI-2-RF score profile is to be used to comment on the future parenting abilities of the parties over the course of several years, and if the observed fluctuation in MMPI-2-RF scores across measurement occasions separated by a much shorter interval is such that the use of the test to forecast years into the future is in doubt, this fact is crucial (Faust, 2012). In a case such as this, the size of the unexplained variance is vitally important to the fair application of psychology in forensic settings, and speculations on a qualitative basis and causal thinking (in Kahneman's terms) may be largely misleading.

2.3.4 Clinical and Statistical

This vignette also highlights the potential value of training clinical psychologists to engage in nuanced analyses of their measurement procedures and the ways their findings are processed by end users. If applied psychologists were systematically trained to avoid focusing so narrowly on the justification of their measures with specific coefficients and instead were taught to think of their measures as intrinsically influenced by the contexts of measurement and the motivations and cognitions test takers have in relation to their performance, the quantitative indications of error variance would not be interpreted as the final word. What seems to be error – from a statistical point of view – may correspond to meaningful events in a person's life.

The challenge for decision-making may be to integrate across multiple sources of information. It is well-known from the judgment and decision-making literature that simple methods such as equal weighting of standardized scores (called improper linear models) could be useful, as illustrated, for example, by Dawes (1979). To

formulate an improper linear model, a set of judges rates a set of objects on a set of attributes. Ratings for each attribute are standardized across judges, and sums or means of standardized scores across attributes form the overall score. It is incumbent on psychology to educate end users of high stakes tests about the many sources of unexplained variance and imperfect validity. We also need to help test users find ways to reconcile a statistical approach for an optimization of prediction across a set of persons with an awareness that unforeseeable variation may invalidate predictions and decisions. Yet the complexity of this task is daunting.

Statistical reasoning may be the optimal way from a global perspective and across the set of individuals under consideration, but this does not guarantee it is the optimal way in individual cases where idiographic information is available about an individual's specific circumstances. From the perspective of causal reasoning in Kahneman's (2011) terms, which is more idiographic than statistical, the error and unexplained variance reflect meaningful information with consequences for how a test result should be interpreted. The problem with such an approach is that human judgment suffers from various shortcomings as amply described by Kahneman (2011) and Kahneman and Tversky (1996), which also explains why a statistical (actuarial) approach frequently works better than clinical judgment for predictive purposes (Dawes et al., 1989).

2.3.5 An Idiographic Alternative

A possible alternative for purely clinical judgment without giving up on individualized information is quantitative idiographic measurement, with different variables for each individual person, and within-person relationships of those variables across situations or stimuli. For example, people may be asked to rate their feelings toward important others in their life, while they each choose their own feeling terms as well as the important others. The data can then be analyzed in an objective way, for example, using cluster analysis or a dimensional analysis. Such approaches may be a way to counter the subjectivity and biases inherent to human judgment. Examples of such an approach can be found in Kelly's (1955) personal construct theory approaches based on the repertory grid and in Herman's self-confrontation method (Hermans, 1991; Lamiell, 1991; Lyddon et al., 2006). A method of Boolean factor analysis and cluster analysis for within-person data matrices that may help to understand the particularity of individual persons can be found in De Boeck and Rosenberg (1988) and Van Mechelen and De Boeck (1989). In addition, the application of idiographic data collection as for the repertory grid and for the self-confrontation method (the measurement tool) may have an effect, hopefully a beneficial effect, on the individuals being measured (the objects of measurement). Although more useful for understanding than for prediction, these methods may help to have a meaningful view on individualized factors that may contribute to unexplained variance in measurement models.

Most methods and the whole field of psychometrics are focused on an interindividual variance paradigm. The inherent complexity of psychological phenomena may require a somewhat different paradigm, with a stronger focus on intraindividual approaches. This may lead to a better understanding of what shows as measurement error in methods based on interindividual variance. While a qualitative way makes sense in the context of discovery, a more quantitative intraindividual approach can take care of the justification.

2.4 Discussion

It was not within the scope of this article to provide a compendium of statistical solutions, but rather to exemplify the psychology side of psychometrics. Our goal here was in the first place to shift perception. Janus faces were posted on gates, so that travelers coming and going saw different faces. As a result, Janus was thought to see both the past and the future. While entering a domain, one would see a particular face, and while leaving one would see a different face. At the start of an investigation, the researcher sees one of the Janus faces, and when the results are in, the researcher may see the other, one that could be disappointing from a measurement quality point of view but informative from a psychological point of view.

Psychology can continue its traditional attempt to separate the measurement tool from the human proclivities being measured, or it can turn around and regard Janus's other face: the face that might smile on us as we change course, perhaps even reverse course partly. Just as we have tended to regard test instruments as objective reflections of behavioral proclivities and to try to develop instruments that achieve this purpose, we have also tended to regard statistical procedures in the same way, and we have developed a reflex (and trained it into our students) according to which certain forms of reliability or measurement invariance have to be established before a measure can be considered worthy of use, and typically this demonstration relies on standard statistical methods such as confirmatory factor analysis, item response theory, or the computation of reliability coefficients such as alpha and test-retest reliability. But there is another path, which is to learn about psychology from so-called psychometric shortcomings and to use these indications and psychological theory to formulate models in line with measurement principles but also with psychological processes. Such models could incorporate shifts in the judgments (such as in the arachnophobia example) and more complicated shifts in the meaning of measures. Although one way to improve the meaningfulness of clinical methods would be with sophisticated, tailored mathematical models, that would not be the only way.

If perception shifted, and we began to regard the measurement as double-faced, we would be less apt to offer sweeping generalizations about human behavior that ultimately undermine our credibility when they are swept away by the facts in an individual person's life or in the next round of generalization in research. What

might result would be an approach to measurement that equally respects the two sides, that respects and provides a place for insights developed through looking into possible violations of measurement qualities and qualitative sources of error variance, to proceed more cautiously to conclusions. We suggest that this might also reduce some of the tendency psychologists have shown toward acrimonious debate and would provide legitimacy for researchers seeking to diversify the range of cultural contexts in which psychological research findings can be applied.

We do not want to replace quantitative approaches with qualitative ones. In the context of predicting specific outcome variables in an individual-differences paradigm, a statistical approach is clearly superior to a clinical approach, and adding qualitative information may not improve predictive accuracy, most likely because clinical judgment is vulnerable to distortions of various kinds. However, qualitative information may contain hints about prediction errors rooted in people's individual contexts. Hints are not proofs, but they help explain the omnipresence of errors and how such errors reflect the complex psychology of individual persons.

Reflecting on the Janus face of psychometrics may help us admit that our understanding of the world is only very partially captured by the current quantitative models we use and that deviations can refer to (1) meaningful but deviating models as discussed in the first vignette, (2) meaningful violations of measurement invariance as discussed in the second vignette, and (3) meaningful content of what is commonly called measurement error. The result might be a more investigative attitude, a stronger awareness of the two-sidedness of psychometric models, an openness to alternatives for the most prominent measurement models (CTT, confirmatory factor models, item response theory), and an awareness that replication and prediction failures do not necessarily stem from measurement shortcomings but are inherent to the meaningful complexity of the psychological reality (De Boeck & Jeon, 2018; De Boeck et al., 2019, 2021).

In his article on Cronbach's alpha, Sijtsma (2009) describes the unfortunate gap between psychology and psychometrics, which shows in misunderstandings and lack of interest from both sides. The gap has also led to the perception of psychometrics as an extraneous technical discipline with its own criteria and to the perception of psychometricians as gatekeepers and law enforcement agents. This view is not surprising, because psychometric models are usually not inspired by psychology (Borsboom, 2006). However, we believe that psychometric models are psychological models by implication, although primarily inspired by metric principles, and that psychometrics cannot be just a toolbox kind of discipline. The two faces of psychometrics cannot be separated.

References

Atkinson, J. W., & Birch, D. (1970). *The dynamics of action*. Wiley.

Boeck, P. D., & Rosenberg, S. (1988). Hierarchical classes: Model and data analysis. *Psychometrika, 53*(3), 361–381. https://doi.org/10.1007/BF02294218

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality-social, clinical, and health psychology. *Psychological Bulletin, 92*(1), 111–135. https://doi.org/10.1037/0033-2909.92.1.111

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Dawes, R. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin, 144*(7), 757–777. https://doi.org/10.1037/bul0000154

De Boeck, P., Jeon, M., & Gore, L. (2019). Beyond registration pre and post. *Computational Brain & Behavior, 2*(3), 183–186. https://doi.org/10.1007/s42113-019-00063-w

De Boeck, P., DeKay, M. L., Gore, L. R., & Jeon, M. (2021). The trees and the forest: Investigating variability surrounding an aggregate result. *Theory and Psychology, 31*(3), 399–404. https://doi.org/10.1177/09593543211016084

Emery, R. E., Otto, R. K., & O'Donohue, W. T. (2005). A critical assessment of child custody evaluations: Limited science and a flawed system. *Psychological Science in the Public Interest, 6*(1), 1–29. https://doi.org/10.1111/j.1529-1006.2005.00020.x

Faust, D. (2012). *Coping with psychiatric and psychological testimony* (6th ed.). Oxford University Press.

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment, 25*(2), 520–531. https://doi.org/10.1037/a0031669

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105*(3), 387–396. https://doi.org/10.1037/0033-2909.105.3.387

Garb, H. N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychological Assessment, 15*(4), 508–520. https://doi.org/10.1037/1040-3590.15.4.508

Gerstmann, E., & Streb, M. J. (2004). Putting an end to push polling: Why it should be banned and why the first amendment lets congress ban it. *Election Law Journal: Rules, Politics, and Policy, 3*(1), 37–46. https://doi.org/10.1089/153312904322739916

Hermans, H. J. M. (1991). The person as co-investigator in self-research: Valuation theory. *European Journal of Personality, 5*(3), 217–234. https://doi.org/10.1002/per.2410050304

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Starus, and Giroux.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*(3), 582–591. https://doi.org/10.1037/0033-295X.103.3.582

Kahneman, D., Sibony, O., & Sunstein, C. R. (2020). *Noise: A flaw in human judgment*. Little.

Kelly, G. (1955). *On the psychology of personal constructs*. Norton.

Lamiell, J. T. (1991). Valuation theory, the self-confrontation method, and scientific personality psychology. *European Journal of Personality, 5*(3), 235–244. https://doi.org/10.1002/per.2410050305

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). Oxford University Press.

Luce, R. D. (1986). *Response times: Their role in inferring elementary psychological organization*. Oxford University Press.

Lyddon, W. J., Yowell, D. R., & Hermans, H. J. M. (2006). The self-confrontation method: Theory, research, and practical utility. *Counseling Psychology Quarterly, 19*(1), 27–43. https://doi.org/10.1080/09515070600589719

MacPhillamy, D. J. & Lewinsohn, P. M. (1982). The pleasant events schedule: Studies on reliability, validity, and scale intercorrelation. *Journal of Consulting and Clinical Psychology, 50*(3), 363–380. https://doi.org/10.1037/0022-006X.50.3.363

McFall, R. M. (1970). Effects of self-monitoring on normal smoking behavior. *Journal of Consulting and Clinical Psychology, 35*(2), 135–142. https://doi.org/10.1037/h0030087

McFall, R. M., & Townsend, J. T. (1998). Foundations for psychological assessment: Implications for cognitive assessment in clinical science. *Psychological Assessment, 10*(4), 316–330. https://doi.org/10.1037/1040-3590.10.4.316

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Taylor and Francis. https://doi.org/10.4324/9780203821961

Molden, D. C. (2014). Understanding priming effects in social psychology: An overview and integration. *Social Cognition, 32*(Supplement), 243–249. https://doi.org/10.1521/soco.2014.32.supp.243

Murray, H. A. (1943). *Thematic apperception test manual*. Harvard 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

Shane, F. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Tashakkori, A., & Teddie, C. (2003). *Handbook of mixed methods in social & behavioral research*. Sage.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8–14. https://doi.org/10.1111/j.1745-3992.1997.tb00603.x

Tuerlinckx, F., De Boeck, P., & Lens, W. (2002). Measuring needs with the thematic apperception test: A psychometric study. *Journal of Personality and Social Psychology, 82*(3), 448–461. https://doi.org/10.1037/0022-3514.82.3.448

Van Bork, R., (2019). *Interpreting psychometric models*. Unpublished doctoral dissertation. University of Amsterdam.

Van Mechelen, I., & De Boeck, P. (1989). Implicit taxonomy in psychiatric diagnosis: A case study. *Journal of Social and Clinical Psychology, 8*(2), 276–287. https://doi.org/10.1002/per.2410040207

Wilson, M. (1989). Saltus: A psychometric model or discontinuity in cognitive development. *Psychological Bulletin, 105*(2), 276–289. https://doi.org/10.1037/0033-2909.105.2.276