

Exploring and Validating Construct Interpretations of Psychological Measurements

Dissertation zur Erlangung des akademischen Grades
Doktor der Philosophie (Dr. phil.)

Vorgelegt im Fachbereich 01 Humanwissenschaften
der Universität Kassel

Von Leon Patrick Wendt

Kassel, Februar 2024

Tag der Disputation: 20. Juni 2024

Examiners [Gutachter]: Prof. Dr. Johannes Zimmermann (Universität Kassel)
Prof. Dr. Ulrich Schroeders (Universität Kassel)
Prof. Dr. Cord Benecke (Universität Kassel)

Acknowledgements

My gratitude goes to my family—Hanna, Hermann, Marlon, Lena, Viktoria, Adrian, and Ulla—for their support and encouragement. Special thanks to Eugenia, whose encouragement and support have been pivotal throughout this journey. I appreciate my friends—Chris, Andi, Jan, Daniel, Maj-Lisa, Niko, and others—and my colleagues—Steffen, Kristin, Gabriel, and others—for their camaraderie and friendship. I am deeply thankful to my advisor, Johannes, for his invaluable guidance and insight. I am also grateful to my co-authors for their essential contributions and insights. To all those mentioned and not mentioned, my sincere thanks.

Table of Contents

Part 1 Exploring and Validating Construct Interpretations of Psychological Measurements [Manteltext]

How to cite: Wendt, L. P. (2024). *Exploring and validating construct interpretations of psychological measurements* [Doctoral dissertation, University of Kassel]. KOBRA [Kasseler Online Bibliothek, Repository und Archiv]. <https://doi.org/10.17170/kobra-2024070110450>

Part 2 Article 1. Indicators of Affect Dynamics: Structure, Reliability, and Personality Correlates

How to cite: Wendt, L. P., Wright, A. G., Pilkonis, P. A., Woods, W. C., Denissen, J. J., Kühnel, A., & Zimmermann, J. (2020). Indicators of affect dynamics: Structure, reliability, and personality correlates. *European Journal of Personality*, 34(6), 1060-1072. <https://doi.org/10.1002/per.2277>

Part 3 Article 2. Mapping Established Psychopathology Scales Onto the Hierarchical Taxonomy of Psychopathology (HiTOP)

How to cite: Wendt, L. P., Jankowsky, K., Schroeders, U., London Personality and Mood Disorder Research Consortium, Nolte, T., Fonagy, P., Montague, P. R., Zimmermann, J., & Olaru, G. (2023). Mapping established psychopathology scales onto the Hierarchical Taxonomy of Psychopathology (HiTOP). *Personality and Mental Health*, 17(2), 117-134. <https://doi.org/10.1002/pmh.1566>

Part 4 Article 3. Mindreading Measures Misread? A Multimethod Investigation Into the Validity of Self-Report and Task-Based Approaches

How to cite: Wendt, L. P., Zimmermann, J., Spitzer, C., & Müller, S. (2024). Mindreading measures misread? A multimethod investigation into the validity of self-report and task-based approaches. *Psychological Assessment*, 36(5), 365-378. <https://doi.org/10.1037/pas0001310>

Exploring and Validating Construct Interpretations of Psychological Measurements

Abstract: In psychological research, as in other scientific disciplines, developing accurate measures of the phenomena under study is fundamental to maintaining the integrity of scientific conclusions. Psychological researchers are often criticized for a lack of rigor when it comes to measurement. On closer inspection, however, these problems may be due not only to the lack of implementation of common standards, but also to fundamental methodological flaws in those standards. Measurement methodology within psychological research has been plagued by many challenges, including conceptual complexity and imprecise terminology, lack of consensus, misplaced emphasis on aspects of psychometric purity, a prevailing inclination toward confirmationism over falsificationism, a lack of distinction between exploratory and confirmatory research, and the absence of a framework fully dedicated to exploratory work, depriving researchers of essential tools. This dissertation aims to address some of these challenges and further presents studies that examine construct interpretation of psychological measurements in three research areas: affect dynamics (Wendt et al., 2020), psychopathology (Wendt et al., 2023), and mindreading (Wendt et al., 2024).

Zusammenfassung: In der psychologischen Forschung wie auch in anderen wissenschaftlichen Disziplinen ist die Verfügbarkeit präziser Messungen der zu untersuchenden Phänomene von grundlegender Bedeutung für die Genauigkeit wissenschaftlicher Schlussfolgerungen. Psychologieforschende sehen sich jedoch häufig mit dem Vorwurf mangelnder Sorgfalt konfrontiert. Bei näherer Betrachtung sind diese Probleme jedoch nicht nur auf die mangelnde Umsetzung gängiger Standards zurückzuführen, sondern auch auf grundsätzliche Probleme der Methodik. Die vorherrschende Methodologie in der psychologischen Diagnostik ist mit zahlreichen Problemen konfrontiert: Unpräzise Terminologie bei konzeptueller Komplexität, mangelnder wissenschaftlicher Konsens, Überbetonung von Aspekten psychometrischer Reinheit, eine vorherrschende Tendenz zum Konfirmationismus anstelle des Falsifikationismus, mangelnde Differenzierung zwischen explorativer und konfirmatorischer Forschung sowie das Fehlen eines Frameworks für explorative Studien. Diese Dissertation versucht, einige dieser Herausforderungen zu adressieren, und stellt drei Studien vor, die Konstruktinterpretationen psychologischer Messungen in den Forschungsbereichen Affektdynamik (Wendt et al., 2020), Psychopathologie (Wendt et al., 2023) und Mentalisierungsfähigkeit (Wendt et al., 2024) untersuchen.

Introduction

In psychological research, as in other fields of science, the development of measurements for the phenomena under study is foundational, and ensuring these measurements are accurate—namely, free from error—is critical to preserving the integrity of scientific conclusions (Flake & Fried, 2020; Schimmack, 2021; Vazire et al., 2022; Zumbo, 2007). Various types of measurement error and their adverse impact on psychological research have been extensively documented (e.g., Bagozzi et al., 1991; Cole & Preacher, 2014; Podsakoff et al., 2024; Schmidt & Hunter, 1996, 1999). Despite this, concerns remain that researchers often do not adequately address the validity of measurements (Borsboom, 2006; Cronbach, 1989; Flake & Fried, 2020; Kane, 2017; Maul, 2017; Schimmack, 2021; Strauss & Smith, 2009). Particularly Schimmack (2021) warned of a potential validation crisis that could undermine the credibility of psychological research, drawing parallels to the replication crisis and the ensuing open science reforms of the 2010s, and called for a focus on measurement in the 2020s. This dissertation addresses Schimmack's call by examining interpretations of psychological measurements in three areas: affect dynamics (Wendt et al., 2020), psychopathology (Wendt et al., 2023), and mindreading ability (Wendt et al., 2024).

In its broadest meaning, the term validity encompasses all sorts of methodological issues that can affect the quality of research outcomes (e.g., Shadish et al., 2002; Vazire et al., 2022). This includes the validity of statistical conclusions (e.g., Starns et al., 2019), the generalizability of findings across different contexts (e.g., Yarkoni, 2022), and the accuracy of causal inferences (e.g., Rohrer, 2018). Within the specific context of measurement, the term validity has several distinct meanings (Newton & Shaw, 2013). It can refer to the accuracy with which a test captures the intended psychological characteristic (Cronbach & Meehl, 1955), its practical utility in predicting significant outcomes (Cureton, 1951), or a global evaluative judgment that integrates these two aspects (e.g., AERA, APA, & NCME, 2014). Recent discourse in the measurement field reveals further nuances in the use of the term validity, often broadening the meaning and scope of established validity concepts.

Two principles apply to all validity concepts in the context of measurement. First, while validity may be simplistically described as a property of a test, a more precise understanding recognizes the *test score* as the actual bearer of validity (e.g., Cureton, 1951; Cronbach & Meehl, 1955; Kane, 2013). The term *test score* can refer to any numerical value derived from empirical observations, extending beyond the confines of formally developed tests, as targeted in the second and third article of this dissertation (Wendt et

al., 2023, 2024), to include spontaneously composed scales, as part of the second article (Wendt et al., 2023), and person-specific summary statistics of individual time series, as targeted in the first article (Wendt et al., 2020). Test scores are the result of scoring procedures that use methods of aggregation, coding schemes, or diagnostic algorithms (Messick, 1995). The distinction between test and test score is critical for two reasons: first, it is the test score, not the test itself, that is the subject of empirical analysis or evaluation; and second, different scoring procedures applied to the same set of observations can differ markedly in validity (Markon et al., 2011; Müller et al., 2022; Wendt et al., 2019). The second principle, common to all conceptions of validity, is that a validity judgment is bound to a particular interpretation or intended use of a test score for a particular population of individuals (e.g., Appelbaum et al., 2018; Cronbach & Meehl, 1955; Cureton, 1951; Kane, 2013).

Researchers can be confused by the conceptual and terminological intricacies of validity (e.g., Borsboom & Wijsen, 2016; Newton & Shaw, 2013). Consequently, the first chapter of this dissertation seeks to elucidate validity concepts, with particular emphasis on construct validity (e.g., Smith, 2005), given its significance to the articles included in this cumulative dissertation. The second chapter takes a critical view of the widespread belief that any research activity that examines the construct meaning of psychological measures can be considered construct validation. It is argued that many studies are more akin to what might be better described as construct exploration. The third chapter highlights the need for, and current lack of, a framework for construct exploration and briefly outlines how such a framework could be developed. The fourth chapter reviews the three articles included in this cumulative dissertation, contextualizing them within the methodology outlined in previous chapters. The fifth chapter provides a general discussion of how methodology and practice in psychological measurement can be advanced to move the field forward.

I. What is Validity?

One of the earliest documented definitions of validity is attributed to Kelley (1927), who stated that a measure is valid if it "measures what it purports to measure" (p. 14). Kelley's definition, notable for its intuitiveness and appeal, is still frequently referenced in the contemporary literature. Although Kelley's description was concise, subsequent definitions have provided more depth, clarified *what* is to be measured, and developed specialized methodologies for *how* to conduct a validation. Nonetheless, these later definitions essentially adhere to the form of Kelley's classic definition, which can thus be said to mark a common ground for all validity concepts (Loevinger, 1957).

The Criterion Validity Model

The criterion model, recognized as the first validity model, defines validity in terms of the utility of a test score in predicting an important criterion variable, typically tied to a practical purpose (e.g., Cureton, 1951). This model is particularly relevant in applied psychological fields where decision-making based on assessments of individuals is paramount. Applications include, but are not limited to, diagnostic evaluations and treatment planning in clinical psychology (Kamphuis et al., 2021), personnel selection and performance evaluations in organizational psychology (Sackett et al., 2022), admissions processes in educational settings (Woo et al., 2023), and evaluations of risk or criminal responsibility in forensic psychology (Rogers et al., 2023; Singh et al., 2011). Recent developments in the criterion model have incorporated the social consequences of test use into the validity definition, thereby recognizing the ethical implications of psychological assessment (Kane, 2013). The significance of these considerations becomes evident, for example, in jury selection processes using verbal ability tests (e.g., Cronbach, 1988) or in hiring decisions based on mental health evaluations.

In the early days of the criterion model, validity testing was straightforward: it involved demonstrating that a test score predicted a criterion variable by statistical association, thus establishing the *criterion validity* of the measure for its designated purpose (e.g., Cronbach, 1988; Cureton, 1951). Terms such as concurrent validity or predictive validity specify whether the criterion variable is measured at the same time or at a future time. In addition, one can assess whether a measure is more effective or provides unique utility over and above other measures in predicting the same criterion, which is sometimes referred to as incremental validity (e.g., AERA, APA, & NCME, 2014; Cronbach, 1988). More recently, however, with newfound attention to social consequences, validity testing in criterion models has expanded beyond mere criterion prediction to include a wider range of research designs (e.g., Kane, 2013; Messick, 1989b).

The Construct Validity Model

When researchers seek to understand psychological phenomena, they often develop theories about them. This requires measuring the building blocks of theories in psychology, such as psychological traits, states, or processes, which can be conceptualized independently of any specific method of measurement (e.g., Cronbach & Meehl, 1955). Cronbach & Meehl (1955) recognized that the criterion model was not well-suited for determining whether measures accurately capture the psychological phenomena of interest. This shortcoming arises because, in the research

context, measures must be theoretically understood, not just practically useful, so that they can be used not only to predict but also to explain human experience and behavior (e.g., Cronbach, 1988). To complement the criterion model, Cronbach & Meehl (1955) introduced the concept of construct validity, which focuses on the quality of a measure from an explanatory perspective. In the construct validity model, the things being measured are theoretical constructs, and a measure is considered valid if it accurately reflects the intended theoretical construct (e.g., Loevinger, 1957; McGrath, 2005).

Constructs. Some scholars have advocated for a realist interpretation of theoretical constructs, positing that these aim to represent natural, causally effective entities that, although not directly observable, can be inferred from psychological measurements (e.g., Borsboom et al., 2004; Meehl, 1979; Smith, 2005). Borsboom et al. (2003, 2004) specifically endorsed the reflective measurement model (e.g., Bollen & Lennox, 1991; Edwards & Bagozzi, 2000) as a statistical approach to embody this perspective, which can be realized through factor analysis (e.g., Flora & Flake, 2017; Sellbom & Tellegen, 2019). Proponents of constructs realism argue that scientific methodology should aim to carve nature at its joints (Meehl, 1979), and that researchers often naturally adopt a realist interpretation of constructs (Borsboom et al., 2004). Moreover, many concepts and practices in psychometrics, such as reliability estimation (Nunnally, 1978) and testing the fit of factor models (e.g., McNeish & Wolf, 2023), are generally based on assumptions that are best understood from the viewpoint of construct realism (Edwards & Bagozzi, 2000; Borsboom et al., 2004; Hood, 2013; but see Borsboom, 2023).

However, other scholars have highlighted alternative perspectives on constructs (see, e.g., Fried, 2017; Messick, 1989b; Slaney, 2017), some of which reject the notion that constructs should be viewed as real causal entities. These critics propose more pragmatic views, suggesting that constructs serve as convenient abstractions (Yarkoni, 2020), parsimonious summaries (Markon & Jonas, 2016), or organizing principles (Borsboom, 2023; Sijtsma, 2006). For example, Yarkoni (2020) places the idea of real constructs in the realm of metaphysics, arguing that psychological causes cannot coexist with known physical laws that give rise to a material world, nor can two types of causes, physical and psychological, be meaningfully conceptualized together.

However, cybernetic theory (e.g., DeYoung, 2015; DeYoung & Krueger, 2018) provides an example of how psychological constructs can be conceptualized as causes. In cybernetic theory, personality traits represent parameters within a dynamic, goal-directed, and self-regulating internal system that governs behavior (DeYoung, 2015). Psychopathology, characterized by negative affectivity, antagonism, and disinhibition, results from the chronic failure of

the system to achieve goal satisfaction (DeYoung & Krueger, 2018, 2023). Therefore, to measure these constructs means to identify the system's parameter configuration and state. In support of construct realism, it simplifies the integration of constructs into theoretical models, which is essential for elucidating causality, often considered the primary goal of scientific inquiry (e.g., Lundberg et al., 2021). In contrast, the role of pragmatic constructs within theoretical models may be more ambiguous (but see Borsboom, 2023).

Nomological Network Approach. According to Cronbach & Meehl (1955), construct validity can be assessed using the nomological network approach. This method examines the extent to which empirical findings related to a test score are consistent or inconsistent with theoretical predictions about the targeted construct. Predictions may relate to test content and response processes, the relationships among different indicators of a construct, and how the measure relates to other variables (AERA, APA, & NCME, 2014; Loevinger, 1957). The nomological network is the full set of theoretical propositions postulated for the construct from which predictions can be derived for testing construct validity. If a measure is valid, then empirical patterns should closely match predictions. If the measure is invalid, the empirical pattern would deviate from the predicted pattern, indicating that the test score does not purely reflect the target construct but is contaminated by measurement error (e.g., Smith, 2005).

Cronbach (1989) noted that the nomological network approach is similar to confirmatory testing of hypotheses using the hypothetico-deductive model, where empirical tests are designed to either support or dispute a hypothesis. In construct validation, the *validity hypothesis* is tested, asserting that the measure is valid for measuring the intended construct. Crucially, the nomological network approach can only produce meaningful results for testing the validity hypothesis under two conditions. First, the theoretical assumptions about the nomological network must be clearly articulated so that they can be translated into concrete predictions in terms of specific effect sizes or effect size intervals. This requires a certain level of theory formalization, which underscores the importance of conceptual clarity for all elements of a theory (see, e.g., Bringmann et al., 2022; Flake & Fried, 2020; Fried, 2020). Second, the theoretical predictions must be accurate; otherwise, discrepancies between predicted and actual outcomes may simply reflect flaws in the theory (e.g., Smith, 2005).

Although construct validation is often approached as checking items off a checklist (e.g., Borsboom, 2006; Cronbach, 1989; Maul, 2017), validity tests can be designed in any way that data and theory allow it. Any empirical observation can serve as the basis for a validity test. Here, a key distinction is between the theoretical estimand, which

is the theoretical question the researcher is attempting to address, and the empirical estimand, the statistic or model parameter estimated to answer that question (see, e.g., Lundberg et al., 2021). In any test of construct validity, the theoretical estimand is construct validity, but empirical estimands can vary widely.

Empirical estimands for construct validity need not be limited to correlation coefficients, but may include partial correlations, the shape of a distribution, or any other parameter or statistic. Regardless of the empirical estimand chosen, the basic principle is the assessed convergence between theory and data (Cronbach & Meehl, 1955). Commonly used terms such as convergent validity, discriminant validity, and structural validity—sometimes referred to as validity modifier labels (e.g. Newton & Shaw, 2013)—delineate relatively broad classes of empirical estimands, each with its own rationale. However, these labels are often misinterpreted as indicating different theoretical estimands that can be evaluated independently, when in fact they should be considered together to form a single judgment of construct validity (see, e.g., AERA, APA, & NCME, 2014; Lawshe, 1985; Smith, 2005). For example, an ambiguous use of terminology would be that a test has "excellent convergent validity" but "lacks structural validity." A more precise formulation would say that evidence of convergent and structural validity together provide an inconclusive picture of the test's construct validity. To avoid misunderstandings, some scholars recommend avoiding validity modifier labels (e.g., Newton & Shaw, 2013).

While validity tests can take many shapes and forms, structural equation modeling (e.g., Anderson & Gerbing, 1988; Jöreskog, 1970) generally provides a well-suited and flexible statistical framework for construct validation (Schimmack, 2021). First, it allows for the specification of latent variables consistent with the reflective measurement model, thereby facilitating the control of random measurement error. Second, it allows for the disentangling of sources of systematic measurement error, such as method-specific error, for example, using multimethod designs that involve multiple assessment methods for the same construct (e.g., Campbell & Fiske, 1959; Eid et al., 2022; Schimmack, 2021).

Evidential Value of Validity Tests. Not all validity tests provide equally valuable insights into construct validity (e.g., Cronbach, 1989; Kane, 2001; Smith, 2005). Empirical observations that are consistent with the validity hypothesis for an intended construct interpretation do not conclusively prove it, because alternative construct interpretations may be similarly consistent with the data. Thus, a single validity test is unlikely to be sufficient to firmly establish construct validity, suggesting that construct validation is a lengthy process with no clear endpoint (e.g., Cronbach, 1988; Messick, 1989a, 1995). In a similar vein, validity tests are more informative when they take a falsificationist rather

than a confirmationist approach, aiming to rigorously challenge the intended interpretation (e.g., Fidler et al., 2018; Meehl, 1978; Popper, 1962). Thus, high-quality validity tests are characterized by the consideration of plausible rival hypotheses to distinguish the intended interpretation from feasible alternatives (Campbell, 1957; Cronbach, 1989; Meehl, 1978; Messick, 1975; Smith, 2005).

The Unified Model

Current discussions (e.g., Cizek, 2012; Kane, 2013) characterize the two validity models as one concerned with the validity of *test score interpretations* (i.e., construct model) and the other concerned with the validity of *test score uses* (i.e., criterion model). This bifurcation in the concept of validity has been attributed to divergent priorities in different areas of psychological research (e.g., Borsboom & Wijsen, 2016; Newton & Shaw, 2013; Schimmack, 2021; Zumbo, 2007). Applied research emphasizes the use of tests for individual decision-making and encourages consideration of practical utility, social consequences, and ethical test applications (e.g., Iliescu & Greiff, 2021; Woo et al., 2023). Conversely, basic research focuses on measurements that accurately represent theoretical constructs for research applications (e.g., Schimmack, 2021).

While there is general agreement that both models raise important research questions for psychologists, the debate over which aspects should fall under the umbrella of validity has remained a surprisingly persistent and contentious issue (see, e.g., Anastasi, 1986; Borsboom & Wijsen, 2016; Kane, 2013, 2016; Kane & Bridgeman, 2021; Newton & Baird, 2016; Newton & Shaw, 2013; Popham, 1997; Russell, 2022; Shepard, 1997). Efforts to resolve this debate have included proposals to unify these two concepts into one overarching concept of validity (Guion, 1980; Hubley & Zumbo, 2011; Kane, 2013; Loevinger, 1957; Messick, 1989a), typically positioning construct validity as the unifying principle (e.g., Clark & Watson, 2019; Loevinger, 1957; Messick, 1975). In this unified model, validity is defined as the degree to which intended inferences and actions based on a test score are justified, a definition also adopted by current version of *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014).

Criticisms of the Unified Model. Critics of the unified concept argue that it leads to ambiguous terminology by conflating distinct issues that should be assessed separately, potentially causing misconceptions and miscommunications among psychological researchers (see, e.g., Borsboom et al., 2004; Borsboom & Markus, 2013; Borsboom & Wijsen, 2016; Cizek, 2012; Maul, 2017; Mehrens, 1997; Newton & Shaw, 2013, 2016; Popham, 1997). These critics assert that a test can be useful for decision-making without

necessarily capturing any particular construct, and conversely, accurately capturing a construct does not imply its practical utility. In addition, the unified model's definition of validity blurs the distinction between evidence and truth, suggesting that validity is based on the evidential basis that supports an interpretation rather than the veracity of that interpretation (Borsboom et al., 2004; Borsboom & Markus, 2013). Borsboom and Wijsen (2016) succinctly summarized these criticisms, coining the term "Frankenstein's validity monster" to highlight the problematic amalgam of ontological, ethical, and epistemological issues.

Dissatisfaction with the unified concept of validity has led some to suggest moving away from the use of validity as an all-inclusive term to more precisely defined terms (Newton & Shaw, 2016), such as redesignating criterion validity as the utility of a test (Lissitz & Samuelson, 2007). Other scholars, however, argue for the preservation of the unified concept, in part for political reasons, arguing that reserving the term validity to construct validity may diminish considerations of ethical test use (e.g., Kane, 2013; Linn, 1997). The debate over the scope of the term validity continues, with literature reviews and studies indicating a lack of consensus among experts (Camargo et al., 2018; Newton & Shaw, 2013, 2016). The three articles included in this dissertation focus exclusively on construct interpretations of psychological measures.

II. Construct Validation: The Gap Between Ideal and Practice

In the first chapter, I introduced construct validation as a confirmatory testing framework for validating intended construct interpretations of psychological measurements (Cronbach & Meehl, 1955). However, psychologists typically use the term "construct validation" to refer to any research activity aimed at understanding the construct meaning of test scores, including exploratory studies that lack a well-defined construct or theory related to that construct. Narrative accounts of measurement practices even suggest that the field predominantly uses exploratory research designs (e.g., Alexandrova & Haybron, 2016; Benson, 1998; Borsboom et al., 2004; Cronbach, 1988, 1989; Kane, 2001; Sijtsma, 2012, 2013; Zumbo, 2007; Smith, 2005), often deriving latent factors through exploratory factor analysis and freely exploring the correlation patterns of these factors with other variables (e.g., Borsboom, 2006; Maul, 2017; Schimmack, 2010). However, such activities may be more in the spirit of discovering new constructs rather than validating existing ones. Although the factors identified in exploratory studies are commonly treated as constructs, they may not necessarily achieve the status of a full-fledged construct ready for validation as envisioned by Cronbach & Meehl

(1955). Rather, such exploratory analyses may operate with constructs-in-progress, initially empty shells that gradually gain meaning by acquiring a label, a definition, a description, and an evolving theory, all of which can be open to ongoing revision.

Weak vs. Strong Construct Validation

Cronbach (1988, 1989) thought of exploratory approaches to investigating construct interpretations as a weak form of construct validation, whereas the originally proposed nomological network approach was the strong form. Cronbach's terminology seems to suggest that the two forms are not *qualitatively* different, in that both aim for the same theoretical estimand (i.e., construct validity), but are *quantitatively* different, with the weaker form providing a poorer empirical estimand in terms of less robust evidence. This perspective seems to be echoed by other validity theorists, who view the exploratory approach simply as a less rigorous version of the confirmatory approach (see, e.g., Alexandrova & Haybron, 2016; Cronbach, 1989; Kane, 2001, 2013; Schimmack, 2010; Sijtsma, 2013; Strauss & Smith, 2009). Some scholars have suggested that construct validation has an inherently exploratory flavor because it tests construct validity and theory simultaneously (Cronbach & Meehl, 1955; Smith, 2005). However, some have also expressed skepticism about whether all studies of construct meaning, including purely exploratory ones, should qualify as construct validation, although explicit criteria for what counts as a test of construct validity have not been proposed or discussed (e.g., Cronbach, 1989; Kane, 2001; Smith, 2005; Strauss & Smith, 2009). In the following sections, I present and compare hypothetical extreme examples to further elucidate the weak and strong forms of construct validation. The example of the weak form is drawn from narrative accounts of measurement practices, while the example of the strong form is based on methodological discussions of ideal construct validation (see, e.g., Borsboom, 2006; Cronbach, 1988, 1989; Kane, 2001; Maul, 2017; Schimmack, 2010; Zumbo, 2007; Smith, 2005).

The Strong Form. Ideally, the strong form starts with a thoroughly conceptualized construct, including a precise definition and detailed description that clearly conveys what the construct is and is not, and a theory that addresses the causal mechanisms in which the construct is involved (e.g., Borsboom, 2006; Schimmack, 2010; Strauss & Smith, 2009). This allows predictions to be made that can be pinpointed to a narrow range of plausible effect sizes. These predictions are sufficiently construct-specific to be easily distinguished from those related to other constructs, especially those of plausible alternative interpretations (e.g., Cronbach, 1988, 1989; Messick, 1975). Consequently, va-

lidity tests are carefully selected for their potential to distinguish between competing interpretations (e.g., Smith, 2005; Strauss & Smith, 2009). In accordance with contemporary standards for confirmatory research, the research design and hypotheses are preregistered (e.g., Nosek et al., 2018; Wagenmakers et al., 2012).

To understand why the strong form has significant evidential value for testing construct validity, it is crucial to understand the derivation chain in a hypothesis test (e.g., Meehl, 1990; Scheel et al., 2021). This chain includes the tested hypothesis, positioned at the end, and all auxiliary assumptions, forming the rest of the chain. During a hypothesis test, the chain can either remain intact or break. If it remains intact, then the entire chain has held together, but if it breaks, it can only be determined that a break occurred, not the specific location of the break within the chain (see, e.g., Cronbach & Meehl, 1955; Smith, 2005). For a validity test to be truly informative about the hypothesis being tested, it must be ensured that no other segment of the derivation chain is prone to failure, so that the result of the test—whether it holds or breaks—depends solely on the veracity of the hypothesis being tested, not on the veracity of the auxiliary assumptions (e.g., Meehl, 1978, 1990; Scheel et al., 2021; Weimer, 1979). It seems common sense that a hypothesis test should provide a minimum of evidential value regarding the hypothesis to be even considered a test of that hypothesis. To this end, the auxiliary assumptions need to have a relatively high plausibility *a priori*, or at least, a higher plausibility than the hypothesis being tested.

In strong construct validation, the derivation chain is robust because the theoretical predictions are highly plausible. Thus, a broken chain indicates that the measure is invalid rather than that the nomological network is misspecified. In the strong form, validity tests are truly informative because they can provide evidence *for* and *against* the intended construct interpretation of a test score.

The Weak Form. In contrast, the weak form often initiates the investigative process with a (sometimes arbitrary) set of indicators and a limited grasp of the content they tap. There may be no clear intention as to what construct interpretation is being pursued with these indicators (e.g., Borsboom, 2006; Schimmack, 2010). There may simply be an unstated assumption that the indicators reflect an as yet undetermined number of constructs. The lack of intent regarding which *particular* construct is to be measured severely limits the potential for construct-specific validity tests (e.g., Kane, 2001; Smith, 2005). Consequently, in the weak form, the statistical analyses performed are generic (e.g., exploratory factor analysis, internal consistency estimation, correlation matrix analysis). This makes them broadly applicable to any construct regardless of its unique nomological network. For the same reason, however, they are inherently un-

suitable for adjudicating among plausible alternative interpretations. For example, a factor analysis might suggest unidimensionality of a measure, leading to the conclusion that its indicators assess a common construct (e.g., [Flora & Flake, 2017](#); [Sellbom & Tellegen, 2019](#)). Yet, this alone provides little to no evidence of a particular construct interpretation, since (reflective) constructs are supposed to be (essentially) unidimensional by their very definition. Similarly, it is self-evident that many constructs exhibit positive or negative correlations with other constructs, so that examining a correlation matrix without detailed theoretical guidance is insufficient to claim support for a particular construct interpretation (e.g., [Borsboom, 2006](#)).

In the weak form, researchers may theorize after obtaining empirical data and possibly offer a tentative post hoc interpretation of the construct. However, such a post hoc interpretation would likely lack depth and may not rise to the status of a full-fledged construct as required by the nomological network approach (e.g., [Schimmack, 2010, 2021](#)). Moreover, such post hoc theories are unlikely to provide a reasonable basis for a serious test of construct validity, given that the theoretical underpinnings of the nomological network may still be rather speculative. This problem would be exacerbated if the nomological network was largely constructed based on the same measure for which the construct's meaning is to be inferred ([Schimmack, 2010](#)). As a result, the weak form is unlikely to provide a robust derivation chain for testing construct validity. Even if the data appear to be consistent with a particular interpretation, the evidential value of such intuitions may be severely limited. In essence, the weak form may not be able to provide a sufficient level of evidence for a particular interpretation of the construct, although it may narrow the range of possible construct interpretations somewhat.

The Scope of Construct Validation

Here I have illustrated weak and strong construct validation with two hypothetical extreme examples. While there is certainly a broad spectrum of exploratory and confirmatory approaches, I would argue that a significant number of studies in the measurement literature fall closer to the exploratory side. I offer three distinct but interconnected arguments for why the weak form, especially in its extreme, should not be considered construct validation.

First, construct validation means testing the hypothesis that a test score measures a *particular* construct. This seems impossible without a construct, and remains elusive for a construct-in-progress. Second, claims should be proportionate to the evidence provided. However, the weak form is likely to fall short of providing a minimum level of evidence due to the lack of a highly plausible nomological network and/or the reliance on generic statistical approaches.

Third, in no other area of psychological research would exploration be considered a "weak form of validation". On the contrary, the distinction between exploration and validation is typically considered fundamental (e.g., [Nosek et al., 2018](#); [Wagenmakers et al., 2012](#)). Given these considerations, I suggest that many studies that would be considered construct validation according to current standards and practices may not live up to expectations. Rather, exploratory techniques may be more appropriately termed construct exploration, as they allow the study of the construct meaning of psychological measures without necessarily being able to provide evidence for a particular construct interpretation. Whereas construct validation involves testing construct validity as the theoretical estimand, construct exploration may be better suited to testing a range of different, albeit less ambitious, theoretical estimands.

To elucidate this point, consider the bogus construct validation study by [Maul \(2017\)](#), where the author demonstrated that a set of semantically meaningless items, each referencing the fictional term "gavagai," could pass common validation procedures. Specifically, eight such items were rated on a Likert-type scale by 400 participants, resulting in two extracted factors that showed significant, albeit weak, associations with the personality traits agreeableness and openness. As [Maul \(2017\)](#) notes, it would be a widely accepted conclusion that these analyses serve to demonstrate the construct validity of the gavagai questionnaire.

Maul's study serves as a memorable demonstration of the absurdity with which the term validation can be applied in contemporary psychological measurement. It is obvious that construct validity cannot possibly be a meaningful theoretical estimand in this example because gavagai is a made-up term, it is not a construct. This is not to say that nothing can be learned about the gavagai questionnaire from the analyses. Maul's analyses can be considered appropriate for another theoretical estimand: *constructness*, which I would define as the hypothesis that a set of items reflects a common construct, regardless of what that construct may represent. If the items in the gavagai measure assessed a specific construct rather than no construct at all, this alone would lead to the expectation of a positive manifold between the items and correlations with other variables greater than zero ([Rhemtulla et al., 2017](#)). Conversely, if the items measured absolutely nothing, no correlations would be observed because nothing does not correlate with anything. Therefore, the analytic approach in Maul's study can be viewed as assessing the constructness of the gavagai questionnaire, but not its construct validity.

III. Outlining a Framework for Construct Exploration

Existing guidelines for research on the construct meaning of psychological measurements focus heavily on the nomological network approach, which is confirmatory (see, e.g., Cronbach & Meehl, 1955; Clark & Watson, 2019; Simms, 2008; Ziegler, 2020). However, this confirmatory approach requires a high degree of conceptual clarity and a solid theoretical foundation (Alexandrova & Haybron, 2016; Borsboom et al., 2004; Borsboom, 2006), conditions that are often not met due to the vague and tentative nature of theories in psychological research (e.g., Eronen & Bringmann, 2021; Fried, 2020; Oberauer & Lewandowsky, 2019). Consequently, researchers often turn to exploratory methods, but find themselves without substantial guidance on how to do so (e.g., Cronbach, 1989; Kane, 2001; Maul, 2017).

In this chapter, I outline a potential framework for construct exploration, briefly describing its rationale and scope, and then discussing various theoretical estimands that might be included. I propose that construct exploration can be useful whenever there is no stated intention about which well-defined construct a measure is thought to reflect or when tests of construct validity are otherwise not feasible, for example, in the absence of highly plausible or consensually accepted theoretical propositions. In such scenarios, research designs guided by a construct exploration methodology may be the most efficient allocation of resources. Through a program of construct exploration, researchers can aim to gradually refine their understanding of what construct a measure may reflect, operating with a minimum of theoretical assumptions. The goal of construct exploration is to methodically narrow the range of possible construct interpretations, eventually arriving at a set narrow enough to be subjected to construct validation methodology.

Many existing methodological and statistical approaches lend themselves well to a construct exploration framework, especially those that require minimal theoretical assumptions. In the following section, I discuss what these methods are and how they can be used for construct exploration, given their potential for narrowing the range of possible construct interpretations. In contrast to construct validation, which is concerned with the construct validity of a particular construct interpretation for a fixed set of indicators, construct exploration is concerned with the underlying construct and *all* of its indicators. Theoretical estimands for construct exploration include, but are not limited to, whether items measure a construct at all (constructness),

whether the construct is dimensional or categorical (structural type), which set of indicators exhaustively captures the construct (content exhaustiveness), the construct's location within construct taxonomies (construct location), and whether the construct is distinct from previously established constructs (non-redundancy).

Constructness

Before considering which construct might be reflected in an item pool, it is essential to determine whether the items do in fact capture a common construct. Tests of constructness are useful for this purpose; they aim to identify a construct in a pool of items, thus ruling out the possibility that the items do not reflect any construct at all (e.g., Rhemtulla et al., 2017). From a causal perspective, a set of items possessing constructness means that the causal influence of at least one construct on all items is greater than zero. Importantly, constructness does not require that items reflect only a single construct; they may reflect multiple constructs, either in their shared or unique variance. Nor is homogeneity of indicators (McGrath, 2005; Smith et al., 2009) a requirement for constructness (e.g., Bollen & Lennox, 1991). Nevertheless, demonstrating constructness is certainly facilitated by high-quality indicators that strongly and purely reflect a construct.

Constructness of a set of items can be evidenced through their positive manifold (e.g., Bollen & Lennox, 1991; Edwards & Bagozzi, 2000) and their relatively consistent statistical relationships with external variables (Thielmann & Hilbig, 2019). Researchers can apply various latent variable models, including factor analysis, to assess model fit and factor loading patterns for a range of plausible candidate models (e.g., Fabrigar et al., 1999; Brown, 2015). For heterogeneous indicators of a common construct, bifactor modeling (e.g., Morin et al., 2016) is a useful approach (e.g., Rodriguez et al., 2016). However, caution should be taken when relying on model fit indices alone, because the same covariance matrix can fit different data-generating models that do not all involve common factors (e.g., van Bork et al., 2017; Watts et al., 2023). Therefore, model fit indices are informative but do not lend themselves to definitive conclusions about the veracity of models and should be supplemented by other sources of evidence (e.g., Greiff & Heene, 2017; Stanton et al., 2023).

Examining nomological consistency serves as an important complement, indicating whether patterns of associations between potential indicators of a construct and external variables are relatively consistent (e.g., Thielmann & Hilbig, 2019). In the context of testing constructness, the precise choice of external variables and the expected magnitude of the associations are less important, if the range of

external variables is sufficiently broad to detect inconsistencies. The analyses presented in Müller et al. (2022) can be reframed as an examination of constructness. The authors found that although the items of the *Reflective Functioning Questionnaire* (Fonagy et al., 2016) fit a unidimensional model, their associations with external variables showed substantial variation, leaving some doubt as to whether all eight items reflect the same construct.

It should be noted that even if items appear unidimensional and exhibit similar patterns with external variables, this does not guarantee that their factor can be equated with a single construct (e.g., Markon & Jonas, 2016; Wood et al., 2015; Zumbo, 2007). In fact, multiple constructs may produce a single factor if the constructs have similar effects on the indicators, so their different influences cannot be easily disentangled (Savalei & Falk, 2014). A noteworthy example for such phenomena is common method bias (e.g., Podsakoff et al., 2024). Nevertheless, it is not necessary to equate the factor with a construct to build an argument for constructness.

Structural Type

A second theoretical estimand is the structural type of the construct, namely, whether its latent values represent quantitative or qualitative differences (e.g., Hopwood et al., 2023; Markon & Krueger, 2006; Meehl, 1992). In other words, structural type addresses whether the construct is more accurately represented as a continuous dimension or as discrete categories (e.g., Haslam et al., 2020). By determining the structural type of the construct, interpretations that assume a different structural type can be ruled out, thereby refining the range of potential construct interpretations.

To examine the structural type of a construct, one approach is to use taxometric methods (e.g., McGrath & Walters, 2012; Ruscio et al., 2011), while another is to compare the fit of latent variable models, which may include factor models, latent class models, or hybrid models (e.g., Aslinger et al., 2018; Hallquist & Wright, 2014; Wendt et al., 2019). It is critical to recognize that categorical constructs may appear dimensional if their error terms are dimensional, and conversely, dimensional constructs may appear categorical without systematic comparison to dimensional alternatives (Bauer & Curran, 2003; Lubke & Neale, 2006; Wendt et al., 2019). Thus, one should not fully rely on fit indices for establishing the structural type of a construct.

A complementary way to assess structural type is to estimate which model's latent variables explain more variance in external variables (e.g., Markon et al., 2011; Wendt et al., 2019). This strategy is based on the idea that the model that more accurately captures the construct's true structure will

more faithfully reproduce its patterns of association with external variables, whereas a model that less accurately represents the construct will exhibit diminished associations (e.g., Markon et al., 2011; Preacher et al., 2005).

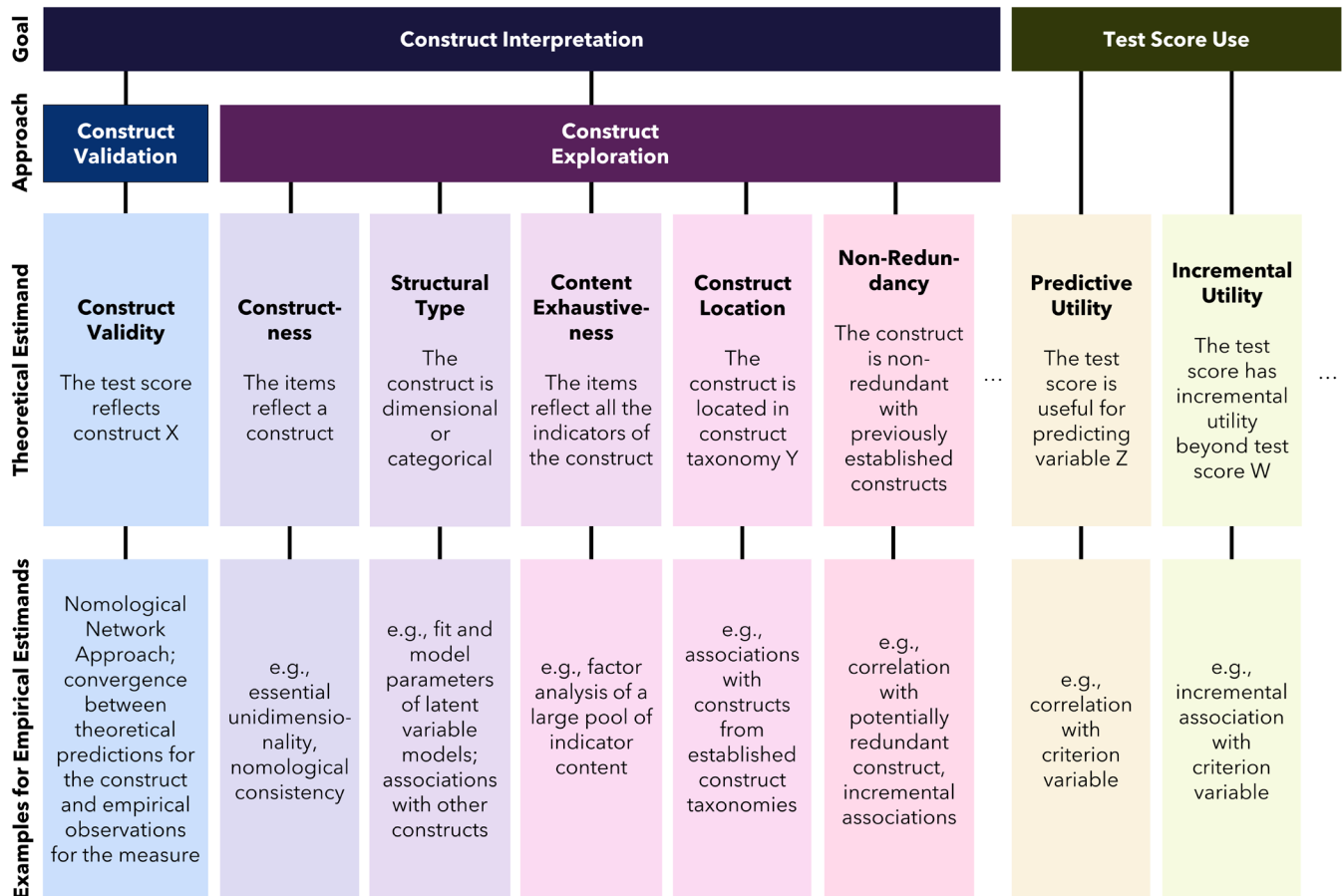
Content Exhaustiveness

While constructness is concerned with identifying a construct within a set of items, content exhaustiveness aims to uncover *all* indicators of that construct. From a causal perspective, a set of indicators is content exhaustive if it contains all indicators for which the construct has a causal influence greater than zero. Thus, content exhaustiveness attempts to delineate the boundaries of the constructs, allowing for a detailed examination of the content of its indicators. Such attempts are essential to clarify, for example, whether the construct is broad or specific, which in turn helps to limit the range of possible interpretations.

Latent variable modeling methods, such as factor analysis (e.g., Flora & Flake, 2017; Sellbom & Tellegen, 2019), again may serve as the primary tool for assessing content exhaustiveness. Here, however, the inclusion of a large and diverse set of content is instrumental. For example, Müller et al. (2023) created 40 self-report items to explore the domain of perceived ability to understand mental states of human beings, systematically sampling varied content regarding the target (self or others) and the type of mental state (emotions, thoughts and attitudes, goals, and motives). Their factor analysis showed that factors emerged for different targets, but not for different types of mental states. This finding contrasts with other studies that have sampled items from the same general domain but limited the consideration of content to emotional mental states, resulting in a distorted picture of the constructs' nature (e.g., Vachon & Lynam, 2016).

It is important to distinguish content exhaustiveness from construct underrepresentation, the latter of which refers to a test score that lacks the content necessary to validly measure a target construct (e.g., Messick, 1989a; Steger et al., 2023). Construct underrepresentation implies that the omission of content can alter the meaning of the construct when indicators are not all mutually interchangeable (e.g., AERA, APA, & NCME, 2014). Content exhaustiveness, on the other hand, aims to identify all indicators of the construct, without considering interchangeability. The purpose of examining content exhaustiveness is to gain insight into the meaning of the construct, not to ensure valid measurement.

Figure 1. Situating Construct Exploration in Psychological Measurement Research



Construct Location

The importance of locating constructs within established taxonomies is increasingly recognized as an effective way to better understand their meaning (e.g., [Bainbridge et al., 2022](#); [Kotov et al., 2021](#)). Moreover, integrating constructs into preexisting frameworks facilitates the synthesis of findings across disciplines, thereby promoting a cumulative approach to science and stimulating theoretical advances ([Baliatti et al., 2015](#); [John et al., 2008](#); [Le et al., 2010](#)). The taxonomies considered may vary depending on the research area. For example, in individual differences research, notable frameworks include the Big Five framework ([Goldberg, 1993](#)), the Hierarchical Taxonomy of Psychopathology (HiTOP; [Kotov et al., 2021](#)), the Interpersonal Circumplex (e.g., [Wright et al., 2023](#)), and the Social, Emotional, and Behavioral Skills Model (e.g., [Soto et al., 2021, 2022](#)). The utility of examining construct location lies in the ability to extrapolate certain information from the taxonomy

in which a construct is located to refine the range of plausible construct interpretations. This is because constructs within the same taxonomy may share some characteristics.

Taxonomies are sometimes hierarchically organized, with constructs placed at different levels of abstraction and specificity (e.g., [Clark & Watson, 2019](#); [Forbes et al., 2021](#); [Lahey et al., 2021](#); [Markon et al., 2005](#)). A key aspect of hierarchical structures is that higher-level components may reflect overarching characteristics shared by (measures of) constructs at the lower levels. This circumstance facilitates the incorporation of prior knowledge about potential common sources of variance that may also pertain to (measures of) the target construct. For example, the placement of a construct within HiTOP may suggest that some of its variance is related to broad factors such as the p-factor (e.g., [Smith et al., 2020](#); [Watts et al., 2023](#)) or psychopathology spectra (e.g., [Kotov et al., 2020](#); [Krueger et al., 2021](#); [Watson et al., 2022](#)), which could point to common substantive or non-substantive influences (e.g., [Bäckström et al., 2009](#); [Watts et al., 2021, 2023](#)).

The location of a target construct within a construct taxonomy can be examined through patterns of association, for example, using correlation or regression analysis, factor analysis (e.g., [Bainbridge et al., 2022](#)), or the back-backward method (e.g., [Forbes, 2023](#); [Goldberg, 2006](#)). Strong associations suggest that the target construct may be appropriately located in proximity, while the absence of associations suggests that the target construct may be better located in a different taxonomy.

Non-Redundancy

The measurement field has witnessed an influx of seemingly novel constructs and measures without sufficient examination of their distinctiveness from established ones (e.g., [Elson et al., 2023](#); [Lawson & Robins, 2021](#); [Le et al., 2010](#)). It is important to recognize that the label given to a construct may not convey its true meaning ([Lilienfeld & Strother, 2020](#); [Weidman et al., 2017](#)). Addressing the question of whether a construct is redundant with a previously established construct is essential to leveraging existing knowledge and consolidating scientific evidence (e.g., [Block, 1995](#); [Elson et al., 2023](#); [Kelley, 1927](#)).

To demonstrate non-redundancy, researchers may show that a new construct does not correlate too strongly with an existing construct, in other words, to ensure that they "rank persons differently, or in some other way give distinctive reports" ([Cronbach, 1989, p. 153](#)). Nonetheless, there are certain caveats: redundancy may be masked by the attenuating effect of random measurement error at the level of observed variables ([Campbell & Fiske, 1959](#); [DeShon, 1998](#)). At the level of latent variables, it may be masked by subtle methodological differences between measures, such as rating scale or item wording (e.g., [Le et al., 2010](#); [Schmidt & Hunter, 1999](#); [Shaffer et al., 2016](#)). In a second approach, redundancy can be indicated by identical nomological patterns with external variables (e.g., [Altgassen et al., 2024](#); [Thielmann & Hilbig, 2019](#)). It is worth noting that measures may appear empirically redundant but differ in content (e.g., [Rosenbusch et al., 2020](#)). In such situations, it may be concluded that the measures are in fact redundant, even though neither exhaustively covers the construct's content.

[Figure 1](#) provides a schematic overview that positions construct exploration alongside established methodologies in research on psychological measurements. The figure includes the theoretical estimands that were discussed in this chapter. However, these are not intended to provide an exhaustive list. It merely underscores the wide array of useful estimands and highlights some that may generally be worth considering. This list may be further extended to examining a construct's consistency or context dependency across samples, methods (e.g., [Levin-Aspensson et al., 2021](#);

[Cote & Buckley, 1987](#)), and populations (e.g., [Meredith, 1993](#)), its substantive significance (e.g., [Greenwald et al., 2009](#); [Kurdi et al., 2019](#); [Soto, 2019](#)), the distribution of its latent values (e.g., [Hester et al., 2023](#)), and its trait- or state-likeness (e.g., [Steyer et al., 2015](#); [Zimmermann et al., 2017](#)).

IV. Summary of the Articles

In previous chapters, I have argued in favor of a clear distinction between two approaches to investigating construct interpretations of psychological measures: construct validation and construct exploration. Against this background, this chapter presents and summarizes the contributions of the three articles included in this dissertation. I suggest that the first and the second article are best viewed through the lens of a construct exploration methodology, while the third article has all the features that one would expect from a construct validation study (e.g., [Cronbach, 1989](#); [Cronbach & Meehl, 1955](#); [Smith, 2005](#)).

Indicators of Affect Dynamics: Structure, Reliability, and Personality Correlates

Research Background. The first article ([Wendt et al., 2020](#)) explores the measurement of individual differences in affective dynamics, the ebb and flow of emotions over time (e.g., [Vaugh & Kuppens, 2021](#)). To this end, researchers often use indicators of affect dynamics (IADs), which are person-specific summary statistics derived from time-series data collected through intensive longitudinal research designs (e.g., [Pirla et al., 2023](#); [Wright & Zimmermann, 2019](#)). It is commonly hypothesized or assumed that different IADs capture distinct (trait-like) constructs that warrant unique interpretations.

Here are examples of constructs, along with the IADs that are sometimes used to measure them. Trait affect, conceptualized as the "set point" or "home base" of the affective system ([Kuppens et al., 2010](#)), is often assessed by the mean of an individual's emotion time series. Emotional variability, conceptualized as general sensitivity to internal or external stimuli, can be measured by the individual standard deviation (e.g., [Kalokerinos et al., 2020](#); [Mader et al., 2023](#)). Emotional inertia, which is the resistance to change in emotional states, can be represented by the individual autocorrelation (e.g., [Koval et al., 2021](#); [Koval & Kuppens, in press](#)). In addition, emotion differentiation, the ability to discriminate between emotions, can be measured by individuals' average intercorrelation between different emotion items ([Thompson et al., 2021](#)).

Previous research has identified mathematical dependencies among IADs, leading to skepticism about their potential to delineate distinct constructs (e.g., [Jahng et al.,](#)

2008; Mestdagh et al., 2018). While the most basic IADs, namely mean and standard deviation, have been established as distinct constructs (e.g., Eid & Diener, 1999), the distinctiveness of more complex IADs remains less certain (e.g., Bos et al., 2019; Dejonckheere et al., 2019; Houben & Kuppens, 2020). Therefore, the aim of this study was to examine whether IADs, especially the more complex ones, can serve as reliable indicators of non-redundant, trait-like individual differences.

Methods and Results. Diary data from Denissen & Kühnel (2008) and Wright et al. (2015), with one measurement occasion per day, and ecological momentary assessment data from Wright et al. (2017), with multiple measurements per day, were used for secondary analyses, totaling 1192 participants and 51,278 measurement occasions. The investigation included IADs ranging from those that can be calculated using simple formulas—such as mean, standard deviation, mean square successive differences, and average item intercorrelation (e.g., Ebner-Priemer et al., 2009; Pirla et al., 2023)—to those that require the estimation of individual vector autoregressive network models, providing individual estimates for contemporaneous correlations, autoregressions, and cross-lagged regressions (Epskamp et al., 2018). The study considered IADs across positive affect, negative affect, and hostile affect scales, each represented by a set of emotion items identified through multilevel factor analysis (Reise et al., 2005).

To assess redundancies among IADs, the study used varimax rotated principal component analysis (see, e.g., Fabrigar et al., 1999), which revealed substantial commonalities among IADs. The study also used regression analysis (Cohen et al., 2004), which showed that simple IADs, namely mean and standard deviation, accounted for a substantial amount of variance in many of the more complex IADs. The split-half reliability of IADs was assessed by dividing the time series into odd and even days (e.g., Mejía et al., 2014). For more complex IADs, the reliability analysis focused on residualized versions in which the mean and standard deviation were partialled out. The reliability of more complex IADs varied, with some reaching acceptable levels and others showing little or no reliable variance. To explore whether IADs might indicate trait-like individual differences, the study correlated IADs with measures of broad personality traits from the Big Five framework (e.g., John et al., 2008) and synthesized results from all three samples in a mini random-effects meta-analysis (e.g., Borenstein et al., 2021), again focusing on residualized versions for the more complex IADs. There were substantial associations for simple IADs, but they were largely absent for complex IADs.

Discussion. The study is best described as a construct exploration study, as it investigates IADs' constructness (whether any constructs are captured), non-redundancy (whether constructs are distinct from each other), and trait-

likeness (whether the constructs represent psychological traits). In contrast, this study should not be thought of as a construct validation study because it does not perform confirmatory tests of the hypothesis that IADs provide valid measurements of particular constructs.

This article contributes to a growing body of literature highlighting the measurement challenges in assessing affect dynamics, including statistical dependencies (e.g., Dejonckheere et al., 2019; Jahng et al., 2008; Mestdagh et al., 2018), low reliability (e.g., Du & Wang, 2018; Pirla et al., 2023; Schneider & Junghaenel, 2022; Wenzel & Brose, 2023), and unclear substantive significance (e.g., Bos et al., 2019; Dejonckheere et al., 2019; Houben & Kuppens, 2020; Koval et al., 2013). The findings caution against prematurely ascribing construct interpretations to IADs as distinct trait-like individual differences without thoroughly considering aforementioned issues.

Since the publication of the 2020 article, researchers have proposed new strategies to capture constructs more accurately in the domain of affect dynamics. These include proposals for alternative statistics (Ringwald & Wright, 2022) or modeling techniques such as dynamic structural equation modeling (e.g., Hamaker et al., 2021; Wenzel et al., 2023; Wenzel & Brose, 2023) and other advanced models (Koslowski & Holtmann, 2023; Mader et al., 2023). However, the use of increasingly complex models or statistical adjustments does not guarantee better measurements (Baral & Curran, 2023; Hoyle et al., 2023; Ringwald & Wright, 2022; Wenzel & Kubiak, 2020). Further research is needed to explore which constructs are captured by emotional time series and which models and statistics are best suited to teasing them apart.

Mapping Established Psychopathology Scales Onto the HiTOP

Research Background. The second article (Wendt et al., 2023) examines the placement of scales from established psychopathology scales within the Hierarchical Taxonomy of Psychopathology (HiTOP). Such scales may, for example, include measures of depression (e.g., Fried et al., 2022), impulsivity (e.g., Hook et al., 2021), emotion regulation (e.g., Agako et al., 2022), paranoia (e.g., Statham et al., 2019), or schizotypy (e.g., Mason, 2015). HiTOP serves as a comprehensive framework that organizes signs, symptoms, and traits associated with mental disorders in a hierarchical taxonomy, from broad, general constructs to more specific ones, delineating different levels of specificity and abstraction based on phenotypic similarity (e.g., Kotov et al., 2021).

Established psychopathology scales are typically aligned with traditional diagnostic concepts, which portray psychopathology as a collection of relatively independent

categorical entities that are either present or absent in an individual. However, contemporary research has firmly challenged this view, pointing to the existence of broad constructs that transcend categorical diagnoses (e.g., Forbes et al., 2021, 2023; Wright et al., 2013; Ringwald et al., 2023). These include the HiTOP spectra of internalizing, thought disorder, detachment, disinhibited externalizing, antagonistic externalizing, and somatoform, as well as a general factor of psychopathology. While the specific meaning of these constructs-in-progress are still being clarified, emerging evidence suggests their potential, highlighting the need for ongoing research in these areas (e.g., Conway et al., 2019; Kotov et al., 2020; Krueger et al., 2021; Smith et al., 2020; Watson et al., 2022).

Historically, established psychopathology scales have been studied in isolation, like research on categorical concepts of mental disorders has been treated (Tabb, 2019). However, with advances in the nosology of psychopathology, there is a need to reassess traditional self-report questionnaires developed in the pre-HiTOP era. Specifically, such reevaluation aims to determine whether these measures also tap into the broader constructs described by the HiTOP taxonomy (e.g., Brown et al., 2023; Sellbom et al., 2021; Wright & Simms, 2015). If so, this would suggest that research findings previously attributed to specific scales and constructs may in fact reflect overarching features of psychopathology, enabling new insights for refining construct interpretations of these scales. Consequently, the second article set out to locate 92 established psychopathology scales within the HiTOP framework.

Methods and Results. Cross-sectional data were used from 909 participants, including 260 healthy community members and 649 outpatients with predominantly severe mental illness. Participants completed a battery of established psychopathology scales with a total of 685 items and were interviewed using the *Structured Clinical Interview for DSM-IV Axis 2 Disorders* (SCID-II; First & Gibbon, 2004).

To map the established scales onto the top two levels of HiTOP, specifically the spectra level and the p-factor, it was necessary to establish a measurement model. This process involved a three-step approach. First, expert ratings of item content were solicited to identify potential indicators for the HiTOP spectra (e.g., Colquitt et al., 2019). Second, the indicator selection for each spectrum was refined using exploratory bifactor analysis (Mansolf & Reise, 2016), thereby ensuring that each conformed to an essentially unidimensional structure (Rodriguez et al., 2016). Third, the final content-based scales were combined into a bifactor-(S-1) model (Eid et al., 2017) using homogeneous parcels (Little et al., 2013). To test the accuracy of the content-based scales in representing the HiTOP spectra, their correlations with SCID-II diagnoses were compared with meta-analytic

results (Ringwald et al., 2023), showing satisfactory consistency between the content-based scales and the meta-analytic results.

To estimate the construct location of the established scales, their factors were regressed on the HiTOP factors from the bifactor-(S-1) model, after removing overlapping items to avoid inflated associations. The HiTOP factors accounted for a substantial proportion of the variance in the established scales' latent factors, with the p-factor explaining an average of 54% and the spectra an additional 14%, leaving 32% of the factor variance unexplained. Most scales were clearly assigned to a single HiTOP spectrum (54 scales), while a smaller number were interstitial, reflecting blends between spectra (23 scales). In addition, there were 12 scales that were purely indicative of the p-factor, with no link to a specific spectrum.

Discussion. This study is best described as a construct exploration study because it locates existing psychopathology scales within the comprehensive HiTOP taxonomy, providing greater opportunities for a more complete understanding of their construct meaning. Specifically, by demonstrating that established scales map coherently onto the hierarchy, the study underscores the need for distinguishing their unique meaning from what shared characteristics may be captured (e.g., Kotov et al., 2020; Krueger et al., 2021; Smith et al., 2020; Watson et al., 2022). This highlights the benefits of a comprehensive psychopathology assessment, which can be used to tease apart these different levels of the hierarchy (e.g., Conway et al., 2019; Simms et al., 2022; Stanton et al., 2020; Vize & Wright, 2024). In addition, the study promotes the integration of different measures of psychopathology and, more broadly, their underlying schools of thought, exploiting their potential for further theoretical advances (DeYoung & Krueger, 2018; DeYoung et al., 2022; Fonagy & Campbell, 2021; Fonagy et al., 2021; Remmers et al., 2023).

Mindreading Measures Misread? A Multimethod Investigation Into the Validity of Self-Report and Task-Based Approaches

Research Background. The third article (Wendt et al., 2024) investigates the construct validity of self-report and task-based measures of mindreading ability by testing their widely accepted interpretation against critical alternative interpretations using the nomological network approach (e.g., Cronbach & Meehl, 1955; Smith, 2005). Notable examples of mindreading ability questionnaires include the *Reflective Functioning Questionnaire* (Fonagy et al., 2016), the *Empathy Quotient* (Baron-Cohen & Wheelwright, 2004), and the *Affective and Cognitive Measure of Empathy* (Vachon & Lynam, 2016). Among task-based measures, the

Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001) and the *Movie for the Assessment of Social Cognition* (Dziobek et al., 2006) are prominent examples. In these tasks, the mental states of fictional characters are inferred from media such as videos, photographs, or written stories.

All these measures, including both self-reports and tasks, were originally developed with the intention of measuring mindreading ability, also known as cognitive empathy or mentalizing ability (e.g., Olderbak & Wilhelm, 2020). Mindreading ability involves understanding and accurately interpreting the mental states of others, such as thoughts, feelings, or motivations (e.g., Bateman & Fonagy, 2019). The research community has largely accepted this construct interpretation, as evidenced by recent meta-analyses using these measures to examine the genetic and environmental bases of mindreading ability (Abramson et al., 2020) and its associations with other constructs such as emotion regulation (Salazar Kämpf et al., 2023), attachment (Kivity et al., 2024), and mental disorders (Bora, 2021; Johnson et al., 2022).

Nevertheless, the construct validity of these measures has sometimes been questioned, with criticism directed at both self-report (Ickes, 1993; Müller et al., 2023; Murphy & Lilienfeld, 2019; Realo et al., 2003) and task-based methods (Dodell-Feder et al., 2013; Kittel et al., 2022; Oakley et al., 2016; Osborne-Crowley, 2020; Quesque & Rossetti, 2020). Critics have suggested alternative interpretations: self-reports may reflect perceived rather than actual mindreading ability, with high scores possibly reflecting a maladaptive form of mindreading-specific overconfidence (Müller et al., 2023). Task-based measures, on the other hand, may not target the specific skills required for mindreading in real-world scenarios; rather, differences in task performance may be due to broader cognitive abilities (Osborne-Crowley, 2020).

Methods and Results. Cross-sectional data were collected online from 700 participants, broadly representing the general U.S. population, in two sessions held a few days apart. The study included eight self-report measures (e.g., de Lima & de Lima Osório, 2021; Luyten et al., 2019) and four tasks (e.g., Quesque & Rossetti, 2020; Yeung et al., 2024) designed to measure mindreading ability. To test construct validity, additional measures were administered, including tasks assessing general cognitive ability (e.g., Condon & Revelle, 2014), ratings of perceived task performance, self-reports of various behavioral, emotional, and social skills (Soto et al., 2022), and indicators of psychosocial functioning (e.g., Ro & Clark, 2009).

Four validity tests were used to assess the plausibility of validity hypotheses, advocating for the construct validity of mindreading ability measures, against rival hypotheses, advocating for alternative construct interpretations. Structural equation modeling (Anderson & Gerbing, 1988; Jöreskog,

1970) served as the estimation strategy (Lundberg et al., 2021), and the study was fully preregistered. For the first validity test, the validity hypotheses propose that self-reports and mindreading tasks reflect the same construct, predicting a medium-to-large correlation between them (e.g., Campbell & Fiske, 1959). However, in stark contrast to this expectation, the latent correlation between self-reports and tasks was negligible and statistically insignificant, $r = .05$, 95% CI [-0.04, .15], lending more support to rival interpretations suggesting that different constructs are being captured.

The second validity test was based on the proposition that mindreading ability is a distinct ability separate from general cognitive ability, suggesting that its association with tasks assessing general cognitive ability should not be too strong (e.g., Cronbach, 1989). However, the latent correlation was very strong, $r = .85$, 95% CI [.76, .94], more consistent with the rival interpretation that mindreading tasks reflect general cognitive ability, although not entirely so, as this would imply a (near) perfect association. For the third validity test, the validity hypotheses predict that self-reported mindreading ability would show a stronger association with actual than with perceived performance on mindreading tasks. Contrary to this prediction, and more consistent with the rival hypotheses, the latent correlation with perceived performance was substantial, $r = .29$, 95% CI [.20, .38], and significantly greater than the correlation with actual performance, $p < .001$, a result more consistent with the rival hypotheses.

The fourth validity test was based on the proposition that greater mindreading ability acts as a causal antecedent of better psychosocial functioning, contributing to more favorable life outcomes such as robust mental health (e.g., Luyten et al., 2020). According to the validity hypotheses, measures of mindreading ability are expected to be associated with indicators of psychosocial functioning, even after accounting for potential confounders of this statistical relationship including general cognitive ability (e.g., Pettersson et al., 2021) and general positive self-evaluation (e.g., Bono & Judge, 2003). Contrary to this prediction, negative partial regression coefficients emerged for both mindreading self-reports, $\beta = -0.21$, 95% CI [-0.29, -0.12], and tasks, $\beta = -0.65$, 95% CI [-1.28, -0.02], a result more in line with rival hypotheses.

Discussion. This study can be considered a construct validation study because it meets the description of strong construct validation outlined in the second chapter. It conducts tests of construct validity for particular construct interpretations, also considering plausible rival hypotheses (e.g., Cronbach, 1989; Meehl, 1978; Messick, 1975), and anchoring these tests in consensually established theoretical propositions (e.g., Cronbach & Meehl, 1955; Smith, 2005; Strauss & Smith, 2009). Four preregistered validity

tests failed to support the validity hypotheses that self-reports and tasks are valid measures of mindreading ability. Rather, results were more consistent with alternative interpretations, suggesting that mindreading self-reports may reflect the self-concept of mindreading rather than actual ability, and that mindreading tasks predominantly reflect general cognitive ability rather than specific mindreading skills.

The validation failure demonstrated in this study calls for caution in using existing mindreading self-reports and tasks to study the construct of mindreading ability. Specifically, the study suggests that these measures may contain substantial systematic measurement error and perhaps even little variance related to the intended construct. Thus, the use of these measures in empirical research studies may lead to biased results and potentially inaccurate or misleading conclusions (e.g., Bagozzi et al., 1991; Messick, 1989a; Podsakoff et al., 2024; Schmidt & Hunter, 1996, 1999; Sijtsma, 2013). Decades of prior research has used measures of mindreading ability under the assumption that they measure mindreading ability, resulting in a substantial body of literature, including recent meta-analyses that have synthesized this literature (e.g., Abramson et al., 2020; Bora, 2021; Johnson et al., 2022; Salazar Kämpf et al., 2023). Given the potential validity problems, it would therefore be advisable to reevaluate previous studies by considering alternative construct interpretations. These may yield different conclusions about the phenomena under study.

However, the cautionary findings from this study require further corroboration by independent researchers to solidify the newly proposed construct interpretations of these measures. This could include replication of the validity tests conducted here, conducting additional tests, assessing generalizability, and critically re-evaluating the conceptualization and theory of the mindreading ability construct.

V. General Discussion

If the measurement is flawed, the ability to make substantiated claims about the constructs under study can be severely compromised given the uncertainties about what is being measured. Construct validity is therefore essential to maintaining the integrity of psychological science (Flake et al., 2017; Flake & Fried, 2020; Schimmack, 2021; Vazire et al., 2022; Zumbo, 2007). However, measurement practices have often been found lacking, deviating from the rigorous approach outlined by Cronbach & Meehl (1955), thereby failing to ensure that test scores are interpreted in accordance with the constructs they reflect (Borsboom, 2006; Cronbach, 1989; Hayden, 2022; Kane, 2017; Flake & Fried, 2020; Flake et al., 2022; Slaney, 2017; Maul, 2017).

This dissertation presents three studies that critically examine common interpretations of constructs in the areas of affect dynamics (Wendt et al., 2020), psychopathology (Wendt et al., 2023), and mindreading ability (Wendt et al., 2024). The findings of these studies underscore that the true nature of the constructs assessed by such measures may be less well understood than is often believed. These studies may serve as anecdotal examples of why caution is warranted when interpreting the test scores of psychological measurements. Researchers cannot always rely solely on a measure's label to infer its meaning (Lilienfeld & Strother, 2020); similarly, they cannot simply assume that a measure is safe to use based on claims that it has been "validated." Ultimately, claims of construct validity must be firmly grounded in empirical evidence (Flake & Fried, 2020; Schimmack, 2021).

Schimmack (2021) is generally skeptical of existing psychological measures, suggesting that many may lack construct validity. Schimmack contends that if this is the case, empirical results may be difficult to interpret to the point of being completely uninformative. Higgins et al. (2024) reached a similarly strong conclusion regarding the *Reading the Mind in the Eyes Test* (RMET), after reviewing its validity evidence and finding it insufficient. They concluded that the RMET should no longer inform theories of social cognition, be used in clinical diagnosis and practice, or be disseminated to the public.

While these critiques may seem excessive at first glance, psychologists will undoubtedly recall instances in which research practices that were once widely accepted in the field were later found to be wholly inadequate (e.g., Klein et al., 2018; Open Science Collaboration, 2015). Nevertheless, these situations presented opportunities for essential reforms that can revitalize psychological science (e.g., Schiavone & Vazire, 2023; Vazire, 2018; Vazire et al., 2022; Rodgers & Shrout, 2018; Nelson et al., 2018). Even if, in hindsight, concerns prove to be less serious than some researchers currently believe, adopting a skeptical perspective can be beneficial in cultivating an ethos of intellectual humility that recognizes the inherent limitations of psychological research (Hoekstra & Vazire, 2021). Moreover, reevaluating and enhancing methodological practices is always a valuable endeavor for advancing the field. In the concluding section of this paper, I will explore potential avenues for refinement, including those recently discussed in the literature and others I personally consider critical.

Reporting Practices. Recent discussions within the psychological measurement literature have emphasized the need for more rigorous and transparent reporting practices (Barry et al., 2014; Flake & Fried, 2020; Flake et al., 2022; Higgins et al., 2024; Shaw et al., 2020). For instance, it is recommended that validity evidence from prior studies be thoroughly reported to justify the selection of measures

(e.g., Hussey & Hughes, 2020; Barry et al., 2014; Flake & Fried, 2020). Furthermore, there is an argument for the inclusion of new, sample-specific validity evidence in all new studies (Appelbaum et al., 2018). Although transparency and rigor are fundamentally beneficial, these calls conflict with the practical limitations of journal page limits. It would undoubtedly be more efficient for reviews or meta-analyses to synthesize the existing literature on validity evidence. However, such studies are rare, with a few exceptions (see, e.g., Higgins et al., 2024; Yao et al., 2022).

Psychometric Purity. The field of measurement often places considerable emphasis on the psychometric properties of measures. These include, but are not limited to, reliability (Nunnally, 1978), measurement invariance (Meredith, 1993), and adherence to a simple structure (e.g., Marsh et al., 2020). While construct validity research is rare, discussions and studies of psychometric properties are widespread. Furthermore, psychometric properties are typically assessed using stringent criteria (e.g., Chen, 2007; Hu & Bentler, 1999), whereas none are applied to validity tests. All this gives the impression that psychometric purity is given priority over construct validity in psychological research.

Reliability, which assesses whether a test score consistently measures the same thing (e.g., Osburn, 2000), is routinely reported in research studies. When reliability evidence is lacking, editors or reviewers may request it and consider its omission a significant flaw. However, strong evidence of construct validity is often not treated as critical (e.g., Flake et al., 2017). Intense debates about reliability and related issues are rampant. These include debates about whether to use Cronbach's alpha or an alternative internal consistency estimator (e.g., Sijtsma, 2009; McNeish, 2018; Flora, 2020; Raykov & Marcoulides, 2019; Savalei & Reise, 2019), or the choice between weighted and unweighted test scores in statistical analyses (McNeish & Wolf, 2020; McNeish, 2022; Widaman & Revelle, 2023a, 2023b).

Reliability is a basic requirement for construct validity, but it does not, by itself, provide evidence for the construct validity of a measure (e.g., Schimmack, 2021; Zumbo, 2007). A measure may have high reliability yet lack any variance related to the target construct (Schimmack, 2021). In this case, the reliability of an invalid measure—whether the test score contains 90%, 70%, or 50% reliable variance—becomes irrelevant. Reliability thus merely acts as a fire detector for construct validity issues: if internal consistency is close to zero, further consideration of the construct becomes unwarranted. Schimmack (2021) stresses that presenting reliability estimates as evidence of construct validity is dubious, however, it remains a widespread practice in the measurement literature (see, e.g., Flake et al., 2017, 2022; Hussey & Hughes, 2020; Higgins et al., 2024).

Further psychometric approaches are often similarly presented as evidence of construct validity (e.g., Dong & Dumas, 2020; Hussey & Hughes, 2020). These include the use of confirmatory factor analysis (CFA) to assess simple structure (Cizek et al., 2008; Hublely et al., 2014; Slaney, 2017). It tests the extent to which indicators that happen to be included in a test accurately reflect the scales to which they are assigned, i.e., without cross-loadings or correlated errors (e.g., Marsh et al., 2020). Traditional criteria for accepting or rejecting a simple structure are stringent and typically permit a few unmodeled parameters (e.g., Hu & Bentler, 1999; McNeish & Wolf, 2023).

However, a measures' fit to a simple structure tells us little about its construct validity unless simple structure is theoretically predicted. Yet, a perfect simple structure is rarely to be expected anyway because reality is complex, so it remains uncertain whether pure indicators that fully meet this criterion even exist (e.g., Hopwood & Donnellan, 2010; Wetzel & Roberts, 2020). Furthermore, small departures from simple structure have a modest effect on test scores, and this effect diminishes in larger models (Marsh et al., 2004).

Similar considerations apply to the evaluation of fit indices in tests of measurement invariance, which assess how consistently a measure behaves across different contexts or populations (e.g., Meredith, 1993). Clearly, it would be problematic if the covariance structure of a personality questionnaire differed dramatically between two subgroups within a population for which the measure is supposed to be valid, such as between men and women. However, although violations of measurement invariance can range from minor to severe, standard practice lends itself to testing for (near) perfect measurement invariance (e.g., Chen, 2007), disregarding variations across the full spectrum of invariance.

In typical scenarios, the psychometric methods discussed here do not offer a whole lot of evidence about whether a test accurately measures a particular construct. Instead, they typically seem to focus on psychometric purity, which could be described as a luxury concern. The strong emphasis on issues of psychometric purity may be indicative of researchers' aversion to rigorous construct validation. While psychometric issues have clear merit in certain situations, such as test optimization (see, e.g., Jankowsky et al., 2020; Olaru et al., 2019), their importance diminishes when the construct meaning of a measure is not well understood. The field's preoccupation with psychometric precision risks pursuing the wrong ideals, tailoring constructs to models rather than the other way around (Wolf, 2023).

Exploratory vs. Confirmatory Research: Blurred Lines.

In contemporary psychological science, the distinction between exploratory and confirmatory research is considered

crucial (Nosek et al., 2018; Wagenmakers et al., 2012). Nevertheless, the term "construct validation" is often used as a catch-all term for studies of psychological measures, including purely exploratory research where constructs are not fully defined or not defined at all (Borsboom, 2006; Maul, 2017). The pervasive tendency to equate exploratory work with confirmatory efforts can hinder the building of a substantive knowledge base and risks overstating evidence of construct validity. Several issues may have contributed to this issue.

The first issue stems from the problematic terminology used by many prominent validity theorists who view construct exploration as a "weak form" of construct validation (Cronbach, 1988, 1989; Kane, 2001; Smith, 2005). The significance of terminology cannot be overstated: researchers will naturally assume that a study provides evidence of construct validity if it is labeled a construct validation study. A second problem is the lack of criteria defining what constitutes construct validation. In contrast, construct validation is sometimes portrayed as an obscure mix of theory testing and construct validity testing (Cronbach & Meehl, 1955; Kane, 2001; Smith, 2005). It would be more fruitful for researchers to decide whether the goal of their study is to test a theory or construct validity, rather than attempting to do both simultaneously. If the analyses do not support either goal, it may be more appropriate to present the study as exploratory.

Third, preregistration, now a cornerstone of confirmatory testing (e.g., Nosek et al., 2018), has not been universally accepted as critical for construct validation. However, its importance in mitigating confirmation bias and motivated reasoning in test development and application cannot be overstated (e.g., Schimmack, 2021; Westen & Rosenthal, 2005). This is especially important in the measurement context, where the distinction between exploration and confirmation is often fuzzy. Furthermore, the deliberate consideration of plausible rival hypotheses should be made a standard practice in construct validation research, as this significantly enhances the evidential value of construct validity tests and makes them true to their name (e.g., Cronbach, 1989; Meehl, 1978; Messick, 1975).

Fourth, the lack of a dedicated construct exploration framework is another issue. Given that many psychological theories are preliminary and underdeveloped (e.g., Eronen & Bringmann, 2021; Fried, 2020; Oberauer & Lewandowsky, 2019), they are not well suited to ambitious construct validation using the nomological network approach (Borsboom, 2023; Cronbach, 1989; Kane, 2016, 2017; Sijtsma, 2006). Indeed, premature reliance on confirmatory tests may hinder scientific progress (e.g., Scheel et al., 2021). In contrast, exploratory work has been instrumental in shaping our conceptualization of individual difference constructs (e.g., Goldberg, 1993; Kotov et al., 2021) and

have the potential to drive theoretical innovations (e.g., Denissen & Penke, 2008; DeYoung, 2015; Fleenor & Jayawickreme, 2015; Roberts, 2018).

Without a framework dedicated to construct exploration, it may be difficult for researchers to make optimal design decisions when construct validation is not a viable option. Conversely, contextualizing exploratory work in a construct validation framework raises issues of interpretation. Researchers may be misled into thinking that their theoretical estimand is construct validity, when in fact their analyses better serve less ambitious estimands such as constructness, structural type, etc., which are not well conceived from the perspective of traditional construct validation. Construct exploration has a different rationale, purpose, and scope, and makes different assumptions that may better illuminate the possibilities and limitations of exploratory measurement work. It provides a context for evaluating exploratory studies that cannot be clearly grasped from the perspective of confirmatory construct validation methodology.

Conceptual Ambiguities with Construct Validity. In the opening chapter of this dissertation, the construct validity model was thoroughly examined. However, one conceptual ambiguity surrounding construct validity remains to be addressed. Although validity theorists have unequivocally emphasized that validity is a matter of degree (e.g., Cronbach & Meehl, 1955; Kane, 2013; Messick, 1989a; Schimmack, 2021; Westen & Rosenthal, 2003), little attention has been paid to how exactly it is to be quantified.

This issue highlights a fundamental ambiguity within the concept of construct validity: how to factor in random and systematic measurement error. Some definitions of construct validity are ambiguous on this issue (e.g., Loevinger, 1957; Strauss & Smith, 2009), while others explicitly consider construct validity as the proportion of construct variance in test scores (e.g., Borsboom et al., 2004; Cronbach & Meehl, 1955; Schimmack, 2021; Westen & Rosenthal, 2003), implying that the composition of error—whether random or systematic—is less important. However, Zumbo (2007) conceptualizes construct validity separately from reliability, characterizing it as the proportion of construct variance within the reliable variance. Zumbo argues that construct validity should be concerned with "inferential quality" and reliability with "data quality" (see also Zumbo & Rupp, 2004).

The primary argument for making a distinction between random and systematic error in conceptualizing construct validity lies in their fundamentally different risks and implications for empirical research (e.g., Bagozzi et al., 1991; Cole & Preacher, 2014; Podsakoff et al., 2024; Schmidt & Hunter, 1996, 1999). Random error, on one hand, can be estimated relatively easily, largely without the need for de-

tailed theoretical assumptions about the construct. Its presence can attenuate bivariate correlations (e.g., DeShon, 1998), hinder control for confounding variables (Westfall & Yarkoni, 2016), and introduce multiple biases in path analyses (e.g., Cole & Preacher, 2014). Structural equation modeling can largely mitigate these biases (Anderson & Gerbing, 1988; DeShon, 1998). On the other hand, detecting systematic measurement error is much more challenging, as it requires theory-informed modeling (e.g., Podsakoff et al., 2024). To control for potential bias, the constructs that induce systematic error must be theoretically understood and themselves measured with construct validity.

For definitions of construct validity that do not differentiate between random and systematic error, the degree of validity can be quantified as the proportion of construct variance within a test score (Schimmack, 2010). It can be estimated using a multi-method approach that aims to control for method-specific variance by using different measures to triangulate the true construct (e.g., Campbell & Fiske, 1959; Eid et al., 2022). The construct validity coefficient is then calculated based on the factor loading on the latent variable (e.g., Schimmack, 2021). However, this method relies on having multiple independent measures of the construct and assumes that their common factor is a close approximation of the construct.

For construct validity concepts that weigh random and systematic error differently, the *r-alerting-cv* coefficient provides a method for quantifying construct validity. This coefficient is calculated using the correlation between the effect sizes of actual validity correlations and those predicted theoretically (Furr & Heuckeroth, 2019; Westen & Rosenthal, 2003). Although this approach bears strong resemblance to validity assessment within the nomological network approach, it was surprisingly not considered by Cronbach & Meehl (1955). A variation of the *r-alerting-cv* coefficient, recalibrated to the level of latent variables, can be used to align it with Zumbo's (2007) concept.

Conclusion. Research on the construct validity of psychological measures is often criticized for suboptimal reporting practices (Flake et al., 2017; Flake & Fried, 2020) and a general lack of rigor (Borsboom, 2006; Schimmack, 2021). However, a closer look reveals that the problems lie not only in implementation, but also in methodology. The measurement field has been plagued by many challenges, including conceptual intricacies and imprecise terminology (Borsboom & Wijsen, 2016; Cizek, 2012), lack of consensus (Camargo et al., 2018; Newton & Shaw, 2013, 2016), misplaced emphasis on aspects of psychometric purity (Hopwood & Donnellan, 2010; Marsh et al., 2020; Wetzel & Roberts, 2020), a prevailing inclination toward confirmationism over falsificationism (Cronbach, 1989; Meehl, 1978; Messick, 1975), a lack of differentiation between exploratory and confirmatory research (e.g., Cronbach, 1988,

1989; Kane, 2001), and the absence of a construct exploration framework, leaving applied researchers without much-needed tools.

This dissertation attempted to shed more light on potential challenges in psychological measurement. The included articles examined measures of affect dynamics (Wendt et al., 2020), psychopathology (Wendt et al., 2023), and mindreading (Wendt et al., 2024), with results suggesting that our understanding of these measures is incomplete and that significant efforts are needed to move forward. The fundamental question of what exactly measurements of psychological phenomena mean may continue to puzzle researchers for decades to come.

References

- Abramson, L., Uzefovsky, F., Toccaceli, V., & Knafo-Noam, A. (2020). The genetic and environmental origins of emotional and cognitive empathy: review and meta-analyses of twin studies. *Neuroscience & Biobehavioral Reviews*, *114*, 113-133. <https://doi.org/10.1016/j.neubiorev.2020.03.023>
- Agako, A., Ballester, P., Stead, V., McCabe, R. E., & Green, S. M. (2022). Measures of emotion dysregulation: A narrative review. *Canadian Psychology*, *63*(3), 376-391. <https://doi.org/10.1037/cap0000307>
- Alexandrova, A., & Haybron, D. M. (2016). Is construct validation valid? *Philosophy of Science*, *83*(5), 1098-1109. <https://doi.org/10.1086/687941>
- Altgassen, E., Geiger, M., & Wilhelm, O. (2024). Do you mind a closer look? A jingle-jangle fallacy perspective on mindfulness. *European Journal of Personality*, *38*(2), 365-387. <https://doi.org/10.1177/08902070231174575>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME) (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*(1), 1-16. <https://doi.org/10.1146/annurev.ps.37.020186.000245>
- Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411-423. <https://doi.org/10.1037/0033-2909.103.3.411>
- Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3-25. <https://doi.org/10.1037/amp0000191>

- Aslinger, E. N., Manuck, S. B., Pilkonis, P. A., Simms, L. J., & Wright, A. G. C. (2018). Narcissist or narcissistic? Evaluation of the latent structure of narcissistic personality disorder. *Journal of Abnormal Psychology, 127*(5), 496-502. <https://doi.org/10.1037/abn0000363>
- Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality, 43*(3), 335-344. <https://doi.org/10.1016/j.jrp.2008.12.013>
- Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly, 36*(3), 421-458. <https://doi.org/10.2307/2393203>
- Bainbridge, T. F., Ludeke, S. G., & Smillie, L. D. (2022). Evaluating the Big Five as an organizing framework for commonly used psychological trait scales. *Journal of Personality and Social Psychology, 122*(4), 749-777. <https://doi.org/10.1037/pspp0000395>
- Baliotti, S., Mäs, M., & Helbing, D. (2015). On disciplinary fragmentation and scientific progress. *PloS One, 10*(3), e0118747. <https://doi.org/10.1371/journal.pone.0118747>
- Baral, S., & Curran, P. J. (2023). Dynamic structural equation models: Promising yet concerning. *American Journal of Undergraduate Research, 20*(3), 69-79. <https://doi.org/10.33697/ajur.2023.096>
- Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders, 34*(2), 163-175. <https://doi.org/10.1023/b:jadd.0000022607.19833.00>
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the mind in the eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42*(2), 241-251. <https://doi.org/10.1111/1469-7610.00715>
- Barry, A. E., Chaney, B. H., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability: Reporting practices in the field of health education and behavior. *Health Education & Behavior, 41*(1), 12-18. <https://doi.org/10.1177/1090198113483139>
- Bateman, A. W., & Fonagy, P. (2019). *Handbook of Mentalizing in Mental Health Practice*. American Psychiatric Publishing.
- Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8*(3), 338-363. <https://doi.org/10.1037/1082-989X.8.3.338>
- Benson, J. (1998). Developing a strong program of construct validation: a test anxiety example. *Educational Measurement: Issues and Practice, 17*(1), 10-17. <https://doi.org/10.1111/j.1745-3992.1998.tb00616.x>
- Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117*(2), 187-215. <https://doi.org/10.1037/0033-2909.117.2.187>
- Bollen, K. A., & Lennox, R. D. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305-314. <https://doi.org/10.1037/0033-2909.110.2.305>
- Bono, J. E., & Judge, T. A. (2003). Core self-evaluations: a review of the trait and its role in job satisfaction and job performance. *European Journal of Personality, 17*(1_suppl), S5-S18. <https://doi.org/10.1002/per.481>
- Bora, E. (2021). A meta-analysis of theory of mind and 'mentalization' in borderline personality disorder: a true neuro-social-cognitive or meta-social-cognitive impairment? *Psychological Medicine, 51*(15), 2541-2551. <https://doi.org/10.1017/s0033291721003718>
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to Meta-Analysis*. John Wiley & Sons.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3). <https://doi.org/10.1007/s11336-006-1447-6>
- Borsboom, D. (2023). Psychological constructs as organizing principles. In *Essays on Contemporary Psychometrics. Methodology of Educational Measurement and Assessment*. Springer. https://doi.org/10.1007/978-3-031-10370-4_5
- Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement, 50*(1), 110-114. <https://doi.org/10.1111/jedm.12006>
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203-219. <https://doi.org/10.1037/0033-295x.110.2.203>
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071. <https://doi.org/10.1037/0033-295x.111.4.1061>
- Borsboom, D., & Wijsen, L. D. (2016). Frankenstein's validity monster: the value of keeping politics and science separated. *Assessment in Education: Principles, Policy & Practice, 23*(2), 281-283. <https://doi.org/10.1080/0969594x.2016.1141750>
- Bos, E. H., De Jonge, P., & Cox, R. F. A. (2019). Affective variability in depression: Revisiting the inertia-instability paradox. *British Journal of Psychology, 110*(4), 814-827. <https://doi.org/10.1111/bjop.12372>

- Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science*, 31(4), 340-346. <https://doi.org/10.1177/09637214221096485>
- Brown, J. R., Hicks, A. D., Sellbom, M., & McCord, D. M. (2023). Further mapping of the MMPI-3 onto HiTOP in a primary medical care and a college student sample. *Psychological Assessment*, 35(7), 547-558. <https://doi.org/10.1037/pas0001218>
- Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. Guilford
- Camargo, S. L., Herrera, A. N., & Traynor, A. (2018). Looking for a consensus in the discussion about the concept of validity. *Methodology*, 14(4), 146-155. <https://doi.org/10.1027/1614-2241/a000157>
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297-312. <https://doi.org/10.1037/h0040950>
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105. <https://doi.org/10.1037/h0046016>
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), 464-504. <https://doi.org/10.1080/10705510701301834>
- Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, 17(1), 31-43. <https://doi.org/10.1037/a0026975>
- Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity Evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68(3), 397-412. <https://doi.org/10.1177/0013164407310130>
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412-1427. <https://doi.org/10.1037/pas0000626>
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2004). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge.
- Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, 19(2), 300-315. <https://doi.org/10.1037/a0033805>
- Colquitt, J. A., Sabey, T., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology*, 104(10), 1243-1265. <https://doi.org/10.1037/apl0000406>
- Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, 43, 52-64. <https://doi.org/10.1016/j.intell.2014.01.004>
- Conway, C., Forbes, M. K., Forbush, K. T., Fried, E. I., Hallquist, M. N., Kotov, R., Mullins-Sweatt, S. N., Shackman, A. J., Skodol, A. E., South, S. C., Sunderland, M., Waszczuk, M. A., Zald, D. H., Afzali, M. H., Bornovalova, M. A., Carragher, N., Docherty, A. R., Jonas, K., Krueger, R. F., . . . Eaton, N. R. (2019). A hierarchical taxonomy of psychopathology can transform mental health research. *Perspectives on Psychological Science*, 14(3), 419-436. <https://doi.org/10.1177/1745691618810696>
- Cote, J. A., & Buckley, M. R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation Studies. *Journal of Marketing Research*, 24(3), 315. <https://doi.org/10.2307/3151642>
- Cronbach, L. J. (1988). Five perspectives on validity argument. In *Test Validity*. Lawrence Erlbaum.
- Cronbach, L. J. (1989). Construct validation after thirty years. In *Intelligence: Measurement, Theory, and Public Policy*. University of Illinois Press.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302. <https://doi.org/10.1037/h0040957>
- Cureton, E. E. (1951). Validity. In *Educational Measurement*. American Council on Education.
- de Lima, F. F., & de Lima Osório, F. (2021). Empathy: Assessment instruments and psychometric quality - A systematic literature review with a meta-analysis of the past ten years. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.781346>
- Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, 3(5), 478-491. <https://doi.org/10.1038/s41562-019-0555-0>
- Denissen, J. J. A., & Kühnel, A. (2008). Handbook for the use of data from the diary study at Humboldt Universität zu Berlin. Retrieved from https://www.psychologie.hu-berlin.de/de/prof/perdev/downloadentwper/diarystudy/Handbook_Diary.pdf
- Denissen, J. J. A., & Penke, L. (2008). Motivational individual reaction norms underlying the Five-Factor model of personality: First steps towards a theory-based conceptual framework. *Journal of Research in Personality*, 42(5), 1285-1302. <https://doi.org/10.1016/j.jrp.2008.04.002>
- DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods*, 3(4), 412-423. <https://doi.org/10.1037/1082-989X.3.4.412>

- DeYoung, C. G. (2015). Cybernetic Big Five theory. *Journal of Research in Personality, 56*, 33-58. <https://doi.org/10.1016/j.jrp.2014.07.004>
- DeYoung, C. G., Kotov, R., Krueger, R. F., Cicero, D. C., Conway, C., Eaton, N. R., Forbes, M. K., Hallquist, M. N., Jonas, K., Latzman, R. D., Rodriguez-Seijas, C., Ruggero, C. J., Simms, L. J., Waldman, I. D., Waszczuk, M. A., Widiger, T. A., & Wright, A. G. (2022). Answering questions about the Hierarchical Taxonomy of Psychopathology (HITOP): analogies to whales and sharks miss the boat. *Clinical Psychological Science, 10*(2), 279-284. <https://doi.org/10.1177/21677026211049390>
- DeYoung, C. G., & Krueger, R. F. (2018). A cybernetic theory of psychopathology. *Psychological Inquiry, 29*(3), 117-138. <https://doi.org/10.1080/1047840x.2018.1513680>
- DeYoung, C. G., & Krueger, R. F. (2023). A cybernetic perspective on the nature of psychopathology: Transcending conceptions of mental illness as statistical deviance and brain disease. *Journal of Psychopathology and Clinical Science, 132*(3), 228-237. <https://doi.org/10.1037/abn0000541>
- Dodell-Feder, D., Lincoln, S. H., Coulson, J. P., & Hooker, C. I. (2013). Using fiction to assess mental state understanding: A new task for assessing theory of mind in adults. *PLoS One, 8*(11), e81279. <https://doi.org/10.1371/journal.pone.0081279>
- Dong, Y., & Dumas, D. (2020). Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Personality and Individual Differences, 160*, 109956. <https://doi.org/10.1016/j.paid.2020.109956>
- Du, H., & Wang, L. (2018). Reliabilities of intraindividual variability indicators with autocorrelated longitudinal data: Implications for longitudinal study designs. *Multivariate Behavioral Research, 53*(4), 502-520. <https://doi.org/10.1080/00273171.2018.1457939>
- Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., & Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders, 36*(5), 623-636. <https://doi.org/10.1007/s10803-006-0107-0>
- Ebner-Priemer, U. W., Eid, M., Kleindienst, N., Stabenow, S., & Trull, T. J. (2009). Analytic strategies for understanding affective (in)stability and other dynamic processes in psychopathology. *Journal of Abnormal Psychology, 118*(1), 195-202. <https://doi.org/10.1037/a0014868>
- Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155-174. <https://doi.org/10.1037/1082-989X.5.2.155>
- Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology, 76*(4), 662-676. <https://doi.org/10.1037/0022-3514.76.4.662>
- Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods, 22*(3), 541-562. <https://doi.org/10.1037/met0000083>
- Eid, M., Koch, T., & Geiser, C. (2022). Multitrait-multimethod models. *Handbook of Structural Equation Modeling* (2nd ed.). Routledge.
- Elson, M., Hussey, I., Alsalti, T., & Arslan, R. (2023). Psychological measures aren't toothbrushes. *Communications Psychology, 1*(1). <https://doi.org/10.1038/s44271-023-00026-9>
- Epskamp, S., Waldorp, L. J., Möttus, R., & Borsboom, D. (2018). The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research, 53*(4), 453-480. <https://doi.org/10.1080/00273171.2018.1454823>
- Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science, 16*(4), 779-788. <https://doi.org/10.1177/1745691620970586>
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299. <https://doi.org/10.1037/1082-989X.4.3.272>
- Fidler, F., Thorn, F. S., Barnett, A., Kambouris, S., & Kruger, A. (2018). The epistemic importance of establishing the absence of an effect. *Advances in Methods and Practices in Psychological Science, 1*(2), 237-244. <https://doi.org/10.1177/2515245918770407>
- First, M. B., & Gibbon, M. (2004). The Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) and the Structured Clinical Interview for DSM-IV Axis II Disorders (SCID-II). In *Comprehensive Handbook of Psychological Assessment, Vol. 2. Personality Assessment*. John Wiley & Sons, Inc.
- Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *American Psychologist, 77*(4), 576-588. <https://doi.org/10.1037/amp0001006>
- Flake, J. K., & Fried, E. I. (2020). Measurement SchMeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456-465. <https://doi.org/10.1177/2515245920952393>

- 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. <https://doi.org/10.1177/1948550617693063>
- Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, 56, 82-92. <https://doi.org/10.1016/j.jrp.2014.10.009>
- Flora, D. B. (2020). Your coefficient Alpha is probably wrong, but which coefficient Omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, 3(4), 484-501. <https://doi.org/10.1177/2515245920951747>
- Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science*, 49(2), 78-88. <https://doi.org/10.1037/cbs0000069>
- Fonagy, P., & Campbell, C. (2021). Future directions in personality pathology. *Current Opinion in Psychology*, 37, 145-151. <https://doi.org/10.1016/j.copsyc.2021.01.001>
- Fonagy, P., Campbell, C., Constantinou, M. P., Higgitt, A., Allison, E., & Luyten, P. (2021). Culture and psychopathology: An attempt at reconsidering the role of social learning. *Development and Psychopathology*, 34(4), 1205-1220. <https://doi.org/10.1017/s0954579421000092>
- Fonagy, P., Luyten, P., Moulton-Perkins, A., Lee, Y. W., Warren, F., Howard, S., Ghinai, R., Fearon, P., & Lowyck, B. (2016). Development and validation of a self-report measure of mentalizing: the Reflective Functioning Questionnaire. *PloS One*, 11(7), e0158678. <https://doi.org/10.1371/journal.pone.0158678>
- Forbes, M. K. (2023). Improving hierarchical models of individual differences: An extension of Goldberg's bass-ackward method. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000546>
- Forbes, M. K., Neo, B., Nezami, O. M., Fried, E. I., Faure, K., Michelsen, B., Twose, M., & Dras, M. (2023). Elemental psychopathology: distilling constituent symptoms and patterns of repetition in the diagnostic criteria of the DSM-5. *Psychological Medicine*, 1-9. <https://doi.org/10.1017/s0033291723002544>
- Forbes, M. K., Sunderland, M., Rapee, R. M., Batterham, P. J., CEAR, A. L., Carragher, N., Ruggero, C. J., Zimmerman, M., Baillie, A., Lynch, S., Mewton, L., Slade, T., & Krueger, R. F. (2021). A detailed hierarchical model of psychopathology: From individual symptoms up to the general factor of psychopathology. *Clinical Psychological Science*, 9(2), 139-168. <https://doi.org/10.1177/2167702620954799>
- Fried, E. I. (2017). What are psychological constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health Psychology Review*, 11(2), 130-134. <https://doi.org/10.1080/17437199.2017.1306718>
- Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, 31(4), 271-288. <https://doi.org/10.1080/1047840X.2020.1853461>
- Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, 1(6), 358-368. <https://doi.org/10.1038/s44159-022-00050-2>
- Furr, R. M., & Heuckeroth, S. (2019). The "Quantifying Construct Validity" Procedure: Its Role, Value, Interpretations, and Computation. *Assessment*, 26(4), 555-566. <https://doi.org/10.1177/1073191118820638>
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48(1), 26-34. <https://doi.org/10.1037/0003-066x.48.1.26>
- Goldberg, L. R. (2006). Doing it all bass-ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality*, 40(4), 347-358. <https://doi.org/10.1016/j.jrp.2006.01.001>
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1), 17-41. <https://doi.org/10.1037/a0015575>
- Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*, 33(5), 313-317. <https://doi.org/10.1027/1015-5759/a000450>
- Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, 11(3), 385-398. <https://doi.org/10.1037/0735-7028.11.3.385>
- Hallquist, M. N., & Wright, A. G. (2014). Mixture modeling methods for the assessment of normal and abnormal personality, Part I: Cross-sectional models. *Journal of Personality Assessment*, 96(3), 256-268. <https://doi.org/10.1080/00223891.2013.845201>
- Hamaker, E. L., Asparouhov, T., & Muthén, B. (2021). Dynamic structural equation modeling as a combination of time series modeling, multilevel modeling, and structural equation modeling. In *The Handbook of Structural Equation Modeling*, Guilford Press.
- Haslam, N., McGrath, M. J., Viechtbauer, W., & Kuppens, P. (2020). Dimensions over categories: a meta-analysis of taxometric research. *Psychological Medicine*, 50(9), 1418-1432. <https://doi.org/10.1017/s003329172000183x>

- Hayden, E. P. (2022). A call for renewed attention to construct validity and measurement in psychopathology research. *Psychological Medicine*, 52(14), 2930–2936. <https://doi.org/10.1017/s0033291722003221>
- Hester, N., Axt, J., Siemers, N., & Hehman, E. (2023). Evaluating validity properties of 25 race-related scales. *Behavior Research Methods*, 55(4), 1758–1777. <https://doi.org/10.3758/s13428-022-01873-w>
- Higgins, W., Kaplan, D., Deschrijver, E., & Ross, R. M. (2024). Construct validity evidence reporting practices for the Reading the mind in the eyes test: A systematic scoping review. *Clinical Psychology Review*, 102378. <https://doi.org/10.1016/j.cpr.2023.102378>
- Hoekstra, R., & Vazire, S. (2021). Aspiring to greater intellectual humility in science. *Nature Human Behaviour*, 5(12), 1602–1607. <https://doi.org/10.1038/s41562-021-01203-8>
- Hood, S. B. (2013). Psychological measurement and methodological realism. *Erkenntnis*, 78(4), 739–761. <https://doi.org/10.1007/s10670-013-9502-z>
- Hook, R., Grant, J. E., Ioannidis, K., Tiego, J., Yücel, M., Wilkinson, P., & Chamberlain, S. R. (2021). Trans-diagnostic measurement of impulsivity and compulsivity: A review of self-report tools. *Neuroscience & Biobehavioral Reviews*, 120, 455–469. <https://doi.org/10.1016/j.neubiorev.2020.10.007>
- Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, 14(3), 332–346. <https://doi.org/10.1177/1088868310361240>
- Hopwood, C. J., Morey, L. C., & Markon, K. E. (2023). What is a psychopathology dimension? *Clinical Psychology Review*, 106, 102356. <https://doi.org/10.1016/j.cpr.2023.102356>
- Houben, M., & Kuppens, P. (2020). Emotion dynamics and the association with depressive features and borderline personality disorder traits: unique, specific, and prospective relationships. *Clinical Psychological Science*, 8(2), 226–239. <https://doi.org/10.1177/2167702619871962>
- Hoyle, R. H., Lynam, D. R., Miller, J. D., & Pek, J. (2023). The questionable practice of partialing to refine scores on and inferences about measures of psychological constructs. *Annual Review of Clinical Psychology*, 19(1), 155–176. <https://doi.org/10.1146/annurev-clinpsy-071720-015436>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Huble, A. M., Zhu, S. M., Sasaki, A., & Gadermann, A. M. (2014). Synthesis of validation practices in two assessment journals: Psychological Assessment and the European Journal of Psychological Assessment. In *Validity and Validation in Social, Behavioral, and Health Sciences*. Springer International Publishing. https://doi.org/10.1007/978-3-319-07794-9_11
- Huble, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, 103(2), 219–230. <https://doi.org/10.1007/s11205-011-9843-4>
- Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, 3(2), 166–184. <https://doi.org/10.1177/2515245919882903>
- Ickes, W. (1993). Empathic accuracy. *Journal of Personality*, 61(4), 587–610. <https://doi.org/10.1111/j.1467-6494.1993.tb00783.x>
- Iliescu, D., & Greiff, S. (2021). On consequential validity. *European Journal of Psychological Assessment*, 37(3), 163–166. <https://doi.org/10.1027/1015-5759/a000664>
- Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological Methods*, 13(4), 354–375. <https://doi.org/10.1037/a0014173>
- Jankowsky, K., Olaru, G., & Schroeders, U. (2020). Compiling measurement invariant short scales in cross-cultural personality assessment using ant colony optimization. *European Journal of Personality*, 34(3), 470–485. <https://doi.org/10.1002/per.2260>
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. In *Handbook of Personality: Theory and Research* (3rd ed.). Guilford Press.
- Johnson, B., Kivity, Y., Rosenstein, L. K., LeBreton, J. M., & Levy, K. N. (2022). The association between mentalizing and psychopathology: A meta-analysis of the reading the mind in the eyes task across psychiatric disorders. *Clinical Psychology: Science and Practice*, 29(4), 423–439. <https://doi.org/10.1037/cps0000105>
- Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika*, 57(2), 239–251. <https://doi.org/10.1093/biomet/57.2.239>
- Kalokerinos, E. K., Murphy, S. C., Koval, P., Bailen, N. H., Crombez, G., Hollenstein, T., Gleeson, J., Thompson, R. J., Van Ryckeghem, D., Kuppens, P., & Bastian, B. (2020). Neuroticism may not reflect emotional variability. *Proceedings of the National Academy of Sciences of the United States of America*, 117(17), 9270–9276. <https://doi.org/10.1073/pnas.1919934117>

- Kamphuis, J. H., Noordhof, A., & Hopwood, C. J. (2021). When and how assessment matters: An update on the Treatment Utility of Clinical Assessment (TUCA). *Psychological Assessment*, 33(2), 122-132. <https://doi.org/10.1037/pas0000966>
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342. <https://doi.org/10.1111/j.1745-3984.2001.tb01130.x>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. <https://doi.org/10.1111/jedm.12000>
- Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, 23(2), 198-211. <https://doi.org/10.1080/0969594x.2015.1060192>
- Kane, M. T. (2017). Causal interpretations of psychological attributes. *Measurement: Interdisciplinary Research & Perspective*, 15(2), 79-82. <https://doi.org/10.1080/15366367.2017.1369771>
- Kane, M., & Bridgeman, B. (2021). The Evolution of the Concept of Validity. In *The History of Educational Measurement*. Routledge.
- Kelley, T. L. (1927). *Interpretation of Educational Measurements*. Macmillan.
- Kittel, A., Olderbak, S., & Wilhelm, O. (2022). Sty in the Mind's Eye: A Meta-Analytic Investigation of the nomological network and internal consistency of the "Reading the Mind in the Eyes" test. *Assessment*, 29(5), 872-895. <https://doi.org/10.1177/1073191121996469>
- Kivity, Y., Levy, K. N., Johnson, B., Rosenstein, L. K., & LeBreton, J. M. (2024). Mentalizing in and out of awareness: A meta-analytic review of implicit and explicit mentalizing. *Clinical Psychology Review*, 102395. <https://doi.org/10.1016/j.cpr.2024.102395>
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzaska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., ..., Nosek, B. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490. <https://doi.org/10.1177/2515245918810225>
- Koslowski, K., & Holtmann, J. (2023, July 6). Unique contributions of dynamic affect indicators - beyond static variability. <https://doi.org/10.31234/osf.io/t6xqk>
- Kotov, R., Jonas, K., Carpenter, W. T., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, K. T., Hobbs, K. A., Reininghaus, U., Slade, T., South, S. C., Sunderland, M., Waszczuk, M. A., Widiger, T. A., Wright, A. G., Zald, D. H., Krueger, R. F., Watson, D., & Workgroup, H. U. (2020). Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): I. Psychosis superspectrum. *World Psychiatry*, 19(2), 151-172. <https://doi.org/10.1002/wps.20730>
- Kotov, R., Krueger, R. F., Watson, D., Cicero, D. C., Conway, C., DeYoung, C. G., Eaton, N. R., Forbes, M. K., Hallquist, M. N., Latzman, R. D., Mullins-Sweatt, S. N., Ruggero, C. J., Simms, L. J., Waldman, I. D., Waszczuk, M. A., & Wright, A. G. (2021). The Hierarchical Taxonomy of Psychopathology (HiTOP): a quantitative nosology based on consensus of evidence. *Annual Review of Clinical Psychology*, 17(1), 83-108. <https://doi.org/10.1146/annurev-clinpsy-081219-093304>
- Koval, P., Burnett, P. T., & Zheng, Y. (2021). Emotional inertia: on the conservation of emotional momentum. In *Affect Dynamics*. Springer International Publishing.
- Koval, P., & Kuppens, P. (in press). Changing feelings: Individual differences in emotional inertia. In *Change in Emotion and Mental Health*. Academic Press.
- Koval, P., Pe, M. L., Meers, K., & Kuppens, P. (2013). Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion*, 13(6), 1132-1141. <https://doi.org/10.1037/a0033579>
- Krueger, R. F., Hobbs, K. A., Conway, C., Dick, D. M., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, K. T., Keyes, K. M., Latzman, R. D., Michelini, G., Patrick, C. J., Sellbom, M., Slade, T., South, S. C., Sunderland, M., Tackett, J. L., Waldman, I. D., Waszczuk, M. A., . . . Kotov, R. (2021). Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): II. Externalizing superspectrum. *World Psychiatry*, 20(2), 171-193. <https://doi.org/10.1002/wps.20844>
- Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, 21, 984-991. <https://doi.org/10.1177/0956797610372634>
- Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, 74(5), 569-586. <https://doi.org/10.1037/amp0000364>
- Lahey, B. B., Moore, T. M., Kaczurkin, A. N., & Zald, D. H. (2021). Hierarchical models of psychopathology: empirical support, implications, and remaining issues. *World Psychiatry*, 20(1), 57-63. <https://doi.org/10.1002/wps.20824>

- Lawshe, C. H. (1985). Inferences from personnel tests and their validity. *Journal of Applied Psychology*, *70*(1), 237-238. <https://doi.org/10.1037/0021-9010.70.1.237>
- Lawson, K. M., & Robins, R. W. (2021). Sibling constructs: what are they, why do they matter, and how should you handle them? *Personality and Social Psychology Review*, *25*(4), 344-366. <https://doi.org/10.1177/10888683211047101>
- Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes*, *112*(2), 112-125. <https://doi.org/10.1016/j.obhdp.2010.02.003>
- Levin-Aspenson, H. F., Watson, D., Clark, L. A., & Zimmerman, M. (2021). What is the general factor of psychopathology? Consistency of the p factor across samples. *Assessment*, *28*(4), 1035-1049. <https://doi.org/10.1177/1073191120954921>
- Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology*, *61*(4), 281-288. <https://doi.org/10.1037/cap0000236>
- Linn, R. L. (1997). Evaluating the validity of assessments: the consequences of use. *Educational Measurement: Issues and Practice*, *16*(2), 14-16. <https://doi.org/10.1111/j.1745-3992.1997.tb00587.x>
- Lissitz, R. W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437-448. <https://doi.org/10.3102/0013189x07311286>
- Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3), 285-300. <https://doi.org/10.1037/a0033266>
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635-694. <https://doi.org/10.2466/pr0.1957.3.3.635>
- Lubke, G. H., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*, *41*(4), 499-532. https://doi.org/10.1207/s15327906mbr4104_4
- Lundberg, I., Johnson, R. A., & Stewart, B. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review*, *86*(3), 532-565. <https://doi.org/10.1177/00031224211004187>
- Luyten, P., Campbell, C., Allison, E., & Fonagy, P. (2020). The mentalizing approach to psychopathology: State of the art and future directions. *Annual Review of Clinical Psychology*, *16*(1), 297-325. <https://doi.org/10.1146/annurev-clinpsy-071919-015355>
- Luyten, P., Malcorps, S., Fonagy, P., & Ensink, K. (2019). Assessment of mentalizing. In *Handbook of Mentalizing in Mental Health Practice*. American Psychiatric Publishing.
- Mader, N., Arslan, R., Schmukle, S. C., & Rohrer, J. M. (2023). Emotional (in)stability: Neuroticism is associated with increased variability in negative emotion after all. *Proceedings of the National Academy of Sciences*, *120*(23). <https://doi.org/10.1073/pnas.2212154120>
- Mansolf, M., & Reise, S. P. (2016). Exploratory bifactor analysis: the Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate Behavioral Research*, *51*(5), 698-717. <https://doi.org/10.1080/00273171.2016.1215898>
- Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin*, *137*(5), 856-879. <https://doi.org/10.1037/a0023678>
- Markon, K. E., & Jonas, K. (2016). Structure as cause and representation: Implications of descriptivist inference for structural modeling across multiple levels of analysis. *Journal of Abnormal Psychology*, *125*(8), 1146-1157. <https://doi.org/10.1037/abn0000206>
- Markon, K. E., & Krueger, R. F. (2006). Information-theoretic latent distribution modeling: Distinguishing discrete and continuous latent variable models. *Psychological Methods*, *11*(3), 228-243. <https://doi.org/10.1037/1082-989X.11.3.228>
- Markon, K. E., Krueger, R. F., & Watson, D. (2005). Delineating the structure of normal and abnormal personality: an integrative hierarchical approach. *Journal of Personality and Social Psychology*, *88*(1), 139-157. <https://doi.org/10.1037/0022-3514.88.1.139>
- Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. (2020). Confirmatory Factor Analysis (CFA), Exploratory Structural Equation Modeling (ESEM), and SET-ESEM: Optimal balance between goodness of fit and parsimony. *Multivariate Behavioral Research*, *55*(1), 102-119. <https://doi.org/10.1080/00273171.2019.1602503>
- Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on Hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320-341. https://doi.org/10.1207/s15328007sem1103_2
- Mason, O. (2015). The assessment of schizotypy and its clinical relevance. *Schizophrenia Bulletin*, *41*(suppl 2), S374-S385. <https://doi.org/10.1093/schbul/sbu194>
- Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research & Perspective*, *15*(2), 51-69. <https://doi.org/10.1080/15366367.2017.1348108>

- McGrath, R. E. (2005). Conceptual complexity and construct validity. *Journal of Personality Assessment*, 85(2), 112-124. https://doi.org/10.1207/s15327752jpa8502_02
- McGrath, R. E., & Walters, G. D. (2012). Taxometric analysis as a general strategy for distinguishing categorical from dimensional latent structure. *Psychological Methods*, 17(2), 284-293. <https://doi.org/10.1037/a0026973>
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412-433. <https://doi.org/10.1037/met0000144>
- McNeish, D. (2022). Psychometric properties of sum scores and factor scores differ even when their correlation is 0.98: A response to Widaman and Revelle. *Behavior Research Methods*, 55(8), 4269-4290. <https://doi.org/10.3758/s13428-022-02016-x>
- 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>
- McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, 28(1), 61-88. <https://doi.org/10.1037/met0000425>
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806-834. <https://doi.org/10.1037/10112-043>
- Meehl, P. E. (1979). A funny thing happened to us on the way to the latent entities. *Journal of Personality Assessment*, 43(6), 563-577. https://doi.org/10.1207/s15327752jpa4306_2
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66(1), 195-244. <https://doi.org/10.2466/pr0.1990.66.1.195>
- Meehl, P. E. (1992). Factors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality*, 60(1), 117-174. <https://doi.org/10.1111/j.1467-6494.1992.tb00269.x>
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16-18. <https://doi.org/10.1111/j.1745-3992.1997.tb00588.x>
- Mejia, S. T., Hooker, K., Ram, N., Pham, T. A., & Metoyer, R. (2014). Capturing intraindividual variation and covariation constructs: using multiple time-scales to assess construct reliability and construct stability. *Research in Human Development*, 11(2), 91-107. <https://doi.org/10.1080/15427609.2014.906728>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525-543. <https://doi.org/10.1007/BF02294825>
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30(10), 955-966. <https://doi.org/10.1037/0003-066x.30.10.955>
- Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5-11. <https://doi.org/10.3102/0013189x018002005>
- Messick, S. (1989b). Validity. In *Educational Measurement* (3rd ed.). American Council on Education.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. <https://doi.org/10.1037/0003-066x.50.9.741>
- Mestdagh, M., Pe, M., Pestman, W. R., Verdonck, S., Kuppens, P., & Tuerlinckx, F. (2018). Sidelineing the mean: The relative variability index as a generic mean-corrected variability measure for bounded variables. *Psychological Methods*, 23(4), 690-707. <https://doi.org/10.1037/met0000153>
- Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, 23(1), 116-139. <https://doi.org/10.1080/10705511.2014.961800>
- Müller, S., Wendt, L. P., Spitzer, C., Masuhr, O., Back, S. N., & Zimmermann, J. (2022). A critical evaluation of the Reflective Functioning Questionnaire (RFQ). *Journal of Personality Assessment*, 104(5), 613-627. <https://doi.org/10.1080/00223891.2021.1981346>
- Müller, S., Wendt, L. P., & Zimmermann, J. (2023). Development and validation of the Certainty about Mental States Questionnaire (CAMSQ): A self-report measure of mentalizing oneself and others. *Assessment*, 30(3), 651-674. <https://doi.org/10.1177/10731911211061280>
- Murphy, B. A., & Lilienfeld, S. O. (2019). Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. *Psychological Assessment*, 31(8), 1062-1072. <https://doi.org/10.1037/pas0000732>
- Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, 69(1), 511-534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Newton, P. E., & Baird, J. (2016). The great validity debate. *Assessment in Education: Principles, Policy & Practice*, 23(2), 173-177. <https://doi.org/10.1080/0969594x.2016.1172871>
- Newton, P. E., & Shaw, S. (2013). Standards for talking and thinking about validity. *Psychological Methods*, 18(3), 301-319. <https://doi.org/10.1037/a0032969>

- Newton, P. E., & Shaw, S. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, 23(2), 178-197. <https://doi.org/10.1080/0969594x.2015.1037241>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600-2606. <https://doi.org/10.1073/pnas.1708274114>
- Nunnally, J. C. (1978). An overview of psychological measurement. *Clinical Diagnosis of Mental Disorders: A Handbook*. Springer.
- Oakley, B., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of Abnormal Psychology*, 125(6), 818-823. <https://doi.org/10.1037/abn0000182>
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5), 1596-1618. <https://doi.org/10.3758/s13423-019-01645-2>
- Olaru, G., Schroeders, U., Härtung, J., & Wilhelm, O. (2019). Ant colony optimization and local weighted structural equation modeling. A tutorial on novel item and person sampling procedures for personality research. *European Journal of Personality*, 33(3), 400-419. <https://doi.org/10.1002/per.2195>
- Olderbak, S., & Wilhelm, O. (2020). Overarching principles for the organization of socioemotional constructs. *Current Directions in Psychological Science*, 29(1), 63-70. <https://doi.org/10.1177/0963721419884317>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). <https://doi.org/10.1126/science.aac4716>
- Osborne-Crowley, K. (2020). Social cognition in the real world: Reconnecting the study of social cognition with social reality. *Review of General Psychology*, 24(2), 144-158. <https://doi.org/10.1177/1089268020906483>
- Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, 5(3), 343-355. <https://doi.org/10.1037/1082-989x.5.3.343>
- Pettersson, E., Lichtenstein, P., Larsson, H., D'Onofrio, B. M., Lahey, B. B., & Latvala, A. (2021). Associations of resting heart rate and intelligence with general and specific psychopathology: a prospective population study of 899,398 Swedish men. *Clinical Psychological Science*, 9(3), 524-532. <https://doi.org/10.1177/2167702620961081>
- Pirla, S., Taquet, M., & Quoidbach, J. (2023). Measuring affect dynamics: An empirical framework. *Behavior Research Methods*, 55(1), 285-300. <https://doi.org/10.3758/s13428-022-01829-0>
- Podsakoff, P. M., Podsakoff, N. P., Williams, L. J., Huang, C., & Yang, J. (2024). Common method bias: it's bad, it's complex, it's widespread, and it's not easy to fix. *Annual Review of Organizational Psychology and Organizational Behavior*, 11(1), 17-61. <https://doi.org/10.1146/annurev-orgpsych-110721-040030>
- Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 9-13. <https://doi.org/10.1111/j.1745-3992.1997.tb00586.x>
- Popper, K. R. (1962). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Harper & Row.
- Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the Extreme Groups Approach: A critical reexamination and new recommendations. *Psychological Methods*, 10(2), 178-192. <https://doi.org/10.1037/1082-989x.10.2.178>
- Quesque, F., & Rossetti, Y. (2020). What do Theory-of-Mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, 15(2), 384-396. <https://doi.org/10.1177/1745691619896607>
- Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient Alpha, we still need you! *Educational and Psychological Measurement*, 79(1), 200-210. <https://doi.org/10.1177/0013164417725127>
- Realo, A., Allik, J., Nõlvak, A., Valk, R., Ruus, T., Schmidt, M., & Eilola, T. M. (2003). Mind-reading ability: Beliefs and performance. *Journal of Research in Personality*, 37(5), 420-445. [https://doi.org/10.1016/s0092-6566\(03\)00021-7](https://doi.org/10.1016/s0092-6566(03)00021-7)
- Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment*, 84(2), 126-136. https://doi.org/10.1207/s15327752jpa8402_02
- Remmers, C., Wester, R. A., Repnik, L. G., Plumböhm, M., Unger, S., & Jauk, E. (2023). Psychodynamic theory meets HiTOP: The nomological network between motivational conflicts and dimensions of the hierarchical taxonomy of psychopathology (HiTOP). *Journal of Research in Personality*, 106, 104418. <https://doi.org/10.1016/j.jrp.2023.104418>
- Rhemtulla, M., Borsboom, D., & Van Bork, R. (2017). How to measure nothing. *Measurement: Interdisciplinary Research & Perspective*, 15(2), 95-97. <https://doi.org/10.1080/15366367.2017.1369785>
- Ringwald, W. R., Forbes, M. K., & Wright, A. G. (2023). Meta-analysis of structural evidence for the Hierarchical Taxonomy of Psychopathology (HiTOP) model. *Psychological Medicine*, 1-14. <https://doi.org/10.1017/s0033291721001902>

- Ringwald, W. R., & Wright, A. G. (2022, February 28). Overcoming the confound of means and variability for measuring everyday emotion dynamics related to neuroticism. <https://doi.org/10.31234/osf.io/nxbyd>
- Ro, E., & Clark, L. A. (2009). Psychosocial functioning in the context of diagnosis: Assessment and theoretical issues. *Psychological Assessment*, 21(3), 313-324. <https://doi.org/10.1037/a0016611>
- Roberts, B. W. (2018). A revised sociogenomic model of personality traits. *Journal of Personality*, 86(1), 23-35. <https://doi.org/10.1111/jopy.12323>
- Rodgers, J. L., & Shrout, P. E. (2018). Psychology's replication crisis as scientific opportunity: A précis for policy-makers. *Policy Insights From the Behavioral and Brain Sciences*, 5(1), 134-141. <https://doi.org/10.1177/2372732217749254>
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, 21(2), 137-150. <https://doi.org/10.1037/met0000045>
- Rogers, R., Tazi, K. Y., & Drogin, E. Y. (2023). Forensic assessment instruments: Their reliability and applicability to criminal forensic issues. *Behavioral Sciences & the Law*, 41(5), 415-431. <https://doi.org/10.1002/bsl.2613>
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27-42. <https://doi.org/10.1177/2515245917745629>
- Rosenbusch, H., Wanders, F., & Pit, I. L. (2020). The Semantic Scale Network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychological Methods*, 25(3), 380-392. <https://doi.org/10.1037/met0000244>
- Ruscio, J., Ruscio, A. M., & Carney, L. (2011). Performing taxometric analysis to distinguish categorical and dimensional variables. *Journal of Experimental Psychopathology*, 2(2), 170-196. <https://doi.org/10.5127/jep.010910>
- Russell, M. (2022). Clarifying the terminology of validity and the investigative stages of validation. *Educational Measurement: Issues and Practice*, 41(2), 25-35. <https://doi.org/10.1111/emip.12453>
- Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, 107(11), 2040-2068. <https://doi.org/10.1037/apl0000994>
- Salazar Kämpf, M., Adam, L., Rohr, M. K., Exner, C., & Wieck, C. (2023). A meta-analysis of the relationship between emotion regulation and social affect and cognition. *Clinical Psychological Science*, 11(6), 1159-1189. <https://doi.org/10.1177/21677026221149953>
- Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, 49(5), 407-424. <https://doi.org/10.1080/00273171.2014.931800>
- Savalei, V., & Reise, S. P. (2019). Don't forget the model in your model-based reliability coefficients: A Reply to McNeish (2018). *Collabra*, 5(1). <https://doi.org/10.1525/collabra.247>
- Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, 16(4), 744-755. <https://doi.org/10.1177/1745691620966795>
- Schiavone, S. R., & Vazire, S. (2023). Reckoning with our crisis: An agenda for the field of social and personality psychology. *Perspectives on Psychological Science*, 18(3), 710-722. <https://doi.org/10.1177/17456916221101060>
- Schimmack, U. (2010). What multi-method data tell us about construct validity. *European Journal of Personality*, 24(3), 241-257. <https://doi.org/10.1002/per.771>
- Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology*, 5. <https://doi.org/10.15626/mp.2019.1645>
- Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, 1(2), 199-223. <https://doi.org/10.1037/1082-989X.1.2.199>
- Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, 27(3), 183-198. [https://doi.org/10.1016/s0160-2896\(99\)00024-0](https://doi.org/10.1016/s0160-2896(99)00024-0)
- Schneider, S., & Junghaenel, D. U. (2022). Estimating reliabilities and correcting for sampling error in indices of within-person dynamics derived from intensive longitudinal data. *Behavior Research Methods*, 55(7), 3872-3891. <https://doi.org/10.3758/s13428-022-01995-1>
- Sellbom, M., Kremyar, A. J., & Wygant, D. B. (2021). Mapping MMPI-3 scales onto the hierarchical taxonomy of psychopathology. *Psychological Assessment*, 33(12), 1153-1168. <https://doi.org/10.1037/pas0001049>
- Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, 31(12), 1428-1441. <https://doi.org/10.1037/pas0000623>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2nd ed.). Houghton, Mifflin and Company.
- Shaffer, J. A., DeGeest, D. S., & Li, A. (2016). Tackling the problem of construct proliferation. *Organizational Research Methods*, 19(1), 80-110. <https://doi.org/10.1177/1094428115598239>

- Shaw, M., Cloos, L., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications: Insights from Many Labs 2. *Canadian Psychology*, *61*(4), 289-298. <https://doi.org/10.1037/cap0000220>
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5-24. <https://doi.org/10.1111/j.1745-3992.1997.tb00585.x>
- Sijtsma, K. (2006). Psychometrics in psychological research: role model or partner in science? *Psychometrika*, *71*(3). <https://doi.org/10.1007/s11336-006-1497-9>
- 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>
- Sijtsma, K. (2012). Future of psychometrics: Ask what psychometrics can do for psychology. *Psychometrika*, *77*(1), 4-20. <https://doi.org/10.1007/s11336-011-9242-4>
- Sijtsma, K. (2013). Theory development as a precursor for test validity. In *New Developments in Quantitative Psychology: Presentations From the 77th Annual Psychometric Society Meeting*. Springer.
- Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass*, *2*(1), 414-433. <https://doi.org/10.1111/j.1751-9004.2007.00044.x>
- Simms, L. J., Wright, A. G., Cicero, D. C., Kotov, R., Mullins-Sweatt, S. N., Sellbom, M., Watson, D., Widiger, T. A., & Zimmermann, J. (2022). Development of Measures for the Hierarchical Taxonomy of Psychopathology (HITOP): a collaborative scale development project. *Assessment*, *29*(1), 3-16. <https://doi.org/10.1177/10731911211015309>
- Singh, J., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, *31*(3), 499-513. <https://doi.org/10.1016/j.cpr.2010.11.009>
- Slaney, K. (2017). *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions*. Palgrave Macmillan. <https://doi.org/10.1057/978-1-137-38523-9>
- Smith, G. T. (2005). On construct Validity: Issues of method and measurement. *Psychological Assessment*, *17*(4), 396-408. <https://doi.org/10.1037/1040-3590.17.4.396>
- Smith, G. T., Atkinson, E. G., Davis, H. A., Riley, E. N., & Oltmanns, J. R. (2020). The general factor of psychopathology. *Annual Review of Clinical Psychology*, *16*(1), 75-98. <https://doi.org/10.1146/annurev-clinpsy-071119-115848>
- Smith, G. T., McCarthy, D. M., & Zapsolski, T. C. B. (2009). On the value of homogeneous constructs for construct validation, theory testing, and the description of psychopathology. *Psychological Assessment*, *21*(3), 272-284. <https://doi.org/10.1037/a0016699>
- Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The Life Outcomes of Personality Replication Project. *Psychological Science*, *30*(5), 711-727. <https://doi.org/10.1177/0956797619831612>
- Soto, C. J., Napolitano, C. M., & Roberts, B. W. (2021). Taking skills seriously: Toward an integrative model and agenda for social, emotional, and behavioral skills. *Current Directions in Psychological Science*, *30*(1), 26-33. <https://doi.org/10.1177/0963721420978613>
- Soto, C. J., Napolitano, C. M., Sewell, M. N., Yoon, H. J., & Roberts, B. W. (2022). An integrative framework for conceptualizing and assessing social, emotional, and behavioral skills: The BESSI. *Journal of Personality and Social Psychology*, *123*(1), 192-222. <https://doi.org/10.1037/pspp0000401>
- Stanton, K., McDonnell, C. G., Hayden, E. P., & Watson, D. (2020). Transdiagnostic approaches to psychopathology measurement: Recommendations for measure selection, data analysis, and participant recruitment. *Journal of Abnormal Psychology*, *129*(1), 21-28. <https://doi.org/10.1037/abn0000464>
- Stanton, K., Watts, A. L., Levin-Aspensson, H. F., Carpenter, R. W., Emery, N. N., & Zimmerman, M. (2023). Focusing narrowly on model fit in factor analysis can mask construct heterogeneity and model misspecification: Applied demonstrations across sample and assessment types. *Journal of Personality Assessment*, *105*(1), 1-13. <https://doi.org/10.1080/00223891.2022.2047060>
- Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A. J., Bröder, A., Cox, G. E., Criss, A. H., Curl, R., Dobbins, I. G., Dunn, J. C., Enam, T., Evans, N. J., Farrell, S., Fraundorf, S. H., Gronlund, S. D., Heathcote, A., Heck, D. W., Hicks, J. L., . . . Wilson, J. H. (2019). Assessing theoretical conclusions with blinded inference to investigate a potential inference crisis. *Advances in Methods and Practices in Psychological Science*, *2*(4), 335-349. <https://doi.org/10.1177/2515245919869583>
- Statham, V., Emerson, L., & Rowse, G. (2019). A systematic review of self-report measures of paranoia. *Psychological Assessment*, *31*(2), 139-158. <https://doi.org/10.1037/pas0000645>
- Steger, D., Jankowsky, K., Schroeders, U., & Wilhelm, O. (2023). The road to hell is paved with good intentions: How common practices in scale construction hurt validity. *Assessment*, *30*(6), 1811-1824. <https://doi.org/10.1177/10731911221124846>

- Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits—revised. *Annual Review of Clinical Psychology*, *11*(1), 71–98. <https://doi.org/10.1146/annurev-clinpsy-032813-153719>
- Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*(1), 1–25. <https://doi.org/10.1146/annurev.clinpsy.032408.153639>
- Tabb, K. (2019). Philosophy of psychiatry after diagnostic kinds. *Synthese*, *196*(6), 2177–2195. <https://doi.org/10.1007/s11229-017-1659-6>
- Thielmann, I., & Hilbig, B. E. (2019). Nomological consistency: A comprehensive test of the equivalence of different trait indicators for the same constructs. *Journal of Personality*, *87*(3), 715–730. <https://doi.org/10.1111/jopy.12428>
- Thompson, R. J., Springstein, T., & Boden, M. T. (2021). Gaining clarity about emotion differentiation. *Social and Personality Psychology Compass*, *15*(3). <https://doi.org/10.1111/spc3.12584>
- Vachon, D. D., & Lynam, D. R. (2016). Fixing the problem with empathy. *Assessment*, *23*(2), 135–149. <https://doi.org/10.1177/1073191114567941>
- van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & Van Der Maas, H. L. J. (2017). What is the p-factor of psychopathology? Some risks of general factor modeling. *Theory & Psychology*, *27*(6), 759–773. <https://doi.org/10.1177/0959354317737185>
- Vazire, S. (2018). Implications of the Credibility Revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. <https://doi.org/10.1177/1745691617751884>
- Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility Beyond Replicability: Improving the four validities in Psychological science. *Current Directions in Psychological Science*, *31*(2), 162–168. <https://doi.org/10.1177/09637214211067779>
- Vize, C., & Wright, A. G. C. (2024). Translating the transdiagnostic: aligning assessment practices with research advances. *Assessment*, *31*(1), 199–215. <https://doi.org/10.1177/10731911231194996>
- Wagenmakers, E., Wetzels, R., Borsboom, D., Van Der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. <https://doi.org/10.1177/1745691612463078>
- Watson, D., Levin-Aspensson, H. F., Waszczuk, M. A., Conway, C., Dalgleish, T., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, K. T., Hobbs, K. A., Michelini, G., Nelson, B. D., Sellbom, M., Slade, T., South, S. C., Sunderland, M., Waldman, I. D., Witthöft, M., Wright, A. G., . . . Author_Id, N. (2022b). Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): III. Emotional dysfunction superspectrum. *World Psychiatry*, *21*(1), 26–54. <https://doi.org/10.1002/wps.20943>
- Watts, A. L., Greene, A. L., Bonifay, W., & Fried, E. I. (2023). A critical evaluation of the p-factor literature. *Nature Reviews Psychology*. <https://doi.org/10.1038/s44159-023-00260-2>
- Watts, A. L., Makol, B. A., Palumbo, I. M., De Los Reyes, A., Olino, T. M., Latzman, R. D., DeYoung, C. G., Wood, P. K., & Sher, K. J. (2021). How robust is the p Factor? Using multi-trait-multimethod modeling to inform the meaning of general factors of youth psychopathology. *Clinical Psychological Science*, *10*(4), 640–661. <https://doi.org/10.1177/21677026211055170>
- Waugh, C. E., & Kuppens, P. (2021). *Affect dynamics*. Springer Nature.
- Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267–295. <https://doi.org/10.1037/emo0000226>
- Weimer, W. B. (1979). *Notes on the Methodology of Scientific Research*. Erlbaum.
- Wendt, L. P., Jankowsky, K., Schroeders, U., London Personality and Mood Disorder Research Consortium, Nolte, T., Fonagy, P., Montague, P. R., Zimmermann, J., and Olaru, G. (2023). Mapping established psychopathology scales onto the Hierarchical Taxonomy of Psychopathology (HiTOP). *Personality and Mental Health*, *17*(2), 117–134. <https://doi.org/10.1002/pmh.1566>
- Wendt, L. P., Wright, A. G., Pilkonis, P. A., Nolte, T., Fonagy, P., Montague, P. R., Benecke, C., Krieger, T., & Zimmermann, J. (2019). The latent structure of interpersonal problems: Validity of dimensional, categorical, and hybrid models. *Journal of Abnormal Psychology*, *128*(8), 823–839. <https://doi.org/10.1037/abn0000460>
- Wendt, L. P., Wright, A. G., Pilkonis, P. A., Woods, W. C., Denissen, J. J. A., Kühnel, A., & Zimmermann, J. (2020). Indicators of affect dynamics: Structure, reliability, and personality correlates. *European Journal of Personality*, *34*(6), 1060–1072. <https://doi.org/10.1002/per.2277>
- Wendt, L. P., Zimmermann, J., Spitzer, C., & Müller, S. (2024). Mindreading measures misread? A multimethod investigation into the validity of self-report and task-based approaches, *36*(5), 365–378. *Psychological Assessment*. <https://doi.org/10.1037/pas0001310>

- Wenzel, M., & Brose, A. (2023). Addressing measurement issues in affect dynamic research: Modeling emotional inertia's reliability to improve its predictive validity of depressive symptoms. *Emotion*, 23(2), 412-424. <https://doi.org/10.1037/emo0001108>
- Wenzel, M., & Kubiak, T. (2020). Neuroticism may reflect emotional variability when correcting for the confound with the mean. *Proceedings of the National Academy of Sciences*, 117(52), 32857-32858. <https://doi.org/10.1073/pnas.2017910117>
- Wenzel, M., Rowland, Z., Mey, L. K., Kurth, K., Tüscher, O., & Kubiak, T. (2023). Variability in negative affect is an important feature of neuroticism above mean negative affect once measurement issues are accounted for. *European Journal of Personality*, 37(3), 338-351. <https://doi.org/10.1177/08902070221089139>
- Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, 84(3), 608-618. <https://doi.org/10.1037/0022-3514.84.3.608>
- Westen, D., & Rosenthal, R. (2005). Improving construct validity: Cronbach, Meehl, and Neurath's ship: Comment. *Psychological Assessment*, 17(4), 409-412. <https://doi.org/10.1037/1040-3590.17.4.409>
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS One*, 11(3), e0152719. <https://doi.org/10.1371/journal.pone.0152719>
- Wetzel, E., & Roberts, B. W. (2020). Commentary on Hussey and Hughes (2020): Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, 3(4), 505-508. <https://doi.org/10.1177/2515245920957618>
- Widaman, K. F., & Revelle, W. (2023a). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, 55(2), 788-806. <https://doi.org/10.3758/s13428-022-01849-w>
- Widaman, K. F., & Revelle, W. (2023b). Thinking about sum scores yet again, maybe the last time, we don't know, oh no. A Comment on McNeish (2023). *Educational and Psychological Measurement*. Advance online publication. <https://doi.org/10.1177/00131644231205310>
- Wolf, M. G. (2023, June 7). The problem with over-relying on quantitative evidence of validity. <https://doi.org/10.31234/osf.io/v4nb2>
- Woo, S. E., LeBreton, J. M., Keith, M. G., & Tay, L. (2023). Bias, fairness, and validity in graduate-school admissions: A psychometric perspective. *Perspectives on Psychological Science*, 18(1), 3-31. <https://doi.org/10.1177/17456916211055374>
- Wood, D., Gardner, M. H., & Harms, P. D. (2015). How functionalist and process approaches to behavior can explain trait covariation. *Psychological Review*, 122(1), 84-111. <https://doi.org/10.1037/a0038423>
- Wright, A. G., Beltz, A. M., Gates, K. M., Molenaar, P., & Simms, L. J. (2015). Examining the dynamic structure of daily internalizing and externalizing behavior at multiple levels of analysis. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.01914>
- Wright, A. G., Krueger, R. F., Hobbs, M. J., Markon, K. E., Eaton, N. R., & Slade, T. (2013). The structure of psychopathology: Toward an expanded quantitative empirical model. *Journal of Abnormal Psychology*, 122(1), 281-294. <https://doi.org/10.1037/a0030133>
- Wright, A. G., Pincus, A. L., & Hopwood, C. J. (2023). Contemporary integrative interpersonal theory: Integrating structure, dynamics, temporal scale, and levels of analysis. *Journal of Psychopathology and Clinical Science*, 132(3), 263-276. <https://doi.org/10.1037/abn0000741>
- Wright, A. G., & Simms, L. J. (2015). A metastructural model of mental disorders and pathological personality traits. *Psychological Medicine*, 45(11), 2309-2319. <https://doi.org/10.1017/s0033291715000252>
- Wright, A. G., Stepp, S. D., Scott, L. N., Hallquist, M. N., Beeney, J. E., Lazarus, S. A., & Pilkonis, P. A. (2017). The effect of pathological narcissism on interpersonal and affective processes in social interactions. *Journal of Abnormal Psychology*, 126(7), 898-910. <https://doi.org/10.1037/abn0000286>
- Wright, A. G., & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological Assessment*, 31(12), 1467-1480. <https://doi.org/10.1037/pas0000685>
- Yao, J., Lim, S., Guo, C. Y., Ou, A. Y., & Ng, J. W. X. (2022). Experienced incivility in the workplace: A meta-analytical review of its construct validity and nomological network. *Journal of Applied Psychology*, 107(2), 193-220. <https://doi.org/10.1037/apl0000870>
- Yarkoni, T. (2020). Implicit realism impedes progress in psychology: Comment on Fried (2020). *Psychological Inquiry*, 31(4), 326-333. <https://doi.org/10.1080/1047840x.2020.1853478>
- Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45. <https://doi.org/10.1017/s0140525x20001685>
- Yeung, E., Apperly, I. A., & Devine, R. T. (2024). Measures of individual differences in adult theory of mind: A systematic review. *Neuroscience & Biobehavioral Reviews*, 157, 105481. <https://doi.org/10.1016/j.neubiorev.2023.105481>

- Ziegler, M. (2020). Psychological Test Adaptation and Development - How papers are structured and why. *Psychological Test Adaptation and Development*, 1(1), 3-11. <https://doi.org/10.1027/2698-1866/a000002>
- Zimmermann, J., Mayer, A., Leising, D., Krieger, T., Holtforth, M. G., & Pretsch, J. (2017). Exploring occasion specificity in the assessment of DSM-5 Maladaptive personality traits. *European Journal of Psychological Assessment*, 33(1), 47-54. <https://doi.org/10.1027/1015-5759/a000271>
- Zumbo, B. D. (2007). Validity: Foundational Issues and Statistical Methodology. In *Handbook of Statistics, Vol. 26: Psychometrics*. Elsevier Science.
- Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In *The SAGE Handbook of Quantitative Methodology for the Social Sciences*. Sage Press.

Article 1:

Indicators of Affect Dynamics: Structure, Reliability, and Personality Correlates

Leon P. Wendt¹, Aidan G.C. Wright², Paul A. Pilkonis³, William C. Woods²,

Jaap J.A. Denissen^{4,5}, Anja Kühnel⁶ & Johannes Zimmermann¹

¹ Department of Psychology, University of Kassel

² Department of Psychology, University of Pittsburgh

³ Department of Psychiatry, University of Pittsburgh School of Medicine

⁴ Department of Developmental Psychology, Tilburg University

⁵ Department of Developmental Psychology, Utrecht University

⁶ Department of Psychology, MSB Medical School Berlin

Status:

Published

Supplemental Materials:

Data, R code, and other materials

<https://doi.org/10.17605/OSF.IO/6GHCX>

Citation:

Wendt, L. P., Wright, A. G., Pilkonis, P. A., Woods, W. C., Denissen, J. J., Kühnel, A., & Zimmermann, J. (2020). Indicators of affect dynamics: Structure, reliability, and personality correlates. *European Journal of Personality*, 34(6), 1060-1072. <https://doi.org/10.1002/per.2277>

Indicators of Affect Dynamics: Structure, Reliability, and Personality Correlates

LEON P. WENDT^{1*}, AIDAN G.C. WRIGHT², PAUL A. PILKONIS³, WILLIAM C. WOODS²,
JAAP J.A. DENISSEN^{4,5}, ANJA KÜHNEL⁶ and JOHANNES ZIMMERMANN¹

¹Department of Psychology, University of Kassel, Kassel, Germany

²Department of Psychology, University of Pittsburgh, Pittsburgh, PA USA

³Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA USA

⁴Department of Developmental Psychology, Tilburg University, Tilburg, The Netherlands

⁵Department of Developmental Psychology, Utrecht University, Utrecht, The Netherlands

⁶Department of Psychology, MSB Medical School Berlin, Berlin, Germany

Abstract: Researchers are increasingly interested in the affect dynamics of individuals for describing and explaining personality and psychopathology. Recently, the incremental validity of more complex indicators of affect dynamics (IADs; e.g. autoregression) has been called into question (Dejonckheere et al., 2019), with evidence accumulating that these might convey little unique information beyond mean level and general variability of emotions. Our study extends the evidence for the construct validity of IADs by investigating their redundancy and uniqueness, split-half reliability based on indices from odd-numbered and even-numbered days, and association with big five personality traits. We used three diverse samples that assessed daily and momentary emotions, including community participants, individuals with personality pathology, and their significant others (total $N = 1192$, total number of occasions = 51 278). Mean and variability of affects had high reliability and distinct nomological patterns to big five personality traits. In contrast, more complex IADs exhibited substantial redundancies with mean level and general variability of emotions. When partialing out these redundancies by using residual variables, some of the more complex IADs had acceptable reliability, but only a few of these showed incremental associations with big five personality traits, indicating that IADs have limited validity using the current assessment practices. © 2020 The Authors. European Journal of Personality published by John Wiley & Sons Ltd on behalf of European Association of Personality Psychology

Key words: affect dynamics; reliability; residual variables; structure; personality traits

INDICATORS OF AFFECT DYNAMICS: STRUCTURE, RELIABILITY, AND PERSONALITY CORRELATES

Major personality models include emotions as part of personality along with motivations, cognitions, and behavioral dispositions (Carver, Sutton, & Scheier, 2000). Researchers have growing interest in how emotions unfold and interact with each other dynamically across time (i.e. affect dynamics) and how these processes relate to diverse psychological phenomena (e.g. Trull, Lane, Koval, & Ebner-Priemer, 2015). For example, affect dynamics have been targeted for describing and explaining well-being (Dejonckheere et al., 2019;

Houben, Van Den Noortgate, & Kuppens, 2015), mood disorders (e.g. Bos, de Jonge, & Cox, 2019), borderline personality pathology (e.g. Mneimne, Fleeson, Arnold, & Furr, 2018), transdiagnostic dimensions of psychopathology (Scott et al., 2020), and normal-range personality differences (e.g. Kalokerinos et al., 2020; Kuppens, Van Mechelen, Nezlek, Dossche, & Timmermans, 2007). Affect dynamics can be measured by various person-specific summary statistics of an emotional time series [e.g. mean of states, standard deviation (*SD*), autoregression, and mean square successive differences (*MSSD*)] using intensive longitudinal research designs (Csikszentmihalyi & Larson, 1987; Hamaker & Wichers, 2017). Hereinafter, we refer to those statistics as indicators of affect dynamics (IADs).


Conceptually, it is presumed that the various IADs each capture distinct and meaningful features of the individuals' emotional experience (in other words, that IADs possess construct validity). The validity of IADs is commonly evaluated by their associations to other relevant constructs (i.e. criterion validity). Critically, past research has demonstrated that manifold redundancies exist between IADs, including mathematical interdependencies (e.g. Jahng, Wood, & Trull, 2008; Mestdagh et al., 2018) and possibly conceptual overlap. These redundancies have to be considered when criterion validity is evaluated, because associations found between IADs and other relevant

*Correspondence to: Leon Wendt, Department of Psychology, University of Kassel, Holländische Str. 36-38, 34127, Kassel, Germany.

E-mail: l.wendt@uni-kassel.de

Ethics committee approval was obtained for data collection (Sample 2, University at Buffalo Institutional Review Board, PRO16080767; Sample 3, University of Pittsburgh Institutional Review Board #12030125) or was not required when data was collected (Sample 1).

Current address: Jaap J. A. Denissen, Utrecht University, Utrecht, The Netherlands

 This article earned Open Data and Open Materials badges through Open Practices Disclosure from the Center for Open Science: <https://osf.io/tyxyz/wiki>. The data and materials are permanently and openly accessible at <https://osf.io/6ghcx/>. Author's disclosure form may also be found at the Supporting Information in the online version.

constructs might be non-specific, thereby undercutting the unique interpretations that presumably justify the use of these metrics. For example, non-specific associations were documented for the criteria of depression (e.g. Koval, Pe, Meers, & Kuppens, 2013), well-being (e.g. Houben *et al.*, 2015), and neuroticism (Kalokerinos *et al.*, 2020). Indeed, the accumulating evidence casts doubt on the incremental validity of more complex IADs (e.g. autoregression) beyond mean level and general variability of emotions (Bos *et al.*, 2019; Dejonckheere *et al.*, 2019). To date, there is still insufficient knowledge about the reliability and validity of more complex IADs using the current assessment practices.

In the following, we describe the IADs considered in the current study along with their common interpretation. First, given a sufficient number of repeated measurements, the individual mean of states (M) is a stable feature of individuals (Watson & Clark, 1999) and can be regarded as a good approximation of *trait affect* (e.g. Watson & Tellegen, 1985). Second, the individual SD is generally referred to as *emotional variability*, with past research indicating that it might be a stable and substantive trait even when controlling for its overlap with the mean (Eid & Diener, 1999). Third, the individual $MSSD$ captures the strength of sudden fluctuations in the process. High values of $MSSD$ have been interpreted as *emotional instability* (Jahng *et al.*, 2008). Fourth, the individual strength of autoregression, which is the likelihood of remaining in a particular affective state from observation to observation, has been interpreted as *emotional inertia* (Kuppens *et al.*, 2012; Kuppens, Allen, & Sheeber, 2010). Fifth, several statistics capture types of *emotion differentiation*, defined as the degree to which individuals report distinct emotional states. This concept may be applied to differentiating between affects (e.g. individual contemporaneous correlation between positive affect (PA) and negative affect (NA): affective bipolarity; Dejonckheere *et al.*, 2018) or differentiating between more fine-grained emotional states within affects (e.g. the individual average item intercorrelation of positively valenced emotions). Sixth, *cross-lagged effects* can be used to describe how distinct affects predict each other across time as operationalized by temporal networks from dynamic network models (Epskamp, Waldorp, Möttus, & Borsboom, 2018). Trait affect and also (to a somewhat lesser extent) emotional variability have received most support for their validity. In addition, M and SD yield the most parsimonious description of an emotional time series when compared against more complex IADs, as the calculation of the former disregards the inherent temporal sequence of repeated measurement. In contrast, more complex IADs do consider the temporal sequence.

The current ambiguity about the validity of IADs impedes research progress on affect dynamics. Three samples were used for secondary analysis in order to shed more light on this issue in several ways: (i) elucidating potential patterns of redundancy by investigating the structure of IADs, (ii) estimating their reliability as a prerequisite for validity, and (iii) extending their nomological network to big five personality traits. Big five personality traits are especially important to consider for tests of criterion validity, as those provide an established framework for capturing major psychological differences between individuals. This study included

heterogenous samples in order to achieve generalizability across populations (i.e. community participants, individuals with personality pathology and their significant others) and sampling frames (i.e. daily and momentary data on emotions). In order to establish the incremental information of IADs, we controlled statistically for redundancies with mean level and general variability of an emotional process by using residual variables. More specifically, the SD statistics had scale means (i.e. residual variable Type I) and more complex IADs had scale means and SD s partialled out (i.e. residual variable Type II). As a result, our residual variables captured the individuals' relative score in relation to what would be expected, given the individuals' mean (and variability) on affect scales.

METHOD

Participants

Participants who completed at least 20 consecutive measurement occasions¹ were selected for the current analyses, resulting in N total participants = 1192 and t total occasions = 51 278. Big five personality traits at baseline were assessed in all samples. A detailed overview of sample characteristics is given in Table 1.

The first sample was based on the Berlin Diary Study (Denissen & Kühnel, 2008). Participants received daily questionnaires containing retrospective measurement of affect over 30 days that were filled out before going to bed. The second sample (Wright *et al.*, 2019) included individuals with a personality disorder diagnosis who completed daily retrospective assessments of affect over 100 days. The third sample (Wright *et al.*, 2017) consisted of dyads who completed a 21-day period of ecological momentary assessment (3.7 assessments per day on average) including individuals who were engaged in outpatient psychiatric treatment and their significant others. In Sample 3, momentary affect was assessed multiple times a day following social interactions (i.e. event-contingent assessment).

Measures

Daily and momentary emotions

Emotion adjectives were used to assess daily and momentary affect. In Sample 1, participants were asked to indicate to what degree emotions were descriptive of how they *generally felt today* on a 5-point scale, ranging from 0 (*not at all*) to 4 (*extreme*). In Sample 2, participants were asked about the extent to which they had *felt this way over the past 24 hours* on a 5-point scale, ranging from 0 (*very slightly*) to 4 (*extremely*). In Sample 3, participants were asked multiple times a day to rate their momentary emotions on a 5-point scale ranging from 1 (*very slightly or not at all*) to 5 (*extremely*). IADs were calculated for scales of PA, NA, and hostile affect (HA), as those were consistently identified

¹The inclusion criterion of 20 measurement occasions is a common, yet arbitrary, threshold.

Table 1. Sample characteristics

Sample #	1	2	3
Further described in	Denissen and Kühnel, 2008	Wright <i>et al.</i> , 2019; Wright, Beltz, Gates, Molenaar, & Simms, 2015	Wright <i>et al.</i> , 2017
Emotion assessment type	Daily Dairy	Daily Dairy	Momentary assessment
Language	German	English (USA)	English (USA)
Population characteristics	~50% students	Individuals diagnosed with any personality disorder	Outpatients screened for personality pathology and their romantic partners
% of participants that satisfied the inclusion criterion (>20 occasions)	41.7%	89.3%	88.1%
<i>N</i> subjects included	870	100	222
<i>M</i> age	30.6	45.0	29.7
% female	87%	65%	77%
Minimum–maximum days	21–29	43–101	7–33
Total <i>t</i> measurement occasions	21709 days	9017 days	20552 beeps
Average <i>t</i> days	25.0	90.2	22.7
Average <i>t</i> beeps per day	/	/	4.0
<i>M</i> measurement occasions per split half for calculating split-half reliability	/	45.1 days	46.2 beeps
<i>N</i> subjects included for calculating split-half reliability	/	100	143
Number of emotion adjectives	26	10	25
Positive affect items	Active Attentive Determined Inspired Enthusiastic Excited Interested Proud Strong Content Pleased Happy Aroused Hyperactivated	Active Attentive Determined Inspired Alert	Active Attentive Determined Inspired Enthusiastic Excited Interested Proud Strong Alert
Negative affect items	Afraid Ashamed Distressed Guilty Scared Alone Miserable Troubled Unhappy	Afraid Ashamed Nervous	Afraid Ashamed Distressed Guilty Scared Alone Nervous Jittery
Hostile affect items	Hostile Upset Irritable	Hostile Upset	Hostile Upset Irritable Loathing Disgusted Angry Scornful
Assessment of personality traits	Big Five Inventory (Lang, Lütke & Asendorpf, 2001)	NEO Five-Factor Inventory (Costa & McCrae, 1992)	Revised NEO Personality Inventory (Costa & McCrae, 1992)
<i>N</i> subjects with assessment of personality traits available	870	99	193

across samples (see Results section). The included emotion adjectives are enlisted in Table 1.

Personality traits

In all samples, we assessed the big five personality traits (i.e. openness to new experiences, conscientiousness, extraversion, agreeableness, and neuroticism). In Sample 1, the

German version of the Big Five Inventory (Lang *et al.*, 2001) was used. Participants rate 42 statements on a 5-point scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). In Sample 2, the 60-item NEO Five-Factor Inventory (Costa & McCrae, 1992) was used. In Sample 3, the Revised NEO Personality Inventory (Costa & McCrae, 1992), consisting of 240 items, was used. For the

latter two NEO instruments, items were rated on a 5-point scale ranging from 0 (*strongly disagree*) to 4 (*strongly agree*).

Statistical analysis

Selection and computation of IADs

Affect scales were included for calculating IADs when they aligned with within-person factors identified by multilevel exploratory factor analysis (ML-EFA; Reise, Ventura, Nuechterlein, & Kim, 2005). The ideal number of within-person factors was selected with respect to interpretability and a combination of model fit indices, including the comparative fit index (CFI), root mean square error of approximation (RMSEA), and the models' improvements in level-specific fit using standardized root mean square residual (SRMR-within and SRMR-between; Kim, Dedrick, Cao, & Ferron, 2016).

The included IADs comprised univariate statistics (Jahng *et al.*, 2008) and model-based statistics (Epskamp *et al.*, 2018). The calculation of univariate IADs was based on rescaled affect scores with values ranging from 0 to 10 for facilitating cross-sample comparisons. Univariate IADs included individual scale mean (M), SD , $MSSD$, and the average item intercorrelation (\bar{r}). We further included corrected indices of emotional variability (i.e. SD^c) and emotional instability (i.e. $MSSD^c$), that have their theoretical maxima adjusted for the observed individual mean (Mestdagh *et al.*, 2018). Model-based IADs were derived from multilevel lag-1 vector autoregressive dynamic networks (Bringmann *et al.*, 2013, 2016; Epskamp *et al.*, 2018). Individual network parameters (also, random effects) including autoregressive effects (i.e. ϕ_{PP} , ϕ_{NN} , ϕ_{HH}), cross-lagged effects (e.g. ϕ_{NP} , ϕ_{PN}), and contemporaneous partial correlations (i.e. r_{PN} , r_{PH} , r_{NH}) were estimated using person-mean centering of z -standardized daily/momentary affect scores.² Non-subsequent measurement occasions were removed from network estimation including overnight lags in case of momentary assessment in Sample 3. Network summary statistics (i.e. node centralities, network density) were not considered, because those were unlikely to be useful for small networks. We evaluated the level and variability of (standardized) individual random effects based on the guidelines of Gignac and Szodorai (2016). The median (Mdn) of individual random effects was regarded as small ($\sim .10$), moderate ($\sim .20$), or large ($\sim .30$), indicating the size of model-based IADs for the average individual. Similarly, the interquartile range (IQR) of individual random effects was regarded as small ($\sim .10$), moderate ($\sim .20$), or large ($\sim .30$), indicating the amount of individual differences in model-based IADs.

Structure of indicators of affect dynamics

Several analytic steps were taken to delineate the structure of the 30 calculated IADs and elucidate their potential redundancies. First, we used parallel analysis and visual inspection

of the scree plot (i.e. elbow criterion).³ Second, the Spearman correlation matrix of IADs was used to extract varimax rotated principal components.⁴ Third, we investigated IADs' particular overlap with affect M s and SD s using the multiple correlation coefficient (R). For emotional variability statistics (i.e. SD , SD^c), we computed their multiple correlation with means, and for more complex statistics (i.e. $MSSD$, $MSSD^c$, \bar{r} , individual network parameters), we computed their multiple correlation with scale means and SD s.

Split-half reliability

Split-half reliability was used to evaluate whether IADs consistently measure the same constructs at the between-person level (e.g. Mejía, Hooker, Ram, Pham, & Metoyer, 2014). To this end, residualized IADs were calculated separately based on time series including only odd or even days, and correlations between split-halves were computed. The correlations between split-halves were then corrected using the Spearman–Brown prophecy formula in order to derive reliability estimates of the IADs based on the complete time series (r_{sb}). In Sample 2, split-half intervals consisted of 45.1 days on average (odd days = 44.5, even days = 45.7). In Sample 3, split-half intervals consisted of 11.4 days and 46.2 momentary occasions on average (odd days = 44.1, even days = 48.4). Split-half reliability was not calculated in Sample 1, as fewer measurement occasions per person were available. Reliability was regarded as low ($\sim .50$), moderate ($\sim .70$), or high ($\sim .90$).

Associations with personality traits

The incremental validity of IADs was evaluated by their bivariate correlations with self-report measures of personality traits using residual variables. Pearson correlations were calculated and Fisher z -transformed in each sample separately, before integrating them in a random effects meta-analysis. In the current study, significant meta-analytic correlations indicate that IADs are incrementally associated with big five personality traits (i.e. across populations and sampling frames) beyond mean level and general variability of affect.

Software packages

Openly accessible R scripts and data are provided that allow reproduction of the reported statistical analyses at <https://osf.io/6ghcx/>. All analyses were executed with the statistical environment R version 3.6.1 (R Core Team, 2019). ML-EFA was estimated using the WLSMV estimator and oblique geomin rotation in Mplus 8.0 (Muthén & Muthén, 2018). SD^c and $MSSD^c$ were calculated using the *relativeVariability* package version 1.0 (Mestdagh *et al.*, 2018). Principal components, scree plots, and parallel analysis were computed by the *psych* package version 1.8.18 (Revelle, 2018). Multilevel dynamic networks were estimated by the *mIVAR* package version 0.4.3 (Epskamp, Deserno, & Bringmann, 2019).

²The full results on network model parameters (i.e. fixed effects, random effect variances) will be made available by the corresponding author upon request.

³Exploratory factor analysis was considered; however, solutions had bad fit or did not converge.

⁴Quartimin rotation was considered and produced similar results.

Table 2. Median, interquartile range, and skew of raw IADs

Statistic	Sample 1			Sample 2			Sample 3		
	Mdn	IQR	Skew	Mdn	IQR	Skew	Mdn	IQR	Skew
PA <i>M</i>	4.49	1.59	-0.23	4.05	2.55	0.26	2.39	2.05	0.76
NA <i>M</i>	1.68	1.98	0.97	1.50	2.39	1.38	0.26	0.54	2.52
HA <i>M</i>	1.77	1.73	0.86	1.65	1.66	1.68	0.35	0.61	2.78
PA <i>SD</i>	1.27	0.57	0.47	1.44	0.70	0.44	1.30	0.64	0.87
NA <i>SD</i>	1.33	0.82	0.20	1.28	0.82	0.17	0.49	0.59	1.20
HA <i>SD</i>	1.77	0.88	0.06	1.64	0.99	0.22	0.74	0.71	0.91
PA <i>SD</i> ^c	0.27	0.12	0.42	0.33	0.15	0.71	0.33	0.12	0.57
NA <i>SD</i> ^c	0.37	0.15	0.39	0.40	0.13	0.49	0.32	0.14	1.10
HA <i>SD</i> ^c	0.49	0.17	0.26	0.51	0.18	0.23	0.42	0.16	0.66
PA <i>MSSD</i>	2.36	2.26	1.71	2.99	2.94	1.93	2.32	2.34	1.93
NA <i>MSSD</i>	2.45	3.15	1.60	2.28	3.02	1.29	0.38	0.68	2.39
HA <i>MSSD</i>	4.85	5.17	1.24	4.06	4.59	1.39	0.86	1.49	2.60
PA <i>MSSD</i> ^c	0.03	0.03	1.81	0.05	0.05	1.84	0.06	0.05	2.26
NA <i>MSSD</i> ^c	0.08	0.07	2.42	0.10	0.08	2.51	0.07	0.06	2.98
HA <i>MSSD</i> ^c	0.15	0.13	2.41	0.17	0.11	0.76	0.12	0.10	1.92
PA \bar{r}	.41	.14	0.57	.68	.36	0.15	.50	.39	0.70
NA \bar{r}	.47	.16	0.59	.66	.46	0.06	.41	.31	0.82
HA \bar{r}	.65	.22	-0.01	1.00	.30	-0.72	.64	.48	0.17
ϕ PP	.17	.08	0.19	.26	.17	0.96	.31	.24	0.57
ϕ NN	.22	.04	0.59	.26	.10	0.70	.32	.27	1.17
ϕ HH	.12	.08	0.53	.13	.09	0.73	.18	.13	1.03
ϕ PN	.00	.04	-0.01	-.01	.04	-0.24	-.01	.03	-0.86
ϕ PH	.01	.05	0.19	-.01	.05	0.63	-.01	.02	-0.72
ϕ NP	-.01	.08	0.59	-.01	.03	0.47	-.02	.07	-0.79
ϕ NH	.02	.01	-0.22	.12	.05	-0.58	.14	.08	1.27
ϕ HP	.01	.06	0.19	.01	.03	-0.11	.00	.03	0.13
ϕ HN	.01	.06	0.10	.02	.04	-0.47	.01	.04	0.76
<i>r</i> PN	-.43	.18	0.14	.04	.14	0.15	-.02	.12	0.47
<i>r</i> PH	-.04	.06	-0.11	.00	.13	0.63	-.05	.08	-0.35
<i>r</i> NH	.41	.20	-0.36	.47	.24	-0.13	.62	.27	-0.15

Note: IADs are based on rescaled values of affect scales in the range of 0 and 10. Mdn, median; IQR, interquartile range; *M*, mean; *SD*, standard deviation; *MSSD*, mean square successive differences; *SD*^c, corrected standard deviation; *MSSD*^c, corrected mean square successive differences; \bar{r} , average item intercorrelation of affect scales; ϕ , autoregressions and cross-lagged effects; *r*, contemporaneous partial correlations.

Random effects meta-analysis was estimated by the *metafor* package version 2.1-0 (Viechtbauer, 2010).

RESULTS

Selection of affect dimensions

Considering both fit and interpretability, ML-EFA solutions were retained that indicated five (Sample 1) or three within-person factors (Samples 2 and 3) achieving acceptable fit, RMSEA \leq .044, CFI \geq .929, SRMR-within \leq .039. PA, NA, and HA were consistently identified in all samples and were therefore used for calculating IADs. Additional factors were identified (i.e. factors of tiredness and calmness) but were not considered for calculating IADs, as those were only present in Sample 1. Further information on ML-EFA models are displayed in the supporting information (see Table S1 for fit statistics and Tables S2-S4 for the estimated within-person factor loadings).

Descriptive statistics

Median, IQR, and skew of raw IADs are displayed in Table 2. Individual mean of affect was high for PA as compared to

NA and HA, indicating that individuals tended to report positive emotions more often than negative emotions. Greater positive skew was observed in NA and HA distributions. *MSSD* statistics had positive skew across the included affect scales (>1.24). Median of the average item intercorrelation of affect scales (i.e. \bar{r}) ranged from .41 to .68 (except for HA \bar{r} that had median of 1.00 in Sample 2), indicating that the respective indicators of affect scales were substantially intercorrelated for the average individual (and perfectly intercorrelated for the average individual in terms of HA in Sample 2).⁵

Median of individual autoregressive parameters was ranging from .12 (small) to .32 (large), indicating that affects carried over to the next day/moment for the average individual. IQR of individual autoregressive parameters was small to moderate in daily data (IQR = .04-.17) and small to large in momentary data (IQR = .13-.27). Median of individual cross-lagged parameters was small (Mdn = -.02-.02), except for ϕ NH, for which small-to-moderate median was observed in Sample 2 (Mdn = .12) and Sample 3 (Mdn = .14). Individual cross-lagged effects had small IQR, IQR = .01-.08. Median of individual contemporaneous

⁵Note that the HA scale comprised only two emotion adjectives (i.e. hostile, upset) in Sample 2.

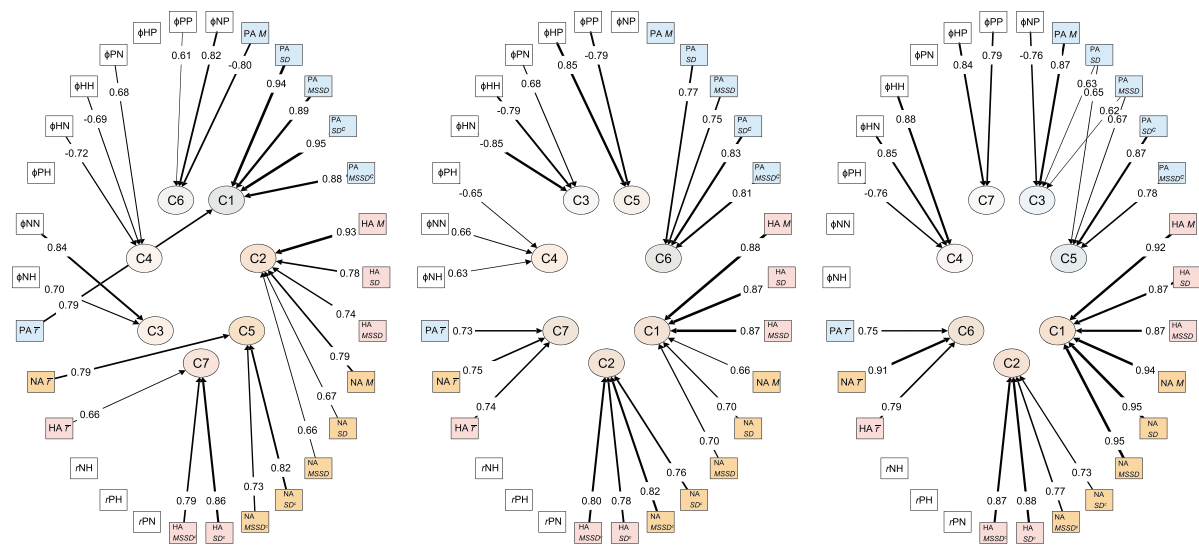


FIGURE 1. Varimax rotated components of IADs across samples. Component loadings < .60 are suppressed. PA, positive affect; NA, negative affect; HA, hostile affect; *M*, mean; *SD*, standard deviation; *MSSD*, mean square successive differences; *SD^c*, corrected standard deviation; *MSSD^c*, corrected mean square successive differences; \bar{r} , average item intercorrelation of affect scales; ϕ , autoregressions and cross-lagged effects; *r*, contemporaneous partial correlations. [Colour figure can be viewed at wileyonlinelibrary.com]

partial correlation between NA and HA was large (Mdn = .41–.62), and IQR was moderate to large, IQR = .20–.27. Median of individual contemporaneous partial correlation between PA and NA was negative in Sample 1 (Mdn = -.43, IQR = .18) and close to zero in Sample 2 (Mdn = .04, IQR = .14) and Sample 3 (Mdn = -.02, IQR = .12). Median of individual contemporaneous partial correlation between PA and HA was close to zero (Mdn = -.05–.00) and IQR was small, IQR = .06–.13. Similar distributions of individual random effects were obtained for odd and even days when compared against the networks that were calculated using the complete data (e.g. *r*NH in Sample 2, complete data: .47, odd days: .47, even days: .45; ϕ HH, complete data: .13, odd days: .13, even days: .17).

Structure

Parallel analysis indicated seven components in Sample 2 and nine components in Samples 1 and 3. Visual inspection of scree plots was inconclusive because there were no clear drops in eigenvalues. We base our interpretation of the structure of IADs on the seven-component resolution as it represented the greatest common denominator across samples. Figure 1 provides a sparse graphical display of the varimax rotated principal components. Commonalities were found between univariate IADs of PA (Sample 1, C1; Sample 2, C6; Sample 3, C3, C5), indicating that those tended to be interrelated. Univariate IADs of NA and HA were reflected in several principal components (Sample 1, C2, C5, C7; Sample 2, C1, C2; Sample 3, C1, C2). Temporal effects including autoregressive and cross-lagged parameters formed additional principal components (Sample 1, C3, C4, C6; Sample 2, C3, C4, C5; Sample 3, C4, C7). The average item intercorrelation of affect scales formed principal components in Sample 2 (C7) and Sample 3 (C6).

Figure 2a–c shows variance decompositions of IADs by sample, indicating the particular degree to which variation in IADs was accounted for by scale means and *SD*s.⁶ Generally, scale means and *SD*s shared plenty of common variance with the more complex IADs, including autoregressions, cross-lagged effects, and contemporaneous partial correlations, thus, highlighting the need to consider these redundancies for extracting their more unique information.

Split-half reliability

Figure 2b–c displays the estimated variance of IADs that was unique and reliable in Samples 2 and 3 as fractions of the total variance observed. These estimates align with the split-half reliability of residualized IADs (see Figures S1 and S2). Split-half reliability of raw IADs is reported in Table S5. Individual *M* of affect scales was highly reliable, *r*_{sb} = .94–.99. Residualized *SD* and *SD^c* had moderate-to-high reliability (*SD*, *r*_{sb} = .77–.96; *SD^c*, *r*_{sb} = .67–.95), indicating that those tend to reliably capture incremental features of emotional time series beyond mean of affect. Partialing out scale means and *SD*s tended to reduce the split-half reliability of more complex IADs, indicating that their reliability estimates were inflated because of redundancies with *M* and *SD*. For example, the split-half reliability of PA *MSSD* decreased from .92 to .69 in Sample 2 and the split-half reliability of ϕ NP decreased from .75 to .06 in Sample 3. Nevertheless, some residualized IADs achieved moderate-

⁶The total variance was decomposed into the part of variance explained by scale means (i.e. the squared multiple correlation with scale means), the part of variance explained by scale standard deviations beyond what had already been explained by scale means (i.e. the squared multiple correlation with scale means and scale standard deviations minus the squared multiple correlation with scale means), the unique variance that was reliable (i.e. split-half reliability of the residualized variable), and the unique variance that was not reliable (one minus the sum of the aforementioned variance parts).

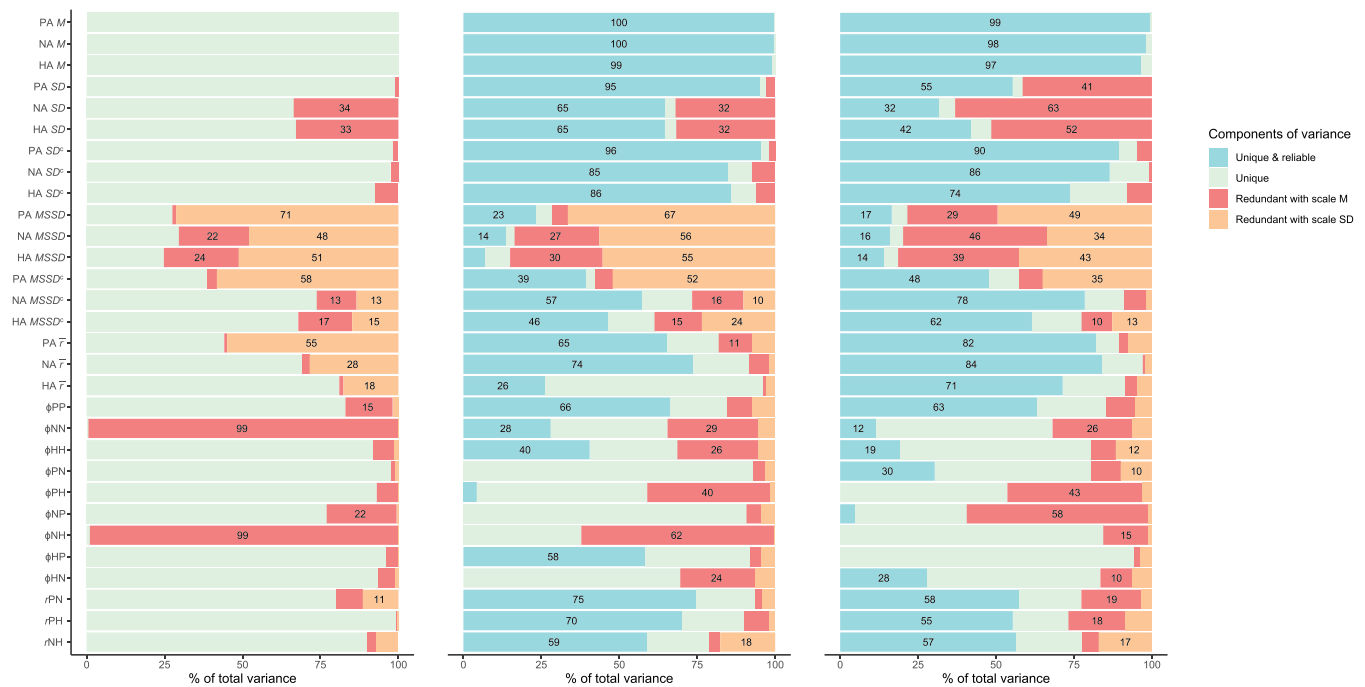


Figure 2. (a–c) Variance decomposition of IADs. From left to right Samples 1–3 are depicted. PA, positive affect; NA, negative affect; HA, hostile affect; *M*, mean; *SD*, standard deviation; *MSSD*, mean square successive differences; *SD*^c, corrected standard deviation; *MSSD*^c, corrected mean square successive differences; \bar{r} , average item intercorrelation of affect scales; ϕ , autoregressions and cross-lagged effects; *r*, contemporaneous partial correlations.

to-high reliability (e.g. *MSSD* and *MSSD*^c, $r_{sb} = .61-.88$). The reliability of residualized HA *MSSD* was low in Sample 2, $r_{sb} = .31$. The residualized average item intercorrelations of affect scales had moderate-to-high reliability, $r_{sb} = .64-.85$, except for HA \bar{r} in Sample 2, $r_{sb} = .16$. The reliability of residualized autoregression was moderate in case of ϕ_{PP} ($r_{sb} = .59-.65$) and low in case of ϕ_{HH} (Sample 2, $r_{sb} = .42$; Sample 3, $r_{sb} = .14$). The residualized autoregression of ϕ_{NN} had reliability that was close to zero, $r_{sb} = .09-.27$. Residualized cross-lagged effects also tended towards reliability estimates that were close to zero, $r_{sb} = .00-.13$, except for ϕ_{HP} in Sample 2, $r_{sb} = .47$. The reliability of residualized contemporaneous partial correlations was moderate, $r_{sb} = .57-.66$.

Associations with personality traits

Figure 3 displays meta-analytic estimates for the bivariate associations between residualized IADs and personality traits. Tables S6–S7 display sample-specific results for raw and residualized IADs. More desirable mean level of affect (i.e. high PA, low NA, and HA) were associated with extraversion (PA *M*, $r = .32$; NA *M*, $r = -.24$; HA *M*, $r = -.14$) conscientiousness (PA *M*, $r = .23$; NA *M*, $r = -.20$; HA *M*, $r = -.20$), agreeableness (PA *M*, $r = .11$; NA *M*, $r = -.16$; HA *M*, $r = -.26$), and openness (PA *M*, $r = .16$). Less desirable mean level of affect were associated with neuroticism (PA *M*, $r = -.26$; NA *M*, $r = .39$; HA *M*, $r = .30$). When controlling for mean level of affects, greater PA variability and

NA variability were incrementally associated with openness (PA *SD*, $r = .15$; PA *SD*^c, $r = .13$; NA *SD*, $r = .06$; NA *SD*^c, $r = .06$) and extraversion (PA *SD*, $r = .15$; PA *SD*^c, $r = .14$; NA *SD*^c, $r = .08$). Lower PA variability was incrementally associated with agreeableness (PA *SD*^c, $r = -.07$). The residualized statistics of NA and HA variability (*SD*), but not *SD*^c, were incrementally associated with neuroticism (NA *SD*, $r = .17$; HA *SD*, $r = .15$).

The bivariate correlations between more complex IADs and personality traits were of smaller size when residual variables were used, indicating that non-specific associations were induced by redundancies with *M* and *SD*. For example, the correlation between NA *SD* and neuroticism decreased from .22 to $-.04$ (Sample 1), and the correlation between ϕ_{NN} and neuroticism decreased from .16 to $-.02$ (Sample 1). Notwithstanding, some incremental associations between more complex IADs and personality traits were found that reached statistical significance ($p < .05$). After controlling for scale means and *SD*s, lower NA instability and HA instability were incrementally associated with neuroticism (NA *MSSD*^c, $r = -.10$; HA *MSSD*^c, $r = -.09$). A less differentiated reporting of hostile states was incrementally associated with agreeableness (HA \bar{r} , $r = .09$) and a more differentiated reporting of negative emotional states was incrementally associated with neuroticism (NA \bar{r} , $r = -.07$). The contemporaneous partial correlation between NA and HA was incrementally associated with agreeableness (r_{NH} , $r = .08$), indicating that highly agreeable individuals exhibit a greater than

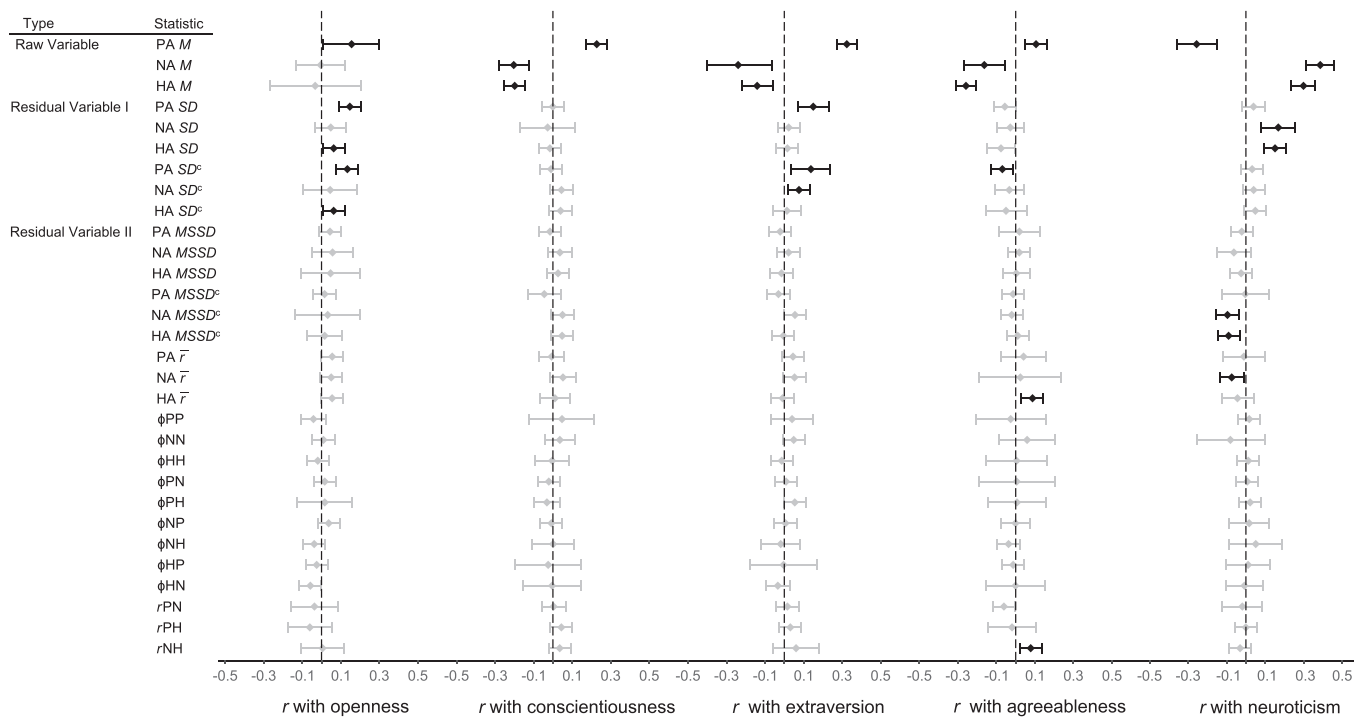


FIGURE 3. Meta-analytic estimates of the bivariate correlations between residualized IADs and big five traits. Residual variable type I were residualized for scale means. Residual variables type II were residualized for scale means and standard deviations. r , meta-analytic estimate of the bivariate correlation; PA, positive affect; NA, negative affect; HA, hostile affect; M , mean; SD , standard deviation; $MSSD$, mean square successive differences; SD^c , corrected standard deviation; $MSSD^c$, corrected mean square successive differences; \bar{r} , average item intercorrelation of affect scales; ϕ , autoregressions and cross-lagged effects; r , contemporaneous partial correlations.

average tendency to experience negative and hostile states in concordance.⁷ Temporal parameters (i.e. autoregressive and cross-lagged effects) had no significant incremental associations with personality traits.

DISCUSSION

Failed tests of incremental validity of more complex IADs raised doubt about their validity and usefulness for studying affect dynamics (Dejonckheere *et al.*, 2019). The current study extends the body of evidence by providing a comprehensive analysis of their structure, split-half reliability, and association with personality traits. In line with earlier results, more complex IADs exhibited substantial redundancies with mean level and general variability of emotions. When these redundancies were controlled statistically, the reliability and validity coefficients of more complex IADs shrunk, and in some cases, they became small or non-significant, indicating that many of the more complex IADs convey limited incremental information on affect dynamics using the current assessment practices.

Structure of indicators of affect dynamics

The covariance structures of IADs elucidated similar patterns of redundancy in the included samples, showing that many of the more complex IADs exhibit marked redundancies with mean level and general variability of emotional time series.

⁷Remember that the contemporaneous partial correlation r_{NH} of the average individual was large, .41–.61.

Our analyses indicate that additional redundancies may exist beyond of that, as was exemplified by principal components that summarized additional common variance between IADs related to emotion differentiation (i.e. \bar{r}) and temporal effects (e.g. ϕ_{PP}).

We discuss possible reasons for the observed redundancies. The statistical overlap between individual M and SD of affect scales was of higher magnitude when the underlying emotion distributions exhibited greater skew, which was the case for negative emotions, hostile emotions, and when affects were measured using momentary assessment. This points to the fact that M and SD are mathematically interdependent in skewed distributions. The redundancies found between $MSSD$ and SD indicate that their raw forms capture similar constructs at the between-person level, that is, general variability. This is not surprising considering that both SD and $MSSD$ may serve as global indices of dispersion (Jahng *et al.*, 2008). However, the residualized form of $MSSD$ and $MSSD^c$ that were used in the current study should have a different interpretation, because their statistical redundancies with M and SD were partialled out (i.e. 10–85% of variance). Thus, the residualized $MSSD$ should mainly reflect the temporal dependency of consecutive measurement occasions, similar to the autoregressive parameter. Thereby, the residualized $MSSD^c$ may delineate a continuum that ranges from emotional instability (i.e. high values) to emotional inertia (i.e. low values).

We have no explanation for the divergent pattern of redundancy with respect to the average item intercorrelation of affect scales (i.e. $PA \bar{r}$, $NA \bar{r}$, $HA \bar{r}$). On the one hand, in Sample 1, greater average item intercorrelations were associated with

greater emotional variability on the respective scales. This might suggest that individuals that respond homogeneously to items of one affect scale (i.e. internal consistency) could have an increased probability of producing true score variability. On the other hand, in Samples 2 and 3, the average item inter-correlation of affect scales formed a component of their own, indicating a general form of emotion differentiation.

Network parameters (i.e. autoregressions, cross-lagged effects, and contemporaneous partial correlations) exhibited redundancies primarily with M , some of those were extremely large (i.e. ϕ_{NN} and ϕ_{NH} in Sample 1). This may be unintuitive, as the networks were estimated using within-person centered variables. However, within-person centering does not change the variance and shape of individual state affect distributions, and, as noted above, many of the affect scales showed little variance (i.e. small IQR) and were skewed, especially the negative items. Thus, the well-known associations between the mean and variability and the role of variance restriction serve as the likely explanation for the observed redundancies between network random effects and individual M (including redundancies of ϕ_{NN} and ϕ_{NH} in Sample 1). Indeed, this has been an issue that has plagued the psychological network literature, which has often found that many of the most conceptually interesting statistics are highly dependent on observed variances in real world data (Rodebaugh *et al.*, 2018).

Split-half reliability of indicators of affect dynamics

In line with prior research, mean and variability of affects had high reliability indicating that those consistently measure the same constructs at the between-person level. Some of the more complex IADs were moderately reliable (e.g. $MSSD^c$, \bar{r} , ϕ_{PP} , contemporaneous partial correlations) after controlling for their overlap with means and SD s. Those IADs may reflect true and unique individual variation; however, they also include substantial measurement error. The extent to which IADs are unreliable puts a ceiling on the strength of their associations with other constructs that can possibly be observed. In consequence, such associations will be attenuated and will require larger sample sizes for detecting signals.

Temporal effects (i.e. autoregressions and cross-lagged effects) tended towards having very poor or no reliability (except for ϕ_{PP}). This suggests that those are not valid IADs for between-person research using the current assessment practices. Considering that most temporal effects yielded little random effect variances, it seems quite plausible that, in some cases, sampling variation and statistical redundancies may occasionally account for the total of their variance. One reason for the unreliability of some IADs could be that the indicators used here (i.e. the selected pool of emotions items) are not optimal or that measurement occasions were too few. Another reason could be that some IADs reflect more volatile psychological conditions that fluctuate rapidly (e.g. because of contextual factors; e.g. Koval & Kuppens, 2012). This would suggest that their assessment should be based on different assessment frames (e.g. more frequent assessments within a smaller time frame) or that such

IADs should be better studied under more controlled conditions (e.g. experimental designs; Dejonckheere, Mestdagh, Kuppens, & Tuerlinckx, 2020).

Associations between indicators of affect dynamics and big five personality traits

Our findings indicate that big five personality traits are characterized by distinct pattern of affect dynamics that primarily encompass individual differences in M s and SD s of PA, NA, and HA. Mean affect had correlations with personality traits that were in line with prior research (e.g. Ching *et al.*, 2014; Howell, Ksendzova, Nestingen, Yerahian, & Iyer, 2017; Watson & Clark, 1999), indicating that more adaptive configurations of personality traits (i.e. emotional stability, openness, conscientiousness, agreeableness, and extraversion) are robustly related to more desirable emotional experiences (i.e. high PA, low NA, low HA).

Positive affect variability was positively associated with extraversion and openness. Given that greater variability of positive emotions was a substantive characteristic of extraverted and open individuals, one could speculate that it reflected the exploratory nature of these traits, in other words, a greater tendency to seek potentially rewarding situations (i.e. greater sensitivity to rewards; DeYoung, 2015). This 'high risk high reward' strategy could result in greater variability in the achievement of rewards, and thus, in greater variability of experiencing positive emotions. In contrast, individuals high in agreeableness had less variability but higher mean level of PA. This is notable, because it indicates that personality traits are differentially associated with mean affect and variability, and thus, it provides evidence for their divergent nomological patterns. A competing account of variability measures argues that those might reflect extreme responding (Baird, Lucas, & Donnellan, 2017), which is the tendency to choose more extreme response categories in self-report questionnaires. However, extreme responding has been mainly associated with high extraversion and high conscientiousness (Austin, Deary, & Egan, 2006), indicating that the current results might not be sufficiently explained by this type of method bias. In the current study, neuroticism did not show consistent incremental associations with greater negative emotional variability, because although SD reached statistical significance, SD^c did not. This corroborates findings from a recent meta-analysis (Kalokerinos *et al.*, 2020). On the one hand, the relative indices used here (i.e. SD^c , $MSSD^c$) may be superior for deriving the more unique information about emotional variability and emotional instability in the presence of non-linear dependencies (Mestdagh *et al.*, 2018). On the other hand, it may be, though, that they overcorrect in skewed distributions thereby reducing their validity, and other methods for accounting for the association between mean and variability should be considered in future work.

Some incremental associations were found between more complex IADs and personality traits. Lower instability of negative and hostile emotional states (i.e. NA $MSSD^c$, HA $MSSD^c$) was observed in individuals high in neuroticism. With respect to our interpretation of the residualized $MSSD^c$

that was discussed earlier, this finding suggests that individuals high in neuroticism might be more resistant to change in negative emotional states; thus, adding to the body of conflicting results on how fluctuations in negative emotions are linked to neuroticism-related constructs (also known as the ‘instability-inertia paradox’, e.g. Bos *et al.*, 2019; Bosley, Soyster, & Fisher, 2019; Koval, Kuppens, Allen, & Sheeber, 2012; Koval *et al.*, 2013). Furthermore, a more differentiated reporting of negative emotions (i.e. NA \bar{r}) was observed in neurotic individuals, and a more differentiated reporting of hostile states (i.e. HA \bar{r}) was observed in disagreeable individuals. One hypothesis could be that emotion differentiation is related to attentional processes, such that, individuals high in neuroticism pay greater attention to their negative emotions and individuals high in disagreeableness pay greater attention to their hostile emotions, and thus, they might experience and report specific affects more nuanced. Finally, a greater association between negative and hostile states (i.e. r_{NH}) was observed in agreeable individuals. One explanation could be that agreeable individuals have a larger aversion of interpersonal conflicts, such that hostility caused more negativity (Suls, Martin, & David, 1998). However, we want to emphasize that the associations between more complex IADs and personality traits were small in magnitude, and as we did not adjust for multiple testing, any of the suggested interpretations need to be taken cautiously.

Limitations

The current study has some important limitations with respect to the samples, the measurement of emotions, the intensive longitudinal designs, the methods for assessing personality traits, and the statistical analyses. First, our study might have compared ‘apples and oranges’ by drawing inferences across samples that were diverse in terms of sampling frames (i.e. daily and momentary emotion data), instruments used, and populations investigated. We regard this limitation a strength, as the heterogeneity across samples contributes to greater generalizability (Yarkoni, 2019). Second, the pool of emotions used might not completely span the affective space. Third, with regard to the involved intensive longitudinal designs, Sample 1 might contain too few measurement occasions for deriving a reliable assessment of network parameters (i.e. 20–30 consecutive measurements per person in Sample 1). Moreover, measurement burst designs (Stawski, MacDonald, & Sliwinski, 2015) can inform researchers about the trait status of IADs, by investigating their stability over more widely spaced temporal intervals, because stability is usually demonstrated over longer periods of time (i.e. 1–2 years) than were used in the current study. Also, IADs might exhibit stronger incremental validity when studied under more controlled contextualized conditions (e.g. experimental designs or event-contingent assessment) because of potentially preferable signal-to-noise ratio (Dejonckheere *et al.*, 2020; Lapate & Heller, 2020). Fourth, we only had access to concurrent self-reported personality data. Some of the dynamic indices could be stronger related to personality facets or nuances, future personality or

personality change, informant reports (Finnigan & Vazire, 2018), or digital footprints of personality (Hinds & Joinson, 2019).

Fifth, our study did not model measurement error. Generally, unmodelled measurement error might have attenuated the reliability and validity estimates of IADs (Rouder & Haaf, 2019; Schuurman & Hamaker, 2019). Error may have been introduced when calculating IADs based on daily/momentary affect scores that may not be perfectly reliable measures of state affect. Measurement error might have led to a downward bias of network parameters in particular (Schuurman, Houtveen, & Hamaker, 2015). Error variance may further accumulate when calculating residualized IADs based on affect means and standard variability that may not be perfectly reliable measures of trait affect and emotional variability. Thereby, validity estimates may be biased when residualized IADs are insufficiently cleared of their redundancies with mean level and general variability (Westfall & Yarkoni, 2016), but this might also occur in the presence of non-linear dependencies (Mestdagh *et al.*, 2018). Further limitations of our statistical approach were that we did not consider alternative approaches that differ in their assumptions or estimation techniques (e.g. dynamic structural equation modelling, Asparouhov, Hamaker, & Muthén, 2018; Geukes *et al.*, 2017; Kuppens, Oravecz, & Tuerlinckx, 2010; Loossens *et al.*, 2019) and that we did not model the dyadic data structure in Sample 3 for calculating IADs.

CONCLUSIONS

The current study highlights that observed links between IADs and other constructs might be non-specific, as arising from redundancies between them. More specifically, in research settings in which affect dynamics are linked to between-person constructs, researchers should rule out more parsimonious explanations (e.g. trait affect, affect variability) before attributing incremental value to more complex IADs. Occasionally, researchers have strived for trait interpretations of IADs—explicitly or implicitly (i.e. by investigating or theorizing on their associations with relatively stable traits as was done in the current study). Our results demonstrate that there is scarce evidence for trait interpretations of many of the more complex IADs with respect to their low reliability and unknown stability.

More generally, researchers should consider the limited validity of IADs for research questions at the between-person level, as their reliability might be low, and any true effects might be obscured or attenuated. Notwithstanding, some of those more complex IADs had unique and somewhat reliable variance, including IADs with respect to emotion differentiation (e.g. \bar{r} , contemporaneous partial correlations) and emotional instability (e.g. $MSSD^c$). For those, small but incremental associations with personality traits were found. These results point out to the possibility that such IADs might contain substantive between-person variance that may be of interest to researchers for studying individual differences, albeit they may not always meet standard psychometric criteria (Wright & Zimmermann, 2019). The current assessment practices might need refinement in order to further improve the

validity of more complex IADs, for example, by increasing the frequency or duration of emotion assessment. However, it is questionable whether a more intensive or longer enduring assessment would be practically feasible, as longer time frames may increase burden of participation and lead to higher non-compliance rates (Eisele *et al.*, 2020). More research is needed for identifying the conditions under which affect dynamics can be assessed most validly.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Fit Statistics of Multi-Level Exploratory Factor Models (Geomin Rotation) by Sample

Table S2. Geomin Rotated Factor Loadings by ML-EFA in Sample 1

Table S3. Geomin Rotated Factor Loadings by ML-EFA in Sample 2

Table S4. Geomin Rotated Factor Loadings by ML-EFA in Sample 3

Table S5. Split-half Reliability of Raw IADs in Sample 2 and Sample 3

Table S6. Bivariate Correlations between Raw IADs and Big Five Traits by Sample

Table S7. Bivariate Correlations between Residualized IADs and Big Five Traits by Sample

Figure S1. Split-half Reliability of IADs in Daily Data (Sample 2).

Figure S2. Split-half Reliability of IADs in Momentary Data (Sample 3)

REFERENCES

- Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 359–388. <https://doi.org/10.1080/10705511.2017.1406803>
- Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, *40*, 1235–1245. <https://doi.org/10.1016/j.paid.2005.10.018>
- Baird, B. M., Lucas, R. E., & Donnellan, M. B. (2017). The role of response styles in the assessment of intraindividual personality variability. *Journal of Research in Personality*, *69*, 170–179. <https://doi.org/10.1016/j.jrp.2016.06.015>
- Bos, E. H., de Jonge, P., & Cox, R. F. (2019). Affective variability in depression: Revisiting the inertia–instability paradox. *British Journal of Psychology*, *110*, 814–827. <https://doi.org/10.1111/bjop.12372>
- Bosley, H. G., Soyster, P. D., & Fisher, A. J. (2019). Affect dynamics as predictors of symptom severity and treatment response in mood and anxiety disorders: Evidence for specificity. *Journal for Person-Oriented Research*, *5*, 101–113. <https://doi.org/10.17505/jpor.2019.09>
- Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., ... Kuppens, P. (2016). Assessing temporal emotion dynamics using networks. *Assessment*, *23*, 425–435. <https://doi.org/10.1177/1073191116645909>
- Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE*, *8*, e60188. <https://doi.org/10.1371/journal.pone.0060188>
- Carver, C. S., Sutton, S. K., & Scheier, M. F. (2000). Action, emotion, and personality: Emerging conceptual integration. *Personality and Social Psychology Bulletin*, *26*, 741–751. <https://doi.org/10.1177/0146167200268008>
- Ching, C. M., Church, A. T., Katigbak, M. S., Reyes, J. A. S., Takanaka-Matsumi, J., Takaoka, S., ... Ortiz, F. A. (2014). The manifestation of traits in everyday behavior and affect: A five-culture study. *Journal of Research in Personality*, *48*, 1–16. <https://doi.org/10.1016/j.jrp.2013.10.002>
- Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PIR) and NEO Five Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.
- Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease*, *175*, 526–536. <https://doi.org/10.1097/00005053-198709000-00004>
- Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., ... Kuppens, P. (2018). The bipolarity of affect and depressive symptoms. *Journal of Personality and Social Psychology*, *114*, 323–341. <https://doi.org/10.1037/pspp0000186>
- Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, *3*, 478–491. <https://doi.org/10.1038/s41562-019-0555-0>
- Dejonckheere, E., Mestdagh, M., Kuppens, P., & Tuerlinckx, F. (2020). Reply to: Context matters for affective chronometry. *Nature Human Behaviour*. Advance Online Publication, 1–4. <https://doi.org/10.1038/s41562-020-0861-6>
- Denissen, J. J. A., & Kühnel, A. (2008). *Handbook for the use of data from the diary study at Humboldt Universität zu Berlin*. Retrieved from https://www.psychologie.hu-berlin.de/de/prof/perdev/downloadentwper/diarystudy/Handbook_Diary.pdf
- DeYoung, C. G. (2015). Cybernetic big five theory. *Journal of Research in Personality*, *56*, 33–58. <https://doi.org/10.1016/j.jrp.2014.07.004>
- Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology*, *76*, 662–676. <https://doi.org/10.1037/0022-3514.76.4.662>
- Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020, February 20). *The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population*. <https://doi.org/10.31234/osf.io/zf4nm>
- Epskamp, S., Deserno, M. K., & Bringmann, L. F. (2019). *mlVAR: Multi-level vector autoregression*. R package version 0.4.3. <https://CRAN.R-project.org/package=mlVAR>
- Epskamp, S., Waldorp, L. J., Möttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*, 453–480. <https://doi.org/10.1080/00273171.2018.1454823>
- Finnigan, K. M., & Vazire, S. (2018). The incremental validity of average state self-reports over global self-reports of personality. *Journal of Personality and Social Psychology*, *115*, 321–337. <https://doi.org/10.1037/pspp0000136>
- Geukes, K., Nestler, S., Hutteman, R., Dufner, M., Küfner, A. C., Egloff, B., Denissen, J. J. A., et al. (2017). Puffed-up but shaky selves: State self-esteem level and variability in narcissists. *Journal of Personality and Social Psychology*, *112*, 769–786. <https://doi.org/10.1037/pspp0000093>
- Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. <https://doi.org/10.1016/j.paid.2016.06.069>

- Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, 26, 10–15. <https://doi.org/10.1177/0963721416666518>
- Hinds, J., & Joinson, A. (2019). Human and computer personality prediction from digital footprints. *Current Directions in Psychological Science*, 28, 204–211. <https://doi.org/10.1177/0963721419827849>
- Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, 141, 901–930. <https://doi.org/10.1037/a0038822>
- Howell, R. T., Ksendzova, M., Nestingen, E., Yerahian, C., & Iyer, R. (2017). Your personality on a good day: How trait and state personality predict daily well-being. *Journal of Research in Personality*, 69, 250–263. <https://doi.org/10.1016/j.jrp.2016.08.001>
- Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological Methods*, 13, 354–375. <https://doi.org/10.1037/a0014173>
- Kalokerinos, E., Murphy, S. C., Koval, P., Bailen, N. H., Crombez, G., Hollenstein, T., Gleeson, J., ... Bastian, B. (2020). Neuroticism may not reflect emotional variability. *Proceedings of the National Academy of Sciences of the United States of America*, 117, 9270–9276. <https://doi.org/10.1073/pnas.1919934117>
- Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multi-level factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research*, 51, 881–898. <https://doi.org/10.1080/00273171.2016.1228042>
- Koval, P., & Kuppens, P. (2012). Changing emotion dynamics: Individual differences in the effect of anticipatory social stress on emotional inertia. *Emotion*, 12, 256–267. <https://doi.org/10.1037/a0024756>
- Koval, P., Kuppens, P., Allen, N. B., & Sheeber, L. (2012). Getting stuck in depression: The roles of rumination and emotional inertia. *Cognition & Emotion*, 26, 1412–1427. <https://doi.org/10.1080/02699931.2012.667392>
- Koval, P., Pe, M. L., Meers, K., & Kuppens, P. (2013). Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion*, 13, 1132–1141. <https://doi.org/10.1037/a0033579>
- Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, 21, 984–991. <https://doi.org/10.1177/0956797610372634>
- Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, 99, 1042–1060. <https://doi.org/10.1037/a0020962>
- Kuppens, P., Sheeber, L. B., Yap, M. B., Whittle, S., Simmons, J. G., & Allen, N. B. (2012). Emotional inertia prospectively predicts the onset of depressive disorder in adolescence. *Emotion*, 12, 283–289. <https://doi.org/10.1037/a0025046>
- Kuppens, P., Van Mechelen, I., Nezlek, J. B., Dossche, D., & Timmermans, T. (2007). Individual differences in core affect variability and their relationship to personality and psychological adjustment. *Emotion*, 7, 262–274. <https://doi.org/10.1037/1528-3542.7.2.262>
- Lang, F. R., Lüdtke, O., & Asendorpf, J. B. (2001). Testgüte und psychometrische Äquivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittelalten und alten Erwachsenen. *Diagnostica*, 47, 111–121.
- Lapate, R. C., & Heller, A. S. (2020). Context matters for affective chronometry. *Nature Human Behaviour: Online advance publication*, 1–2. <https://doi.org/10.1038/s41562-020-0860-7>
- Loossens, T., Mestdagh, M., Dejonckheere, E., Kuppens, P., Tuerlinckx, F., & Verdonck, S. (2019, September 5). *The affective ising model: A computational account of human affect dynamics*. <https://doi.org/10.31234/osf.io/ky23d>
- Mejía, S., Hooker, K., Ram, N., Pham, T., & Metoyer, R. (2014). Capturing intraindividual variation and covariation constructs: Using multiple time-scales to assess construct reliability and construct stability. *Research in Human Development*, 11, 91–107. <https://doi.org/10.1080/15427609.2014.906728>
- Mestdagh, M., Pe, M., Pestman, W., Verdonck, S., Kuppens, P., & Tuerlinckx, F. (2018). Sidelineing the mean: The relative variability index as a generic mean-corrected variability measure for bounded variables. *Psychological Methods*, 23, 690–707. <https://doi.org/10.1037/met0000153>
- Mneimne, M., Fleeson, W., Arnold, E. M., & Furr, R. M. (2018). Differentiating the everyday emotion dynamics of borderline personality disorder from major depressive disorder and bipolar disorder. *Personality Disorders, Theory, Research, and Treatment*, 9, 192–196. <https://doi.org/10.1037/per0000255>
- Muthén, L. K., & Muthén, B. O. (2018). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for statistical computing, Vienna, Austria. URL <https://www.R-project.org/>
- Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment*, 84, 126–136. https://doi.org/10.1207/s15327752jpa8402_02
- Revelle, W. (2018) *psych: Procedures for personality and psychological research*, Northwestern University, Evanston, Illinois, USA, <https://CRAN.R-project.org/package=psychVersion=1.8.12>
- Rodebaugh, T. L., Tonge, N. A., Piccirillo, M. L., Fried, E. I., Horenstein, A., Morrison, A. S., Goldin, P., ... Heimberg, R. G. (2018). Does centrality in a cross-sectional network suggest intervention targets for social anxiety disorder? *Journal of Consulting and Clinical Psychology*, 86, 831–844. <https://doi.org/10.1037/ccp0000336>
- Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, 26, 452–467. <https://doi.org/10.3758/s13423-018-1558-y>
- Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods*, 24, 70–91. <https://doi.org/10.1037/met0000188>
- Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n= 1 psychological autoregressive modeling. *Frontiers in Psychology*, 6, 1038. <https://doi.org/10.3389/fpsyg.2015.01038>
- Scott, L. N., Victor, S. E., Kaufman, E. A., Beeney, J. E., Byrd, A. L., Vine, V., ... Stepp, S. D. (2020). Affective dynamics across internalizing and externalizing dimensions of psychopathology. *Clinical Psychological Science*. Advance online publication, 8, 412–427. <https://doi.org/10.1177/2167702619898802>
- Stawski, R. S., MacDonald, S. W., & Sliwinski, M. J. (2015). Measurement burst design. *The Encyclopedia of Adulthood and Aging*, 1–5.
- Suls, J., Martin, R., & David, J. P. (1998). Person-environment fit and its limits: Agreeableness, neuroticism, and emotional reactivity to interpersonal conflict. *Personality and Social Psychology Bulletin*, 24, 88–98. <https://doi.org/10.1177/0146167298241007>
- Trull, T. J., Lane, S. P., Koval, P., & Ebner-Priemer, U. W. (2015). Affective dynamics in psychopathology. *Emotion Review*, 7, 355–361. <https://doi.org/10.1177/1754073915590617>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 1–48. URL <http://www.jstatsoft.org/v36/i03/>
- Watson, D., & Clark, L. A. (1999). *The PANAS-X: Manual for the positive and negative affect schedule-expanded form*.
- Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, 98, 219–235. <https://doi.org/10.1037//0033-2909.98.2.219>
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, 11, e0152719. <https://doi.org/10.1371/journal.pone.0152719>

- Wright, A. G. C., Beltz, A. M., Gates, K. M., Molenaar, P. C. M., & Simms, L. J. (2015). Examining the dynamic structure of daily internalizing and externalizing behavior at multiple levels of analysis. *Frontiers in Psychology, 6*, 1914. <https://doi.org/10.3389/fpsyg.2015.01914>
- Wright, A. G. C., Gates, K. M., Arizmendi, C., Lane, S. T., Woods, W. C., & Edershile, E. A. (2019). Focusing personality assessment on the person: Modeling general, shared, and person specific processes in personality and psychopathology. *Psychological Assessment, 32*, 502–515. <https://osf.io/nf5me/>
- Wright, A. G. C., Stepp, S. D., Scott, L. N., Hallquist, M. N., Beeney, J. E., Lazarus, S. A., & Pilkonis, P. A. (2017). The effect of pathological narcissism on interpersonal and affective processes in social interactions. *Journal of Abnormal Psychology, 126*, 898–910. <https://doi.org/10.1037/abn0000286>
- Wright, A. G. C., & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological Assessment, 31*, 1467–1480. <https://doi.org/10.1037/pas0000685>
- Yarkoni, T. (2019, November 22). *The Generalizability Crisis*. <https://doi.org/10.31234/osf.io/jqw35>

Article 2:

Mapping established psychopathology scales onto the Hierarchical Taxonomy of Psychopathology (HiTOP)

Leon P. Wendt¹, Kristin Jankowsky¹, Ulrich Schroeders¹, London Personality and Mood Disorder Research Consortium, Tobias Nolte^{2,3}, Peter Fonagy^{3,4}, P. Read Montague^{2,5}, Johannes Zimmermann^{1*} & Gabriel Olaru^{6*}

¹ Department of Psychology, University of Kassel

² Wellcome Trust Centre for Neuroimaging, University College London

³ Anna Freud National Centre for Children and Families

⁴ Research Department of Clinical, Educational and Health Psychology, University College London,

⁵ Fralin Biomedical Research Institute, Department of Psychology, Virginia Tech

⁶ Department of Developmental Psychology, Tilburg University

* Shared last authorship

Status:

Published

Citation:








Wendt, L. P., Jankowsky, K., Schroeders, U., London Personality and Mood Disorder Research Consortium, Nolte, T., Fonagy, P., Montague, P. R., Zimmermann J., & Olaru, G. (2023). Mapping established psychopathology scales onto the Hierarchical Taxonomy of Psychopathology (HiTOP). *Personality and Mental Health*, 17(2), 117-134. <https://doi.org/10.1002/pmh.1566>

Supplemental Materials:

R code, and other materials

<https://doi.org/10.17605/OSF.IO/HKAV3>

Mapping established psychopathology scales onto the Hierarchical Taxonomy of Psychopathology (HiTOP)

Leon P. Wendt¹  | Kristin Jankowsky¹  | Ulrich Schroeders¹  |
 London Personality and Mood Disorder Research Consortium | Tobias Nolte^{2,3}  |
 Peter Fonagy^{3,4}  | P. Read Montague^{2,5} | Johannes Zimmermann¹  |
 Gabriel Olaru⁶ 

¹Department of Psychology, University of Kassel, Kassel, Germany

²Wellcome Trust Centre for Neuroimaging, University College London, London, UK

³Anna Freud National Centre for Children and Families, London, UK

⁴Research Department of Clinical, Educational and Health Psychology, University College London, London, UK

⁵Fralin Biomedical Research Institute, Department of Psychology, Virginia Tech, Roanoke, Virginia, USA

⁶Department of Developmental Psychology, Tilburg University, Tilburg, Netherlands

Correspondence

Leon P. Wendt, Department of Psychology, University of Kassel, Holländische Str. 36-38, 34127 Kassel, Germany.
 Email: l.wendt@uni-kassel.de

Funding information

This work was supported by a National Institute for Health Research (NIHR) Senior Investigator Award (NF-SI-0514-10157) awarded to Peter Fonagy. The work was also supported by NIH-NIDS Grant 5R01NS092701-03 and a Wellcome Trust Principal Research Fellowship awarded to P. Read Montague.

Abstract

The Hierarchical Taxonomy of Psychopathology (HiTOP) organizes phenotypes of mental disorder based on empirical covariation, offering a comprehensive organizational framework from narrow symptoms to broader patterns of psychopathology. We argue that established self-report measures of psychopathology from the pre-HiTOP era should be systematically integrated into HiTOP to foster cumulative research and further the understanding of psychopathology structure. Hence, in this study, we mapped 92 established psychopathology (sub)scales onto the current HiTOP working model using data from an extensive battery of self-report assessments that was completed by community participants and outpatients ($N = 909$). Content validity ratings of the item pool were used to select indicators for a bifactor-(S-1) model of the p factor and five HiTOP spectra (i.e., internalizing, thought disorder, detachment, disinhibited externalizing, and antagonistic externalizing). The content-based HiTOP scales were validated against personality disorder diagnoses as assessed by standardized interviews. We then located established scales within the taxonomy by estimating the extent to which scales reflected higher-level HiTOP dimensions. The analyses shed light on the location of established psychopathology scales in HiTOP, identifying pure markers and blends of HiTOP spectra, as well as pure markers of the p factor (i.e., scales assessing mentalizing impairment and suspiciousness/epistemic mistrust).

Johannes Zimmermann and Gabriel Olaru shared last authorship.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors Personality and Mental Health Published by John Wiley & Sons Ltd.

INTRODUCTION

Traditional psychopathology taxonomies (e.g., diagnostic categories as suggested in the *Diagnostic and Statistical Manual of Mental Disorders* [DSM-5]; American Psychiatric Association [APA], 2013) have severe drawbacks such as artificial comorbidities, arbitrary diagnostic cutoffs, diagnostic instability, and phenotypic heterogeneity that limit their usefulness and practical applicability (e.g., Kotov et al., 2018; Krueger et al., 2018). The *Hierarchical Taxonomy of Psychopathology* (HiTOP) provides an alternative classification system of signs and symptoms of mental disorder as well as maladaptive traits that is built on factor analytic studies of empirical covariation (Kotov et al., 2017). By this means, HiTOP aims to provide a more efficient as well as fine-grained diagnostic conceptualization of mental health problems, following a dimensional rather than a categorical approach to classification (e.g., Markon et al., 2011). Initial results emphasize its potential for providing a better understanding of the nature, scope, and etiology of mental disorders (e.g., Kotov et al., 2020; Krueger et al., 2021; Waszczuk et al., 2020; Watson, Levin-Aspenson, et al., 2022; but see Haefffel et al., 2021), thereby revitalizing clinical psychology research and practice (e.g., Conway et al., 2019; Hopwood et al., 2020; Ruggero et al., 2019).

The hallmark feature of HiTOP is to provide a framework for considering the full breadth and depth of psychopathology in a hierarchical order. HiTOP thus enables the specificity and generality of mental health problems to be located at multiple levels of granularity (e.g., Conway et al., 2019). A general factor of psychopathology, the *p* factor, at the apex of the hierarchy embodies sizeable covariance between most—if not all—indicators of psychological distress (e.g., Caspi & Moffitt, 2018; Constantinou & Fonagy, 2019). The *p* factor is longitudinally stable, moderately heritable, and associated with important functional outcomes (e.g., for a review, see Lahey et al., 2021). Yet the question of whether the *p* factor phenomenon is a mere statistical abstraction (e.g., Fried, 2020; Levin-Aspenson et al., 2021; Watts et al., 2020), captures a substantive construct (e.g., Lahey et al., 2021), or is a mix of both (e.g., Watts et al., 2022) remains controversial. First, as per interpretations of the *p* factor as a common cause, it has been suggested that it reflects general liabilities towards psychopathology such as compromised brain function (e.g., Caspi et al., 2020), impairments in self- and interpersonal functioning (Widiger et al., 2019), or, more specifically, impairments in social learning in terms of problems with mentalizing and epistemic trust (e.g., Fonagy et al., 2021; Fonagy & Campbell, 2021). Other researchers have pointed out that the *p* factor can

also be explained, at least in part, by method-specific causes, such as evaluative biases (e.g., Leising et al., 2020; Pettersson et al., 2014; Smith et al., 2020). Second, some researchers think of the *p* factor as an index of symptomatic distress (Fried et al., 2021) or an index of impairment (McCabe et al., 2022) that is secondary to the disorders themselves. Third, network theorists consider that the *p* factor may not reflect common causes but rather direct causal paths between mutually reinforcing symptoms (e.g., Borsboom & Cramer, 2013; Bringmann et al., 2021). One level down are located the superspectra of emotional dysfunction, psychosis, and externalizing (e.g., Kotov et al., 2020; Krueger et al., 2021; Watson, Levin-Aspenson, et al., 2022). Beneath these, the current working model of HiTOP specifies six dimensions on an intermediate level of the hierarchy—the spectra of internalizing, antagonistic externalizing (hereinafter simply referred to as antagonism), disinhibited externalizing (disinhibition), thought disorder, detachment, and somatoform (Kotov et al., 2017). Lower levels of the hierarchy may consist of finer grained symptom clusters, but fewer studies have investigated their number, nature, or structure (e.g., Cicero et al., 2022; Forbes et al., 2021; Mullins-Sweatt et al., 2022; Sellbom et al., 2022; Watson, Forbes, et al., 2022; Zimmermann et al., 2022).

HiTOP thus offers a comprehensive taxonomy that allows a multidimensional classification of mental health problems and is becoming an increasingly influential alternative to traditional categorical nosologies (Kotov et al., 2021). Conversely, this also means that existing measures of psychopathology that have been used in clinical research and practice for a long time (i.e., in the pre-HiTOP era) should also be mapped to this taxonomy. However, few studies have attempted to locate established self-report questionnaires of psychopathology within the HiTOP model (e.g., Sellbom et al., 2020, 2021; Wright & Simms, 2015). Established psychopathology scales tend to follow more traditional clinical conceptualizations that are often tied to specific diagnostic concepts (e.g., symptoms of depression as measured by the Brief Symptom Inventory [BSI], Derogatis & Spencer, 1993) or are more narrowly circumscribed (e.g., dissociation as measured by the Dissociative Experiences Scale [DES], Bernstein & Putnam, 1986). Self-report measures like these are generally used to provide an economic assessment of psychopathologies for clinical research or to screen for mental disorders in clinical practice. To date, the plethora of tests are studied separately and often have unclear conceptual boundaries (Fried, 2017), thus lacking integration into an overarching conceptualization of psychopathology, which is offered by HiTOP. With accumulating evidence indicating the relevance and utility of higher-level HiTOP dimensions (e.g., Kotov et al., 2020;

Krueger et al., 2021; Smith et al., 2020; Watson, Levin-Aspenson, et al., 2022), it is likely that these superimpose with the more unique information that is conveyed by established psychopathology scales. Established scales therefore need to be re-examined regarding their distinctiveness above and beyond these higher-level dimensions (see Müller et al., 2022, for an example).

This issue is also relevant when investigating nomological networks. Indeed, it is well documented that many risk (e.g., adverse childhood experiences and low cognitive functioning) and outcome variables (e.g., self-harm and incarceration) are similarly related to psychopathology (e.g., Kotov et al., 2020; Krueger et al., 2021; Smith et al., 2020; Watson, Levin-Aspenson, et al., 2022). Statistical associations may thus not be specific to the construct that a test purports to measure. For example, a significant association between concurrent symptoms of dissociation and past childhood experiences (e.g., as reported in Van IJzendoorn & Schuengel, 1996) could just as well be ascribed to the statistical influence of higher-level dimensions, so that such associations likely generalize to many other psychopathological phenomena (e.g., Lahey et al., 2021). Importantly, given that everything is somehow related to everything else (also: crud factor; for a recent discussion, see Orben & Lakens, 2020) and given that this seems to be particularly true for psychopathology constructs, deeper insights can only be obtained when focusing on the magnitude and specificity of effects. Thus, it may only be feasible to determine whether these associations are truly unique if higher-level dimensions are assessed with high fidelity and accounted for statistically.

The aforementioned issues raise questions about isolated interpretations of traditional self-report measures of psychopathology. Given that HiTOP can provide a comprehensive taxonomy to organize psychopathology constructs in an integrated and connected manner, we argue that established psychopathology scales could also be mapped to HiTOP, as has been recently done, for example, for the Minnesota Multiphasic Personality Inventory-3 (Sellbom et al., 2021). In this way, it is possible to examine the measured constructs more thoroughly, to highlight issues of discriminant validity and specificity (Conway et al., 2019; Stanton et al., 2020), to expose similarities and differences between measures, thereby identifying and preventing jingle-jangle fallacies (e.g., Kelley, 1927; Lawson & Robins, 2021), and to foster cumulative research integration. In a similar vein, personality researchers renewed their call for a (more stringent) use of Big Five dimensions to provide a unifying framework for organizing psychological trait dimensions (Bainbridge et al., 2022). Furthermore, existing research

on the structure of psychopathology that has informed the current HiTOP working model is subject to some limitations, as has been pointed out repeatedly (e.g., Kotov et al., 2017). Among these are the reliance on (a) diagnostic categories (e.g., Ringwald et al., 2021) that neglect symptom-level information or on (b) single symptom indicators (e.g., Forbes et al., 2021) that preclude from detecting multidimensionality at lower levels of the hierarchy (e.g., Bollen & Lennox, 1991; Watts et al., 2021). Such considerations have led the HiTOP consortium to start developing a HiTOP measure (Simms et al., 2022) that holds promise in addressing these limitations in terms of realizing a psychometrically optimized multiple indicator assessment. Given the fact that paradigm shifts are implemented only slowly, established psychopathology scales will keep playing an important role to spur additional insights into psychopathology and its structure, due also to their diversity in terms of relying on different clinical conceptualizations and traditions.

Using clinical and community data from an extensive assessment of self-reported psychopathology (i.e., 685 items) including measures of self- and interpersonal functioning, we aimed to map 92 established psychopathology scales from 21 questionnaires onto the current HiTOP working model (Kotov et al., 2017). In a content-based approach to assess HiTOP spectra, we selected indicators from the item pool using expert ratings of item content (e.g., see Colquitt et al., 2019) in a first step, further purified this selection by factor analysis to ensure (essential) unidimensionality in a second step, and realized measurement models in terms of a bifactor-(S-1) model and a correlated factors model in a third step. To test the convergent and discriminant validity of our newly derived HiTOP scales, we evaluated associations with personality disorder (PD) diagnoses following the *Structured Clinical Interview for DSM-IV Axis 2 Disorders* (SCID-II). Clinical diagnoses as assessed by standardized interviews provide a useful validation criterion because (1) they are heteromethod and (2) meta-analytic findings of the associations between HiTOP spectra and DSM diagnostic categories are available (Ringwald et al., 2021). For our main analysis, we estimated the extent to which established scales reflect higher-level HiTOP dimensions (i.e., p factor and HiTOP spectra) to shed more light on which established scales are pure markers or reflect blends of HiTOP dimensions. To this end, we applied bifactor modeling (e.g., Eid et al., 2017) to model the p factor and HiTOP spectra jointly, reflecting the two most prominent upper levels of the psychopathology hierarchy. By mapping many established scales onto HiTOP, we aimed to gain additional insights for understanding psychopathology structure and its measurement.

METHODS

Samples

Participants were recruited in Greater London via the Personality and Mood Disorder Research Consortium consisting of 260 healthy community participants and 649 outpatients ($N = 909$; 66% female; mean age of 30.7, range = 16–65, $SD = 10.4$) from National Health Service Improving Access to Psychological Therapies (NHS IAPT) services for Mood Disorders and secondary or tertiary specialist services for PDs referred from National Health Service specialist PD clinical services. In the total sample, a large number of participants met diagnostic criteria for current Borderline PD (59%), Paranoid PD (27%), Antisocial PD (23%), Narcissistic PD (4%), Schizotypal PD (4%), and Histrionic PD (1%) according to the DSM-5 (APA, 2013). Participants with PDs were oversampled because patients with severe mental health problems were primarily referred who are considered too complex for standard care due to multimorbidity or risk to self or others. The data have been previously used to study various research questions distinct from the current research (Euler et al., 2019; Huang et al., 2020; Rifkin-Zybutz et al., 2021; Wendt et al., 2019).

Measures

Established psychopathology scales

Participants completed a battery of established self-report questionnaires, indicating their agreement to statements about themselves on rating scales. We included a plethora of measures designed to assess current or persistent signs, symptoms, and characteristic traits of mental disorders, including maladaptive personality traits and measures of personality functioning (see DeYoung et al., 2020, for how maladaptive personality traits are linked to HiTOP). The measures were the Autonomous Functioning Index (AFI; Weinstein et al., 2012), Antisocial Process Screening Device (APSD; Frick & Hare, 2001), Barratt Impulsiveness Scale (BIS-11; Patton et al., 1995), Brief Symptom Inventory (BSI; Derogatis & Spencer, 1993), Drugs, Alcohol, and Self-Injury Questionnaire (DASI; Wilkinson et al., 2018), Difficulties in Emotion Regulation Scale (DERS; Gratz & Roemer, 2004), Dissociative Experiences Scale (DES; Bernstein & Putnam, 1986), Experiences in Close Relationships-Revised (ECR-R; Fraley et al., 2000), Empathy Quotient (EQ; Baron-Cohen & Wheelwright, 2004), Green et al. Paranoid Thoughts Scale (GPTS; Green et al., 2008), Inventory of Interpersonal Problems (IIP-32; Horowitz

et al., 2000), Life History of Aggression (LHA; Coccaro et al., 1997), Other as Shamer Scale (OAS; Goss et al., 1994), Operationalized Psychodynamic Diagnosis: Structure Questionnaire (OPD-SQ; Ehrenthal et al., 2012), Personality Assessment Inventory – Borderline Scale (PAI-BOR; Morey, 2014), Personality Assessment Inventory – Antisocial Scale (PAI-ANT; Morey, 2014), Posttraumatic Stress Checklist Scale – Civilian Version (PCL-C; Blanchard et al., 1996), Reflective Functioning Questionnaire – Extended 18-Item Version (RFQ-18; Rogoff et al., 2021), Standardized Assessment of Personality: Abbreviated Scale (SAPAS; Moran et al., 2003), Schizotypal Personality Questionnaire (SPQ; Raine, 1991), and Levenson Self-Report Psychopathy Scale (SRPS; Levenson et al., 1995). For this study, scales were inverted when necessary, so that higher values were geared towards the maladaptive pole of a trait dimension indicating greater severity or impairment in the respective domain. For more detailed information including the number of items and scales, response categories, and internal consistency estimates, see Table S1. In the questionnaires included are 92 (presumably unidimensional) scales. To use scale scores in the subsequent latent variable analysis, we tested unidimensional measurement models (Little et al., 2013), except for DASI Drugs and alcohol, for which we relied on a formative measurement model and used the manifest sum score. We report fit statistics of the unidimensional models for scale scores in Table S2.

Structured Clinical Interview for DSM-IV Axis 2 Disorders (SCID-II)

To assess current symptoms of PD according to DSM-IV (i.e., Paranoid PD, Schizoid PD, Schizotypal PD, Antisocial PD, Borderline PD, Histrionic PD, Narcissistic PD, Avoidant PD, Dependent PD, and Obsessive-compulsive PD), structured interviews were conducted using the SCID-II (First & Gibbon, 2004). The interviews were administered by mental health professionals. For this study, we considered symptom counts of PD diagnoses that are the number of endorsed symptoms in each diagnostic category.

Content-based HiTOP scales

Drawing from the item content of the questionnaires described above, we derived a measurement of HiTOP spectra (i.e., internalizing, antagonism, disinhibition, detachment, thought disorder, and somatoform) and the p factor. We assumed that the item pool provides a

sufficiently broad representation of higher-level psychopathology dimensions (i.e., 685 items in total). HiTOP dimensions are commonly identified in a data-driven way using factor analytic methods. However, in this study, we relied on a content-based approach with expert ratings of the item pool (e.g., Colquitt et al., 2019) in a first step and factor analysis to ensure (essentially) unidimensional scales in a second step. In a third step, we realized a bifactor-(S-1) model and a correlated factors model that were used for estimating the associations between HiTOP dimensions and other variables in the main statistical analyses.

Step 1: Expert ratings

Eight raters (i.e., three of the authors and five trained psychology undergraduate students) were presented all items in randomized order and were asked to evaluate to what extent items are characteristic of each of the HiTOP spectra. The raters familiarized themselves with the original HiTOP publication (Kotov et al., 2017), and it was ensured that raters were knowledgeable of the common definitions of the signs, symptoms, and characteristic traits of mental health problems that are considered by the model. Ratings were provided on a 3-point scale (0 = *not characteristic*, 1 = *possibly characteristic*, 2 = *definitely characteristic*). The interrater reliability for the average of eight judges was acceptable with interclass correlations (ICC[2, 8]; Fleiss & Shrout, 1978) ranging between 0.84 (internalizing) and 0.92 (thought disorder). Items were deemed to be characteristic when the mean rating was >1.2 for one spectrum and <0.8 for other spectra. Overall, we retained a large number of indicators that were evaluated as indicative of HiTOP spectra. However, due to insufficient representation of the somatoform spectrum (i.e., only four items were selected by raters), it was not included in subsequent analyses.

Step 2: Exploratory factor analysis

In the next step, we ensured that the selected items of each spectrum loaded on a common general factor. To this end, we conducted exploratory factor analysis (EFA) on the item pools derived in the previous step. The number of factors was determined based on model fit and the emergence of well-defined factors using geomin rotation.¹ To extract a common general factor, we used orthogonal bifactor rotation (Mansolf & Reise, 2016) and discarded items with weak loadings on the general factor for each spectrum (<0.40). The final number of retained indicators was 76 for internalizing (with 3 items removed due to weak loadings on the general factor), 37 for antagonism (16 items removed), 35 for disinhibition (9 items removed), 39 for detachment (3 items removed), and 49 for thought disorder (3 items removed). These items

represented the content-based HiTOP scales that formed the basis for the measurement models (i.e., bifactor S-1 model and correlated factors model).

Step 3: Bifactor-(S-1) and correlated factors model

Two measurement models were realized to operationalize HiTOP dimensions in a latent variable framework. On the one hand, we used the correlated factors model to estimate associations between HiTOP spectra and PD diagnoses because this facilitates comparison with the results reported by Ringwald et al. (2021). On the other hand, we used bifactor modeling (e.g., Rodriguez et al., 2016) to separate and jointly consider two levels of the HiTOP hierarchy (in terms of the p factor and HiTOP spectra) when mapping established psychopathology scales onto HiTOP. Bifactor modeling is useful for studying external relations of hierarchical constructs (e.g., Bornovalova et al., 2020) because the variance of the indicators can be clearly partitioned into variance common to all indicators (i.e., modeled by a general factor; here: p factor), variance specific to a set of indicators in a given content domain (i.e., modeled by specific factors; here: HiTOP spectra, which are orthogonal to the general trait), and variance not explained by latent factors (i.e., modeled as indicator-specific residual variances). To date, studies have mostly used traditional bifactor models to operationalize HiTOP spectra and the p factor (e.g., Forbes et al., 2021; Lahey et al., 2021). However, traditional bifactor models are prone to estimation problems, such as vanishing specific factors, Heywood cases, or other implausible estimates (Eid et al., 2017). We thus implemented an orthogonal bifactor-(S-1) model (Eid et al., 2017) that has particularly beneficial properties for studying the external relations of hierarchical constructs (Moshagen, 2021; Zhang et al., 2021). In this model, one specific factor is removed so that items of that factor only load on the general factor and thus serve as a *reference* domain for the general factor.

With respect to the indicators for realizing the measurement models, we relied on a homogeneous parceling approach (i.e., opting for item-to-construct balance; Little et al., 2002) to minimize undesirable sources of multidimensionality (e.g., Little et al., 2013; Rhemtulla, 2016). Considering that the included questionnaires differ in the number of response categories, we rescaled item responses from 0 to 100 before creating the parcels (i.e., percent of maximum possible; Cohen et al., 1999). We created three parcels for each HiTOP spectrum using the items that were retained in the previous step. For the correlated factors model, the parcels of each HiTOP spectrum loaded on a corresponding factor. For the bifactor-(S-1) model, one quarter of the items in each HiTOP spectrum were withheld to create statistically

independent parcels for the p factor because, as mentioned earlier, separate indicators are needed that load exclusively on the general factor but not on any specific factor. The remaining items were used to create parcels that loaded on both the general factor and a corresponding specific factor. By using p factor parcels that aggregate across HiTOP spectra, a shortcoming of bifactor-(S-1) models can be circumvented (i.e., equating the p factor with one of the HiTOP spectra) while still facilitating model estimation. Indeed, our model aligns with an operational definition of the p factor in which the p factor simply reflects sum scores of psychopathology indicators (e.g., Fried et al., 2021). The factors of our parcel-based bifactor-(S-1) model have a clear meaning in a descriptive sense: Whereas the p factor reflects the total symptomatic distress (irrespective of content), the specific factors indicate whether symptoms in a HiTOP spectrum are relatively more pronounced or less pronounced than what would be expected given the standing on the p factor.

Fit indices for the correlated factors model (see Figure S1) were acceptable, scaled $\chi^2(160) = 978.6$, Comparative Fit Index (CFI) = 0.96, Root Mean Square Error of Approximation (RMSEA) = 0.08, Standardized Root Mean Square Residual (SRMR) = 0.05. Factor loadings ranged from 0.81 to 0.97 and factor correlations ranged from 0.43 to 0.78. The bifactor-(S-1) model (see Figure S2) had acceptable fit, scaled $\chi^2(120) = 763.1$, CFI = 0.96, RMSEA = 0.07, SRMR = 0.05. The factor loadings of the bifactor-(S-1) model were all positive and model parameters were plausible. Factor loadings on the p factor were highest for indicators of the p factor (ranging from 0.92 to 0.96), followed by internalizing (from 0.87 to 0.89), thought disorder (from 0.76 to 0.85), disinhibition (from 0.71 to 0.76), detachment (from 0.62 to 0.70), and antagonism (from 0.43 to 0.64). The size of factor loadings on the specific factors was in opposite order with antagonism indicators having the strongest loadings on their corresponding specific factor (from 0.58 to 0.75), followed by indicators of detachment (from 0.52 to 0.63), disinhibition (from 0.43 to 0.62), thought disorder (from 0.29 to 0.51), and internalizing (from 0.37 to 0.39). This shows that, in the current study, the p factor was most strongly indicated by internalizing content and less so by antagonism content.

Statistical analysis

Convergent and discriminant validity of content-based HiTOP scales

To test the convergent and discriminant validity of our newly derived content-based HiTOP scales (using the

correlated factors model), we evaluated their associations with PD diagnoses by comparing them against meta-analytic estimates as reported in Ringwald et al. (2021). Specifically, if our content-based HiTOP scales offered an adequate approximation of HiTOP spectra, the correlation patterns to PD diagnoses in the current study should be similar to the meta-analytically derived factor loading patterns of PD diagnoses. Despite methodological differences between the two approaches, they can be compared because they address the same question (i.e., statistical association between PD diagnoses and HiTOP spectra).² Ringwald et al. (2021) regarded PD diagnoses to be markers of HiTOP spectra when factor loadings were equal or larger than the absolute value of 0.30. They reported PD diagnoses to be markers of internalizing (i.e., Avoidant PD and Borderline PD), antagonism (i.e., Antisocial PD, Borderline PD, Histrionic PD, Narcissistic PD, Obsessive–compulsive PD, and Paranoid PD), disinhibition (i.e., Antisocial PD), thought disorder (Paranoid PD, Schizotypal PD, and Schizoid PD), and detachment (i.e., Avoidant PD, Obsessive–compulsive PD, Schizotypal PD, and low Histrionic PD). A schematic model of this analysis is depicted in Figure S3.

Mapping established psychopathology scales onto HiTOP

To map established scales onto HiTOP, we conducted structural equation modeling to regress the factors of established scales on the factors of the bifactor-(S-1) model. An illustration of this model is presented in Figure 1. Separate regression models were used to predict each of the scales (i.e., 92 model estimations in total). Given that content-based HiTOP scales draw from the same item content as the established scales, we needed to prevent unmodeled correlated residual variances from inflating the estimates of association (i.e., criterion contamination). Hence, we excluded items to be considered as indicators for the HiTOP factors when they were part of the criterion scale and reassembled the item parcels for each of the 92 models, thereby ensuring that the same items were not considered in both the criterion and the predictor variables.³

To guide interpretation of the regression models, we regarded standardized regression coefficients of HiTOP spectra (i.e., β_{1-5}) equal or larger to the absolute value of 0.20 as indicating that an established scale reflected a HiTOP spectrum markedly, as this is an effect size typically observed in psychological research (Gignac & Szodorai, 2016). When only one regression coefficient of HiTOP spectra (β_{1-5}) was above the cutoff, we considered an established scale to be a *pure marker* of a HiTOP

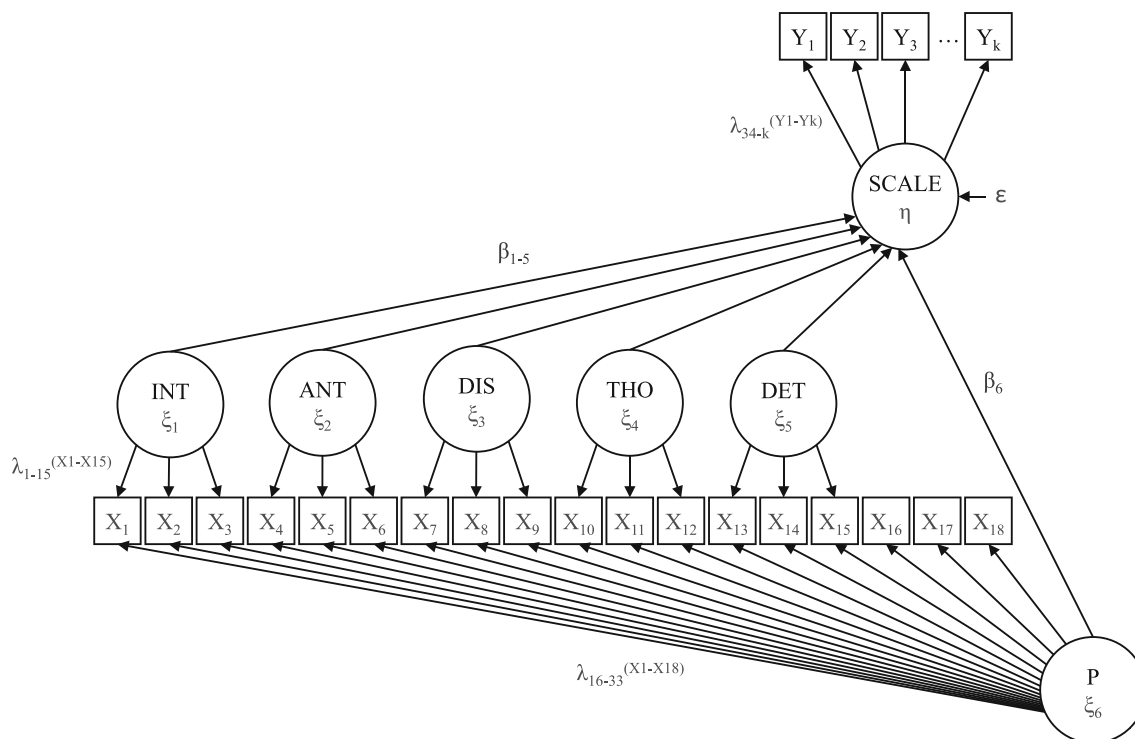


FIGURE 1 Schematic display of the latent regression model used to map established psychopathology scales onto the Hierarchical Taxonomy of Psychopathology (HiTOP). *Note:* Indicator residual variances and latent factor variances are not displayed. HiTOP factors (i.e., INT–DET and P) were modeled using an orthogonal bifactor-(S-1) approach. Target scales (SCALE) were modeled as a unidimensional simple structure (except for DASI Drugs and alcohol). ANT = antagonism; DET = detachment; DIS = disinhibition; INT = internalizing; P = P factor; SCALE = target scale; THO = thought disorder; X_{16} – X_{18} = parcel indicators of the general factor (i.e., P); X_1 – X_{15} = parcel indicators of HiTOP spectra (i.e., INT–DET); Y_1 – Y_k = item indicators of the target scale; β = regression path; ϵ = residual variance of the dependent latent factor; η = dependent latent factor; λ = factor loading; ξ = independent latent factor.

spectrum. When multiple regression coefficients of HiTOP spectra (β_{1-5}) were above the cutoff, we deemed a scale to reflect a *blend* of HiTOP spectra. Also, when all regression coefficients of HiTOP spectra (β_{1-5}) were below the cutoff but the regression coefficient of the p factor (for β_6) was above the cutoff, we deemed a scale to be a pure marker of the p factor. Finally, if neither HiTOP spectra nor the p factor yielded a regression coefficient above the cutoff, we concluded that a scale was not captured by HiTOP at all.

Model estimation and software packages

All analyses were conducted in R Version 4.1.1 (R Core Team, 2021) unless stated otherwise. Models were estimated using robust maximum likelihood (MLR) or weighted least squares mean and variance adjusted (WLSMV). Items with five or more ordinal responses as well as parcels were considered as continuous indicators and items with four or fewer ordinal responses were considered as ordered indicators (Rhemtulla et al., 2012).

Structural equation models and confirmatory factor analysis were estimated with the R package lavaan Version 0.6.9 (Rosseel, 2012), and bifactor-rotated EFA was conducted with Mplus Version 8.4 (Muthén & Muthén, 2017). R code for reproducing the analyses can be accessed at <https://osf.io/hkav3/>. The data are available from the corresponding author on reasonable request.

RESULTS

Convergent and discriminant validity of content-based HiTOP scales

The pattern of associations between interview-based PD diagnoses and self-reported HiTOP spectra appeared to be similar to the pattern of meta-analytic estimates of factor loadings reported in Ringwald et al. (2021), which supported the validity of the newly derived content-based HiTOP scales. The latent correlations are displayed in Table 1. First, we will refer to PDs for which correlation

TABLE 1 Correlations between HiTOP spectra and symptom counts of SCID-II personal disorder diagnostic categories

	INT	ANT	DIS	THO	DET
Narcissistic PD	−0.02	0.27* ^a	0.09	0.06	0.04
Histrionic PD	0.07	0.24* ^a	0.24*	0.10	−0.13* ^a
Borderline PD	0.47* ^a	0.35* ^a	0.55*	0.40*	0.20*
Antisocial PD	0.05	0.43* ^a	0.45* ^a	0.23*	0.08
Paranoid PD	0.43*	0.39* ^a	0.38*	0.40* ^a	0.32*
Schizoid PD	0.08	0.11	0.06	0.07 ^a	0.38* ^a
Schizotypal PD	0.31*	0.24*	0.23*	0.48* ^a	0.26*
OCPD	0.19*	0.10* ^a	0.01	0.09	0.11* ^a
Avoidant PD	0.47* ^a	0.14*	0.25*	0.28*	0.44* ^a
Dependent PD ^b	0.33*	0.06	0.22*	0.27*	0.12*

Note: HiTOP spectra were modeled using a correlated factors model, and personality disorder symptom counts were modeled as manifest variables.

ANT = antagonism; DET = detachment; DIS = disinhibition; INT = internalizing; OCPD = Obsessive–compulsive PD; PD = personality disorder; THO = thought disorder.

^aSalient standardized factor loading (>0.30) of a PD diagnosis with a HiTOP spectrum as reported in Ringwald et al. (2021).

^bDependent PD was not included in Ringwald et al. due to estimation problems.

* $p < 0.05$.

patterns seemed fully consistent with Ringwald et al. (2021). Narcissistic PD was positively associated with antagonism ($r = 0.27$). Histrionic PD was related to antagonism ($r = 0.24$) as well as to low detachment ($r = -0.13$). Antisocial PD was most strongly related to antagonism ($r = 0.43$) and disinhibition ($r = 0.45$). Schizotypal PD was most strongly related to detachment ($r = 0.48$). Avoidant PD was most strongly related to internalizing ($r = 0.47$) and detachment ($r = 0.44$). Second, we will point to results that were not fully consistent with Ringwald et al. (2021), or at least not in every regard. Although, as expected, Borderline PD was strongly related to both internalizing ($r = 0.47$) and antagonism ($r = 0.35$), there were unexpected associations of similar magnitude with disinhibition ($r = 0.40$) and thought disorder ($r = 0.55$). In line with the results of Ringwald et al. (2021), Paranoid PD was in fact strongly associated with antagonism ($r = 0.39$) and thought disorder ($r = 0.40$), but it was also strongly related to the other HiTOP spectra with correlation coefficients between 0.32 and 0.43. In line with expectations, Schizoid PD was strongly related to detachment ($r = 0.38$), though against expectations, it was not significantly related to thought disorder ($r = 0.07$). As in Ringwald et al. (2021), Obsessive–compulsive PD was associated with antagonism ($r = 0.10$) and detachment ($r = 0.10$), but another significant association was found with internalizing ($r = 0.19$). There were no expectations regarding Dependent PD, as no results were reported in the meta-analysis by Ringwald et al. (2021). In sum, our results aligned well with the reported associations by Ringwald et al. (2021), considering that methodological

differences between studies can likely account for moderate deviations (e.g., sample characteristics, methods, and indicators used to operationalize HiTOP spectra).

Mapping established psychopathology scales onto HiTOP

The bifactor-(S-1) models converged normally and the fit was acceptable (see Table S4). The complete list of standardized regression coefficients is displayed in Table S5. To better visualize the results, we used variance decompositions that depict the extent to which HiTOP dimensions are reflected in the established scales (Figure 2). To this end, standardized regression coefficients were taken to the square to indicate the variance explained by each predictor.

Most of the variance in the established scales was explained by HiTOP factors, with the p factor explaining an average of 54% and HiTOP spectra explaining an additional 14% (i.e., 69% in total). With the decision to interpret standardized regression coefficients $> |0.20|$ as marked associations (as described in the Methods section), most scales could be considered pure markers of a single HiTOP spectrum (i.e., 54 scales), whereas fewer scales (i.e., 23) represented blends of HiTOP spectra. This indicates that most scales could be allocated relatively unambiguously to a spectrum when the p factor was taken into account. Among the established scales included in this study, we found 27 scales that were pure markers of internalizing, 5 for thought disorder, 6 for detachment, 9 for disinhibition, and 7 for antagonism.

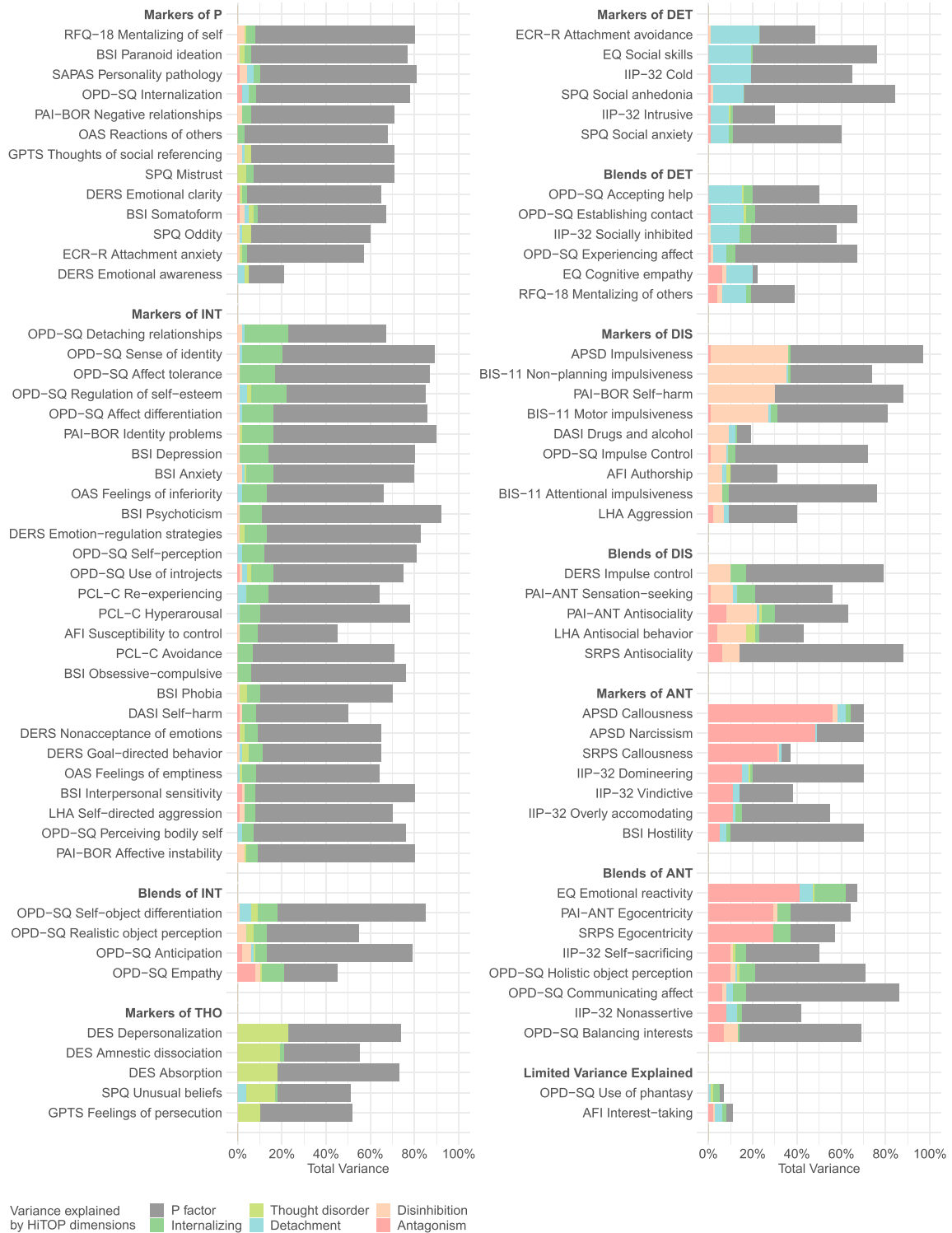


FIGURE 2 Mapping established psychopathology scales onto the Hierarchical Taxonomy of Psychopathology (HiTOP). *Note:* Variance explained by the p factor (modeled as a general factor) and HiTOP spectra (modeled as specific factors) was calculated by taking the square of standardized regression coefficients (noted as β_{1-6} in Figure 1). The order of the scales indicates their estimated location within the HiTOP model as based on our results. ANT = antagonism; DET = detachment; DIS = disinhibition; INT = internalizing; THO = thought disorder; Total Variance = total variance of the latent factor.

Three specific blends were found most frequently: high-externalizing–low-internalizing (as represented by seven scales), high-detachment–high-internalizing (i.e., five scales), and high-antagonism–high-disinhibition (i.e., four scales). In addition, 12 scales were found to exclusively represent the p factor. Finally, it should be noted that limited explained variance was found for only two scales (i.e., OPD-SQ Use of phantasy and AFI Interest-taking).

For readers who are particularly interested in how specific questionnaires are related to HiTOP dimensions, we provide Figure S4 in which the results are visually arranged according to the alphabetical order of the questionnaires. In the following, we will provide some examples for illustrative purposes. For some questionnaires, all scales incorporated therein were mapped to a single HiTOP spectrum. These were the BIS-11 scales (reflecting disinhibition), the DES scales (thought disorder), and the PCL-C scales (internalizing). Other questionnaires had scales predominantly tapping into the internalizing HiTOP spectrum (i.e., BSI, DERS, and OAS). However, most questionnaires had scales tapping multiple HiTOP spectra. To name a few examples, the EQ tapped into both antagonism (e.g., low EQ Emotional reactivity) as well as detachment (e.g., low EQ Social skills). Similarly, the IIP-32 reflected antagonism (e.g., IIP-32 Domineering) and detachment (e.g., IIP-32 Cold). The APSD tapped into disinhibition (e.g., APSD Impulsiveness) and antagonism (e.g., APSD Callousness).

DISCUSSION

In this study, we mapped 92 established psychopathology scales (including signs and symptoms of mental disorder as well as maladaptive traits and indicators of personality functioning) onto the current working model of HiTOP. To this end, we derived content-based scales of HiTOP dimensions, tested their validity, used a bifactor-(S-1) model to separate p factor and spectra statistically, and calculated their associations with established scales in order to estimate the location of scales within the HiTOP framework. The scales tended to be covered well by higher-level HiTOP dimensions and their estimated locations corresponded closely with their current placement in the HiTOP model. These findings underline the capacity of HiTOP to efficiently organize and summarize self-reported psychopathology and it strengthens the notion that established psychopathology measures could and should be integrated into HiTOP.

In previous studies, p factors tended to be saturated with content of the internalizing domain, albeit considerable inconsistencies were documented between studies

that may be related to characteristics of the sample and the indicators used (e.g., Levin-Aspenson et al., 2021; Watts et al., 2020). While we do also find strong empirical overlap between internalizing and the p factor in this study, we also find them to have unique prototypical markers among the included scales that signal their distinctiveness. Our results further demonstrate that, after the p factor is taken out, psychopathology constructs can be linked to single HiTOP spectra or specific blends with clarity and consistency. These findings highlight the utility of the bifactor-(S-1) modeling approach to separate out the general disposition to mental health problems from the specific indications associated with more narrow symptoms of mental disorder, maladaptive personality traits, or indicators of personality functioning.

HiTOP structure

In the following, we will discuss how our findings may further the understanding of psychopathology structure. As pointed out previously, the estimated location of established scales tended to match the current placement of constructs in HiTOP. However, there were some noteworthy deviations that we will also discuss.

The pure markers of internalizing were (1) scales that assess intensely aversive states of negative emotionality (OPD-SQ Affect tolerance and OPD Affect differentiation) that are experienced as uncontrollable (DERS emotion regulation strategies and PAI-BOR Mood instability), including anxiety, phobia, depression (BSI Anxiety, BSI Phobia, and BSI Depression), posttraumatic stress (PCL-C Re-experiencing), and separation anxiety (OPD-SQ Detaching relations); (2) scales that assess adverse physiological or behavioral aspects of intense negative emotionality, such as arousal (PCL-C Hyperarousal), concentration problems (DERS Goal-directed behavior), avoidance (PCL-C Avoidance), and self-harm (DASI Self-harm and LHA Self-harm); and (3) scales that assess unstable or diffuse self-image (OPD-SQ Sense of identity, PAI-BOR Identity problems, and OPD-SQ Self-perception), as well as negative self-evaluation (OPD-SQ Regulation of self-esteem, OAS Feeling of inferiority, BSI Interpersonal sensitivity, OPD-SQ Use of introjects, OAS Feeling of emptiness, and OPD-SQ Bodily self). This pattern is consistent with the current HiTOP working model of the internalizing domain (Watson, Levin-Aspenson, et al., 2022).

Whereas previous studies regularly indicate what features of psychopathology tend to be most strongly related to the p factor, our study is the first to investigate pure markers of the p factor. We find pure markers of the p factor to be (1) scales that assess mentalizing

impairments regarding one's own mental states in general (RFQ-18 Mentalizing of self) and with respect to one's own feelings and emotions in the specific (DERS Emotional clarity and DERS Emotional awareness) and (2) scales that assess suspiciousness and mistrust towards others in terms of feeling negatively evaluated by others (PAI-BOR Negative relationships, GPTS Thoughts of social referencing, and SPQ Mistrust), feeling estranged (SPQ Oddity), feeling unfairly treated or let down (BSI Paranoid Ideation and OAS Reactions of others), or expecting this to happen (OPD-SQ Internalization and ECR-R Attachment anxiety). These results are consistent with views that consider mentalizing impairments and epistemic mistrust as defining features of the p factor (Fonagy et al., 2021; Fonagy & Campbell, 2021) and that place self- and interpersonal functioning at the core of psychopathology (Widiger et al., 2019; Wright et al., *in press*). In fact, whereas we found evidence for suspiciousness to be a pure marker of the p factor, it has previously been placed in the spectra of detachment (Zimmermann et al., 2022), antagonism (Krueger et al., 2021; Mullins-Sweatt et al., 2022), or thought disorder (Cicero et al., 2022). Yet consistent with our results, studies using the *Personality Inventory for DSM-5* (APA, 2013) have indicated that suspiciousness exhibits strong associations with a general PD factor but low domain-specificity (e.g., Somma et al., 2019; Williams et al., 2018).

With respect to the thought disorder spectrum, we found pure markers to be scales assessing unusual or odd beliefs and experiences or perceptual irregularities such as supernatural phenomena (i.e., SPQ Unusual beliefs), dissociative or psychotic experiences (DES Depersonalization, DES Amnestic dissociation, and DES absorption), and feeling persecuted or conspired against by others (GPTS Feelings of persecution). Markers of detachment were measures pertaining to avoiding social contacts and intimacy (i.e., IIP-32 Cold and ECR-R Attachment avoidance), having limited social skills (EQ Social skills), feeling uncomfortable and nervous in social interactions (SPQ Social anxiety), or not feeling rewarded by it (SPQ Social anhedonia). Interestingly, scales that pertain to problems with shyness (OPD-SQ Establishing contact and IIP-32 Socially inhibited) or to making use of social contacts (OPD-SQ Accepting help) appeared to be an interstitial feature between detachment and internalizing. In a similar vein, Ringwald et al. (2021) found that avoidant PD and social phobia precisely reflected this blend, which fits well with our placement of shyness scales, as well as the placement of the shyness scale of the MMPI-3 in HiTOP (Sellbom et al., 2021).

Regarding the disinhibition spectrum of HiTOP, we found pure markers to be scales of impulsiveness (e.g., BIS-11 Non-planning impulsiveness, BIS-11 Motor

impulsiveness, BIS-11 Attentional impulsiveness, OPD-SQ Impulse control, APSD Impulsiveness, and AFI Authorship), substance use (e.g., DASI Drugs and alcohol), impulsive self-directed and other-directed aggression (PAI-BOR Self-harm, LHA Aggression, and OPD-SQ Impulse control), and increased willingness to take risks (APSD Impulsiveness). We found two blends that characterized a combination of disinhibition and internalizing; however, these were only represented by one scale each. A measure of acting impulsively under the influence of negative emotions (also: negative urgency; DERS Impulse control) was specifically related to high disinhibition and high internalizing, whereas sensation-seeking was indicative of high disinhibition and low internalizing (PAI-ANT Sensation-seeking). The most complex pattern of results was observed for the antagonism domain and its various blends. Pure markers of antagonism tapped into willfully ignoring others' feelings and needs (APSD Callousness, SRPS Callousness, IIP-32 Vindictive, and IIP-32 Overly accommodating reversed), caring a lot about oneself instead (APSD Narcissism and IIP-32 Domineering), and having a hostile attitude towards others (BSI Hostility). By contrast, conduct problems such as illegal activities or getting into troubles at work or in school were indicative of interstitial antagonism–disinhibition (PAI-ANT Antisociality, LHA Antisocial behavior, OPD-SQ Balancing interests, and SRPS Antisocial), and scales of cognitive empathy (EQ Cognitive empathy and RFQ-18 Mentalizing others) were placed between detachment and antagonism. The blend of high antagonism and low internalizing was represented by scales of affective empathy (EQ Emotional reactivity and OPD-SQ Empathy) as well as various scales that tap into being egocentric (PAI-ANT Egocentricity, SRPS Egocentricity, and IIP-32 Self-sacrificing reversed). However, what distinguishes pure markers of antagonism from interstitial markers of low-internalizing–high-antagonism seems hard to grasp. We suggest that the latter scales might tap into what the literature on psychopathy refers to as boldness/fearless dominance (for a meta-analysis, see Sleep et al., 2019), which is a construct related to narcissism and dominance-seeking (high antagonism) but also emotional stability (low internalizing).

Limitations

Some limitations of the current study should be considered. First, even though the used item pool is arguably among the more extensive collections of self-reports on psychopathology, some aspects were underrepresented (e.g., somatoform and obsessive–compulsive) or were not assessed at all (e.g., mania, eating pathology, and sexual

problems). Second, the use of extreme groups in sampling (e.g., including both community participants and outpatients) likely bloats the saturation of the p factor in terms of inflating the magnitude of its associations (Fisher et al., 2020), yet we have no reason to believe that it influences their pattern (i.e., sizes of the effects relative to each other). Third, further characteristics of the sample (i.e., oversampling of individuals with pronounced personality pathology) may hamper generalization to other samples. Fourth, although we made the HiTOP factors statistically independent from the predicted scales to avoid inflated associations, there might be additional sources of criterion contamination that we could not control given limitations of the study design (e.g., common method bias). Fifth, when this study was conducted, we relied on the then current version of the HiTOP working model as outlined in Kotov et al. (2017), but the model is subject to ongoing revisions (e.g., Kotov et al., 2020; Krueger et al., 2021; Watson, Levin-Aspenson, et al., 2022). Future studies should replicate our analysis using the official HiTOP measure (e.g., see Simms et al., 2022) once it becomes available. This would allow the analysis to be performed with truly separate scales that would further reduce the risk of criterion contamination. Sixth, we assumed unidimensional measurement models for all established (sub)scales and tested model fit, but we did not further explore misspecifications.

Future directions and practical recommendations

Our study has several implications. First, whenever the aim is to study narrow clinical constructs, we advise researchers to conduct a broad assessment of psychopathology that taps into different hierarchical levels of HiTOP. Using this approach, the meaning and validity of constructs can be better established, specific associations can be studied (i.e., beyond higher-level psychopathology dimensions), jingle-jangle fallacies can be better identified (Lawson & Robins, 2021), and, finally, the treatment utility of clinical assessments may be enhanced (Kamphuis et al., 2021). Currently, an omnibus measure of the HiTOP model is under development (Simms et al., 2022) with initial results being published for preliminary scales of HiTOP spectra (Cicero et al., 2022; Mullins-Sweatt et al., 2022; Sellbom et al., 2022; Watson, Forbes, et al., 2022; Zimmermann et al., 2022). Until the instrument becomes available (and most likely beyond), researchers will need to rely on existing measures that capture psychopathology broadly and are compatible with HiTOP. Alternatively, from the scales included here, some are pure markers that could be used as proxies to

operationalize HiTOP spectra. For example, the SPQ offers multiple pure marker scales of HiTOP dimensions: SPQ Social anhedonia scale could be used as a proxy to assess the detachment spectrum, SPQ Unusual beliefs for thought disorder, and SPQ Mistrust for the p factor. Slightly better, however, would be to approximate HiTOP spectra with multiple proxy scales. Yet in the absence of a truly comprehensive HiTOP measure that should exert higher fidelity in assessing higher-level psychopathology dimensions, inferences with improvised HiTOP measures will be limited but necessary.

The bifactor-(S-1) model offers advantages for modeling multiple levels of the psychopathology hierarchy (Eid et al., 2017; Heinrich et al., 2021), but it requires the specification of reference indicators that instantiate the p factor a priori. Unfortunately, there is little consensus about the meaning of the p factor. In this study, we circumvented this issue by parceling across HiTOP spectra to define the p factor, making use of the sheer mass of items included in this study. However, this is no parsimonious solution to define the p factor in future studies. Our results provide some support for the hypothesis that scales assessing impairments in mentalizing and epistemic trust may be pure markers of the p factor (Fonagy et al., 2021; Fonagy & Campbell, 2021), whereas other candidate constructs that have been proposed (e.g., emotional dysregulation and negative self-evaluation; Smith et al., 2020) were specific to spectra (e.g., internalizing). Although more evidence about the generalizability will be needed to corroborate these results, this raises some optimism that the p factor can be separately identified with selected transdiagnostic constructs. If pure markers (rather than just strong markers) of the p factor could thus be repeatedly identified with reasonable consistency and across different samples, these could be used to define the p factor in bifactor-(S-1) models.

CONCLUSION

Research has documented how symptoms of psychopathology tend to co-occur between individuals. As a result of synthesizing this literature, the HiTOP model proposes a hierarchical system of psychopathology including the p factor and several spectra (i.e., internalizing, thought disorder, detachment, antagonism, and disinhibition) that have exhibited strong validity (e.g., Kotov et al., 2020; Krueger et al., 2021; Watson, Levin-Aspenson, et al., 2022). Herein, we have reported results that help to understand which (sub)scales of established psychopathology questionnaires (a) are pure markers of HiTOP spectra, (b) are pure markers of the p factor, (c) reflect

blends of HiTOP spectra, (d) or—in contrast—do not map onto HiTOP at all. This can enable researchers to form richer and more distinct interpretations of the constructs measured and it facilitates the cumulative integration of various clinical traditions that rely on different conceptualizations and assessments of psychopathology (e.g., OPD-SQ originating from psychodynamic theory) but can be traced into HiTOP as an organizing framework.

ACKNOWLEDGEMENTS

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ETHICS STATEMENT

Ethics committee approval was obtained for data collection (Research Ethics Committee Wales 12/WA/0283). Informed consent was obtained from all participants.

DATA AVAILABILITY STATEMENT

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. R code for reproducing the analyses is permanently and openly accessible at <https://osf.io/hkav3/>. The data are available from the corresponding author on reasonable request.

ORCID

Leon P. Wendt  <https://orcid.org/0000-0003-2229-2860>

Kristin Jankowsky  <https://orcid.org/0000-0002-4847-0760>

Ulrich Schroeders  <https://orcid.org/0000-0002-5225-1122>

Tobias Nolte  <https://orcid.org/0000-0002-6834-7727>

Peter Fonagy  <https://orcid.org/0000-0003-0229-0091>

Johannes Zimmermann  <https://orcid.org/0000-0001-6975-2356>

Gabriel Olaru  <https://orcid.org/0000-0002-7430-7350>

ENDNOTES

¹ We also used a combination of data-driven methods that was recently recommended for dimensionality analysis (i.e., Hull Method, Empirical Kaiser Criterion, traditional parallel analysis, and sequential χ^2 model tests; Auerswald & Moshagen, 2019). However, the methods did not converge on an optimal number of factors (see Table S3 for details) so we considered them to a lesser extent and relied more heavily on substantive considerations.

² Standardized factor loadings denote associations between indicators (e.g., PD diagnoses) and extracted factors (e.g., HiTOP spectra) that are usually rotated towards simple structure

(e.g., geomin rotation). By contrast, in our study, PD diagnoses and HiTOP spectra are each measured independently, so that their association is estimated using the correlation coefficient.

³ Of note, there is the possibility of using another analytic approach. When predicting a scale, all items of the questionnaire from which the criterion scale is taken can be excluded from the HiTOP factors (i.e., not only the items of the criterion scale). This approach could be considered even more conservative in avoiding inflated associations by controlling method variance associated with the specific characteristics of a questionnaire (e.g., number or labels of response options). However, it also has significant shortcomings (i.e., reduced construct coverage), which is why we report this analysis in the supplement (see Note S1) and do not consider it further in the remainder of this article.

REFERENCES

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing. <https://doi.org/10.1176/appi.books.9780890425596>
- Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods, 24*(4), 468–491. <https://doi.org/10.1037/met0000200>
- Bainbridge, T. F., Ludeke, S. G., & Smillie, L. D. (2022). Evaluating the Big Five as an organizing framework for commonly used psychological trait scales. *Journal of Personality and Social Psychology, 122*, 749–777. <https://doi.org/10.1037/pspp0000395>
- Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders, 34*(2), 163–175. <https://doi.org/10.1023/B:JADD.0000022607.19833.00>
- Bernstein, E. M., & Putnam, F. W. (1986). Development, reliability, and validity of a dissociation scale. *The Journal of Nervous and Mental Disease, 174*(12), 727–735. <https://doi.org/10.1097/00005053-198612000-00004>
- Blanchard, E. B., Jones-Alexander, J., Buckley, T. C., & Forneris, C. A. (1996). Psychometric properties of the PTSD Checklist. *Behaviour Research and Therapy, 34*(8), 669–673. [https://doi.org/10.1016/0005-7967\(96\)00033-2](https://doi.org/10.1016/0005-7967(96)00033-2)
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305–314. <https://doi.org/10.1037/0033-2909.110.2.305>
- Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: Appreciating benefits and limitations. *Biological Psychiatry, 88*(1), 18–27. <https://doi.org/10.1016/j.biopsych.2020.01.013>
- Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9*, 91–121. <https://doi.org/10.1146/annurev-clinpsy-050212-185608>
- Bringmann, L. F., Albers, C., Bockting, C., Borsboom, D., Ceulemans, E., Cramer, A., Epskamp, S., Eronen, M. I., Hamaker, E., Kuppens, P., Lutz, W., McNally, R. J., Molenaar, P., Tio, P., Voelkle, M. C., & Wichers, M. (2021). Psychopathological networks: Theory, methods and practice.

- Behaviour Research and Therapy*, 149, 104011. <https://doi.org/10.1016/j.brat.2021.104011>
- Caspi, A., Houts, R. M., Ambler, A., Danese, A., Elliott, M. L., Hariri, A., Harrington, H. L., Hogan, S., Poulton, R., Ramrakha, S., Rasmussen, L. J. H., Reuben, A., Richmond-Rakerd, L., Sugden, K., Wertz, J., Williams, B. S., & Moffitt, T. E. (2020). Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the Dunedin birth cohort study. *JAMA Network Open*, 3(4), e203221. <https://doi.org/10.1001/jamanetworkopen.2020.3221>
- Caspi, A., & Moffitt, T. E. (2018). All for one and one for all: Mental disorders in one dimension. *American Journal of Psychiatry*, 175(9), 831–844. <https://doi.org/10.1176/appi.ajp.2018.17121383>
- Cicero, D. C., Jonas, K. G., Chmielewski, M., Martin, E. A., Docherty, A. R., Berzon, J., Haltigan, J. D., Reininghaus, U., Caspi, A., Grazioplene, R. G., & Kotov, R. (2022). Development of the thought disorder measure for the Hierarchical Taxonomy of Psychopathology. *Assessment*, 29(1), 46–61. <https://doi.org/10.1177/10731911211015355>
- Coccaro, E. F., Berman, M. E., & Kavoussi, R. J. (1997). Assessment of life history of aggression: Development and psychometric characteristics. *Psychiatry Research*, 73(3), 147–157. [https://doi.org/10.1016/S0165-1781\(97\)00119-4](https://doi.org/10.1016/S0165-1781(97)00119-4)
- Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, 34(3), 315–346. https://doi.org/10.1207/S15327906MBR3403_2
- Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology*, 104(10), 1243–1265. <https://doi.org/10.1037/apl0000406>
- Constantinou, M., & Fonagy, P. (2019). Evaluating bifactor models of psychopathology using model-based reliability indices. <https://doi.org/10.31234/osf.io/6tf7j>
- Conway, C. C., Forbes, M. K., Forbush, K. T., Fried, E. I., Hallquist, M. N., Kotov, R., Mullins-Sweatt, S. N., Shackman, A. J., Skodol, A. E., South, S. C., Sunderland, M., Waszczuk, M. A., Zald, D. H., Afzali, M. H., Bornovalova, M. A., Carragher, N., Docherty, A. R., Jonas, K. G., Krueger, R. F., ... Eaton, N. R. (2019). A Hierarchical Taxonomy of Psychopathology can transform mental health research. *Perspectives on Psychological Science*, 14(3), 419–436. <https://doi.org/10.1177/1745691618810696>
- Derogatis, L. R., & Spencer, P. M. (1993). *Brief Symptom Inventory: BSI*. Pearson.
- DeYoung, C. G., Chmielewski, M., Clark, L. A., Condon, D. M., Kotov, R., Krueger, R. F., Lynam, D. R., Markon, K. E., Miller, J. D., Mullins-Sweatt, S. N., Samuel, D. B., Sellbom, M., South, S. C., Thomas, K. M., Watson, D., Watts, A. L., Widiger, T. A., Wright, A. G. C., & the HiTOP Normal Personality Workgroup. (2020). The distinction between symptoms and traits in the Hierarchical Taxonomy of Psychopathology (HiTOP). *Journal of Personality*, 90, 20–33. <https://doi.org/10.1111/jopy.12593>
- Ehrental, J. C., Dinger, U., Horsch, L., Komo-Lang, M., Klinkerfuß, M., Grande, T., & Schauenburg, H. (2012). Der OPD-Strukturfragebogen (OPD-SF): Erste Ergebnisse zu Reliabilität und Validität [The OPD structure questionnaire OPD-SQ. *First results on reliability and validity*]. *Psychosomatik, Psychotherapie, Medizinische Psychologie*, 62(1), 25–32. <https://doi.org/10.1055/s-0031-1295481>
- Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, 22(3), 541–562. <https://doi.org/10.1037/met0000083>
- Euler, S., Nolte, T., Constantinou, M., Griem, J., Montague, P. R., Fonagy, P., & Personality and Mood Disorders Research Network. (2019). Interpersonal problems in borderline personality disorder: Associations with mentalizing, emotion regulation, and impulsiveness. *Journal of Personality Disorders*, 35, 177–193. https://doi.org/10.1521/pedi_2019_33_427
- First, M. B., & Gibbon, M. (2004). The Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) and the Structured Clinical Interview for DSM-IV Axis II Disorders (SCID-II). In M. J. Hilsenroth & D. L. Segal (Eds.), *Comprehensive handbook of psychological assessment, Vol. 2. Personality assessment* (pp. 134–143). John Wiley & Sons, Inc.
- Fisher, J. E., Guha, A., Heller, W., & Miller, G. A. (2020). Extreme-groups designs in studies of dimensional phenomena: Advantages, caveats, and recommendations. *Journal of Abnormal Psychology*, 129(1), 14–20. <https://doi.org/10.1037/abn0000480>
- Fleiss, J. L., & Shrout, P. E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika*, 43(2), 259–262. <https://doi.org/10.1007/BF02293867>
- Fonagy, P., & Campbell, C. (2021). Future directions in personality pathology. *Current Opinion in Psychology*, 37, 145–151. <https://doi.org/10.1016/j.copsyc.2021.01.001>
- Fonagy, P., Campbell, C., Constantinou, M., Higgitt, A., Allison, E., & Luyten, P. (2021). Culture and psychopathology: An attempt at reconsidering the role of social learning. *Development and Psychopathology*. Advance online publication. <https://doi.org/10.1017/S0954579421000092>
- Forbes, M. K., Sunderland, M., Rapee, R. M., Batterham, P. J., Calear, A. L., Carragher, N., Ruggero, C., Zimmerman, M., Baillie, A. J., Lynch, S. J., Mewton, L., Slade, T., & Krueger, R. F. (2021). A detailed hierarchical model of psychopathology: From individual symptoms up to the general factor of psychopathology. *Clinical Psychological Science*, 9(2), 139–168. <https://doi.org/10.1177/2167702620954799>
- Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology*, 78(2), 350–365. <https://doi.org/10.1037/0022-3514.78.2.350>
- Frick, P. J., & Hare, R. D. (2001). *Antisocial Process Screening Device: APSD*. Multi-Health Systems.
- Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, 208, 191–197. <https://doi.org/10.1016/j.jad.2016.10.019>
- Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, 31(4), 271–288. <https://doi.org/10.1080/1047840X.2020.1853461>
- Fried, E. I., Greene, A. L., & Eaton, N. R. (2021). The p factor is the sum of its parts, for now. *World Psychiatry*, 20(1), 69–70. <https://doi.org/10.1002/wps.20814>

- Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78. <https://doi.org/10.1016/j.paid.2016.06.069>
- Goss, K., Gilbert, P., & Allan, S. (1994). An exploration of shame measures—I: The other as Shamer scale. *Personality and Individual Differences, 17*(5), 713–717. [https://doi.org/10.1016/0191-8869\(94\)90149-X](https://doi.org/10.1016/0191-8869(94)90149-X)
- Gratz, K. L., & Roemer, L. (2004). Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the difficulties in emotion regulation scale. *Journal of Psychopathology and Behavioral Assessment, 26*(1), 41–54. <https://doi.org/10.1023/B:JOBA.0000007455.08539.94>
- Green, C. E. L., Freeman, D., Kuipers, E., Bebbington, P., Fowler, D., Dunn, G., & Garety, P. A. (2008). Measuring ideas of persecution and social reference: The Green et al. Paranoid Thought Scales (GPTS). *Psychological Medicine, 38*(1), 101–111. <https://doi.org/10.1017/S0033291707001638>
- Haefel, G. J., Jeronimus, B. F., Kaiser, B. N., Weaver, L. J., Soyster, P. D., Fisher, A. J., Vargas, I., Goodson, J. T., & Lu, W. (2021). Folk classification and factor rotations: Whales, sharks, and the problems with the Hierarchical Taxonomy of Psychopathology (HiTOP). *Clinical Psychological Science*. Advance online publication, 10, 259–278. <https://doi.org/10.1177/21677026211002500>
- Heinrich, M., Geiser, C., Zagorscak, P., Burns, G. L., Bohn, J., Becker, S. P., Eid, M., Beauchaine, T. P., & Knaevelsrud, C. (2021). On the meaning of the “P factor” in symmetrical bifactor models of psychopathology: Recommendations for future research from the bifactor-(S-1) perspective. *Assessment*. Advance online publication. <https://doi.org/10.1177/10731911211060298>
- Hopwood, C. J., Bagby, R. M., Gralnick, T., Ro, E., Ruggero, C., Mullins-Sweatt, S., Kotov, R., Bach, B., Cicero, D. C., Krueger, R. F., Patrick, C. J., Chmielewski, M., DeYoung, C. G., Docherty, A. R., Eaton, N. R., Forbush, K. T., Ivanova, M. Y., Latzman, R. D., Pincus, A. L., ... Zimmermann, J. (2020). Integrating psychotherapy with the hierarchical taxonomy of psychopathology (HiTOP). *Journal of Psychotherapy Integration, 30*(4), 477–497. <https://doi.org/10.1037/int0000156>
- Horowitz, L. M., Alden, L. E., Wiggins, J. S., & Pincus, A. L. (2000). *IIP-64/IIP-32 professional manual*. The Psychological Corporation.
- Huang, Y. L., Fonagy, P., Feigenbaum, J., Montague, P. R., Nolte, T., & London Personality and Mood Disorder Consortium. (2020). Multidirectional pathways between attachment, mentalizing, and posttraumatic stress symptomatology in the context of childhood trauma. *Psychopathology, 53*(1), 48–58. <https://doi.org/10.1159/000506406>
- Kamphuis, J. H., Noordhof, A., & Hopwood, C. J. (2021). When and how assessment matters: An update on the Treatment Utility of Clinical Assessment (TUCA). *Psychological Assessment, 33*(2), 122–132. <https://doi.org/10.1037/pas0000966>
- Kelley, T. L. (1927). *Interpretation of educational measurements*. Macmillan.
- Kotov, R., Jonas, K. G., Carpenter, W. T., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, K. T., Hobbs, K., Reininghaus, U., Slade, T., South, S. C., Sunderland, M., Waszczuk, M. A., Widiger, T. A., Wright, A. G. C., Zald, D. H., Krueger, R. F., Watson, D., & HiTOP Utility Workgroup. (2020). Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): I. Psychosis superspectrum. *World Psychiatry, 19*(2), 151–172. <https://doi.org/10.1002/wps.20730>
- Kotov, R., Krueger, R. F., & Watson, D. (2018). A paradigm shift in psychiatric classification: The Hierarchical Taxonomy Of Psychopathology (HiTOP). *World Psychiatry, 17*(1), 24–25. <https://doi.org/10.1002/wps.20478>
- Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., ... Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology, 126*(4), 454–477. <https://doi.org/10.1037/abn0000258>
- Kotov, R., Krueger, R. F., Watson, D., Cicero, D. C., Conway, C. C., DeYoung, C. G., Eaton, N. R., Forbes, M. K., Hallquist, M. N., Latzman, R. D., Mullins-Sweatt, S. N., Ruggero, C. J., Simms, L. J., Waldman, I. D., Waszczuk, M. A., & Wright, A. G. C. (2021). The Hierarchical Taxonomy of Psychopathology (HiTOP): A quantitative nosology based on consensus of evidence. *Annual Review of Clinical Psychology, 17*, 83–108. <https://doi.org/10.1146/annurev-clinpsy-081219-093304>
- Krueger, R. F., Hobbs, K. A., Conway, C. C., Dick, D. M., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, K. T., Keyes, K. M., Latzman, R. D., Michelini, G., Patrick, C. J., Sellbom, M., Slade, T., South, S. C., Sunderland, M., Tackett, J., Waldman, I., Waszczuk, M. A., ... HiTOP Utility Workgroup. (2021). Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): II. Externalizing superspectrum. *World Psychiatry, 20*(2), 171–193. <https://doi.org/10.1002/wps.20844>
- Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., Simms, L. J., Widiger, T. A., Achenbach, T. M., Bach, B., Bagby, R. M., Bornovalova, M. A., Carpenter, W. T., Chmielewski, M., Cicero, D. C., Clark, L. A., Conway, C., DeClercq, B., DeYoung, C. G., ... Zimmermann, J. (2018). Progress in achieving quantitative classification of psychopathology. *World Psychiatry, 17*(3), 282–293. <https://doi.org/10.1002/wps.20566>
- Lahey, B. B., Moore, T. M., Kaczurkin, A. N., & Zald, D. H. (2021). Hierarchical models of psychopathology: Empirical support, implications, and remaining issues. *World Psychiatry, 20*(1), 57–63. <https://doi.org/10.1002/wps.20824>
- Lawson, K. M., & Robins, R. W. (2021). Sibling constructs: What are they, why do they matter, and how should you handle them? *Personality and Social Psychology Review, 25*(4), 344–366. <https://doi.org/10.1177/10888683211047101>
- Leising, D., Burger, J., Zimmermann, J., Bäckström, M., Oltmanns, J. R., & Connelly, B. S. (2020). Why do items correlate with one another? A conceptual analysis with relevance for general factors and network models. <https://doi.org/10.31234/osf.io/7c895>
- Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology, 68*(1), 151–158. <https://doi.org/10.1037//0022-3514.68.1.151>

- Levin-Aspenson, H. F., Watson, D., Clark, L. A., & Zimmerman, M. (2021). What is the general factor of psychopathology? Consistency of the p factor across samples. *Assessment, 28*(4), 1035–1049. <https://doi.org/10.1177/1073191120954921>
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*(2), 151–173. https://doi.org/10.1207/S15328007SEM0902_1
- Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods, 18*(3), 285–300. <https://doi.org/10.1037/a0033266>
- Mansolf, M., & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate Behavioral Research, 51*(5), 698–717. <https://doi.org/10.1080/00273171.2016.1215898>
- Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin, 137*(5), 856–879. <https://doi.org/10.1037/a0023678>
- McCabe, G. A., Oltmanns, J. R., & Widiger, T. A. (2022). The general factors of personality disorder, psychopathology, and personality. *Journal of Personality Disorders, 36*(2), 129–156. https://doi.org/10.1521/pedi_2021_35_530
- Moran, P., Leese, M., Lee, T., Walters, P., Thornicroft, G., & Mann, A. (2003). Standardised Assessment of Personality–Abbreviated Scale (SAPAS): Preliminary validation of a brief screen for personality disorder. *The British Journal of Psychiatry, 183*(3), 228–232. <https://doi.org/10.1192/bjp.183.3.228>
- Morey, L. C. (2014). *The Personality Assessment Inventory*. Routledge/Taylor & Francis Group.
- Moshagen, M. (2021). When a truly positive correlation turns negative: How different approaches to model hierarchically structured constructs affect estimated correlations to covariates. *European Journal of Personality*. Advance online publication. <https://doi.org/10.1177/08902070211050170>
- Müller, S., Wendt, L. P., Spitzer, C., Masuhr, O., Back, S. N., & Zimmermann, J. (2022). A critical evaluation of the Reflective Functioning Questionnaire (RFQ). *Journal of Personality Assessment, 104*(5), 613–627. <https://doi.org/10.1080/00223891.2021.1981346>
- Mullins-Sweatt, S. N., Bornoalova, M. A., Carragher, N., Clark, L. A., Corona Espinosa, A., Jonas, K., Keyes, K. M., Lynam, D. R., Michelini, G., Miller, J. D., Min, J., Rodriguez-Seijas, C., Samuel, D. B., Tackett, J. L., & Watts, A. L. (2022). HiTOP assessment of externalizing antagonism and disinhibition. *Assessment, 29*(1), 34–45. <https://doi.org/10.1177/10731911211033900>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.
- Orben, A., & Lakens, D. (2020). Crud (re)defined. *Advances in Methods and Practices in Psychological Science, 3*(2), 238–247. <https://doi.org/10.1177/2515245920917961>
- Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology, 51*(6), 768–774. [https://doi.org/10.1002/1097-4679\(199511\)51:6](https://doi.org/10.1002/1097-4679(199511)51:6)
- Petterson, E., Mendle, J., Turkheimer, E., Horn, E. E., Ford, D. C., Simms, L. J., & Clark, L. A. (2014). Do maladaptive behaviors exist at one or both ends of personality traits? *Psychological Assessment, 26*(2), 433–446. <https://doi.org/10.1037/a0035587>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Raine, A. (1991). The SPQ: A scale for the assessment of schizotypal personality based on DSM-III-R criteria. *Schizophrenia Bulletin, 17*(4), 555–564. <https://doi.org/10.1093/schbul/17.4.555>
- Rhemtulla, M. (2016). Population performance of SEM parceling strategies under measurement and structural model misspecification. *Psychological Methods, 21*(3), 348–368. <https://doi.org/10.1037/met0000072>
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. <https://doi.org/10.1037/a0029315>
- Rifkin-Zybutz, R. P., Moran, P., Nolte, T., Feigenbaum, J., King-Casas, B., London Personality and Mood Disorder Consortium, Fonagy, P., & Montague, R. P. (2021). Impaired mentalizing in depression and the effects of borderline personality disorder on this relationship. *Borderline Personality Disorder and Emotion Dysregulation, 8*(1), 15. <https://doi.org/10.1186/s40479-021-00153-x>
- Ringwald, W. R., Forbes, M. K., & Wright, A. G. (2021). Meta-analysis of structural evidence for the Hierarchical Taxonomy of Psychopathology (HiTOP) model. *Psychological Medicine*. Advance online publication. <https://doi.org/10.1017/S0033291721001902>
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. <https://doi.org/10.1037/met0000045>
- Rogoff, S., Moulton-Perkins, A., Warren, F., Nolte, T., & Fonagy, P. (2021). 'Rich' and 'poor' in mentalizing: Do expert mentalizers exist? *PLoS ONE, 16*(10), e0259030. <https://doi.org/10.1371/journal.pone.0259030>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Ruggero, C. J., Kotov, R., Hopwood, C. J., First, M., Clark, L. A., Skodol, A. E., Mullins-Sweatt, S. N., Patrick, C. J., Bach, B., Cicero, D. C., Docherty, A., Simms, L. J., Bagby, R. M., Krueger, R. F., Callahan, J. L., Chmielewski, M., Conway, C. C., de Clercq, B., Dornbach-Bender, A., ... Zimmermann, J. (2019). Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into clinical practice. *Journal of Consulting and Clinical Psychology, 87*(12), 1069–1084. <https://doi.org/10.1037/ccp0000452>
- Sellbom, M., Carragher, N., Sunderland, M., Calear, A. L., & Batterham, P. J. (2020). The role of maladaptive personality domains across multiple levels of the HiTOP structure. *Personality and Mental Health, 14*(1), 30–50. <https://doi.org/10.1002/pmh.1461>
- Sellbom, M., Forbush, K. T., Gould, S. R., Markon, K. E., Watson, D., & Witthöft, M. (2022). HiTOP assessment of the somatoform spectrum and eating disorders. *Assessment, 29*(1), 62–74. <https://doi.org/10.1177/10731911211020825>

- Sellbom, M., Kremyar, A. J., & Wygant, D. B. (2021). Mapping MMPI-3 scales onto the hierarchical taxonomy of psychopathology. *Psychological Assessment*, 33(12), 1153–1168. <https://doi.org/10.1037/pas0001049>
- Simms, L. J., Wright, A. G., Cicero, D., Kotov, R., Mullins-Sweatt, S. N., Sellbom, M., Watson, D., Widiger, T. A., & Zimmermann, J. (2022). Development of measures for the Hierarchical Taxonomy of Psychopathology (HiTOP): A collaborative scale development project. *Assessment*, 29(1), 3–16. <https://doi.org/10.1177/10731911211015309>
- Sleep, C. E., Weiss, B., Lynam, D. R., & Miller, J. D. (2019). An examination of the Triarchic Model of psychopathy's nomological network: A meta-analytic review. *Clinical Psychology Review*, 71, 1–26. <https://doi.org/10.1016/j.cpr.2019.04.005>
- Smith, G. T., Atkinson, E. A., Davis, H. A., Riley, E. N., & Oltmanns, J. R. (2020). The general factor of psychopathology. *Annual Review of Clinical Psychology*, 16(1), 75–98. <https://doi.org/10.1146/annurev-clinpsy-071119-115848>
- Somma, A., Krueger, R. F., Markon, K. E., & Fossati, A. (2019). The replicability of the personality inventory for DSM-5 domain scale factor structure in US and non-US samples: A quantitative review of the published literature. *Psychological Assessment*, 31(7), 861–877. <https://doi.org/10.1037/pas0000711>
- Stanton, K., McDonnell, C. G., Hayden, E. P., & Watson, D. (2020). Transdiagnostic approaches to psychopathology measurement: Recommendations for measure selection, data analysis, and participant recruitment. *Journal of Abnormal Psychology*, 129(1), 21–28. <https://doi.org/10.1037/abn0000464>
- Van IJzendoorn, M., & Schuengel, C. (1996). The measurement of dissociation in normal and clinical populations: Meta-analytic validation of the Dissociative Experiences Scale (DES). *Clinical Psychology Review*, 16(5), 365–382. [https://doi.org/10.1016/0272-7358\(96\)00006-2](https://doi.org/10.1016/0272-7358(96)00006-2)
- Waszczuk, M. A., Eaton, N. R., Krueger, R. F., Shackman, A. J., Waldman, I. D., Zald, D. H., Lahey, B. B., Patrick, C. J., Conway, C. C., Ormel, J., Hyman, S. E., Fried, E. I., Forbes, M. K., Docherty, A. R., Althoff, R. R., Bach, B., Chmielewski, M., DeYoung, C. G., Forbush, K. T., ... Kotov, R. (2020). Redefining phenotypes to advance psychiatric genetics: Implications from hierarchical taxonomy of psychopathology. *Journal of Abnormal Psychology*, 129(2), 143–161. <https://doi.org/10.1037/abn0000486>
- Watson, D., Forbes, M. K., Levin-Aspenson, H. F., Ruggero, C. J., Kotelnikova, Y., Khoo, S., Bagby, R. M., Sunderland, M., Patalay, P., & Kotov, R. (2022). The development of preliminary HiTOP internalizing spectrum scales. *Assessment*, 29(1), 17–33. <https://doi.org/10.1177/10731911211003976>
- Watson, D., Levin-Aspenson, H. F., Waszczuk, M. A., Conway, C. C., Dalgleish, T., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, K. T., Hobbs, K. A., Michelini, G., Nelson, B. D., Sellbom, M., Slade, T., South, S. C., Sunderland, M., Waldman, I., Witthöft, M., Wright, A. G. C., ... Zinbarg, R. E. (2022). Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): III. Emotional dysfunction superspectrum. *World Psychiatry*, 21(1), 26–54. <https://doi.org/10.1002/wps.20943>
- Watts, A. L., Boness, C. L., Loeffelman, J. E., Steinley, D., & Sher, K. J. (2021). Does crude measurement contribute to observed unidimensionality of psychological constructs? A demonstration with DSM-5 alcohol use disorder. *Journal of Abnormal Psychology*, 130(5), 512–524. <https://doi.org/10.1037/abn0000678>
- Watts, A. L., Lane, S. P., Bonifay, W., Steinley, D., & Meyer, F. A. (2020). Building theories on top of, and not independent of, statistical models: The case of the p-factor. *Psychological Inquiry*, 31(4), 310–320. <https://doi.org/10.1080/1047840X.2020.1853476>
- Watts, A. L., Makol, B. A., Palumbo, I. M., de Los Reyes, A., Olino, T. M., Latzman, R. D., DeYoung, C. G., Wood, P. K., & Sher, K. J. (2022). How robust is the p factor? Using multitrait-multimethod modeling to inform the meaning of general factors of youth psychopathology. *Clinical Psychological Science*. Advance online, 10, 640–661. <https://doi.org/10.1177/21677026211055170>
- Weinstein, N., Przybylski, A. K., & Ryan, R. M. (2012). The index of autonomous functioning: Development of a scale of human autonomy. *Journal of Research in Personality*, 46(4), 397–413. <https://doi.org/10.1016/j.jrp.2012.03.007>
- Wendt, L. P., Wright, A. G. C., Pilkonis, P. A., Nolte, T., Fonagy, P., Montague, P. R., & Zimmermann, J. (2019). The latent structure of interpersonal problems: Validity of dimensional, categorical, and hybrid models. *Journal of Abnormal Psychology*, 128(8), 823–839. <https://doi.org/10.1037/abn0000460>
- Widiger, T. A., Bach, B., Chmielewski, M., Clark, L. A., DeYoung, C., Hopwood, C. J., Kotov, R., Krueger, R. F., Miller, J. D., Morey, L. C., Mullins-Sweatt, S. N., Patrick, C. J., Pincus, A. L., Samuel, D. B., Sellbom, M., South, S. C., Tackett, J. L., Watson, D., Waugh, M. H., ... Thomas, K. M. (2019). Criterion A of the AMPD in HiTOP. *Journal of Personality Assessment*, 101(4), 345–355. <https://doi.org/10.1080/00223891.2018.1465431>
- Wilkinson, P. O., Qiu, T., Neufeld, S., Jones, P. B., & Goodyer, I. M. (2018). Sporadic and recurrent non-suicidal self-injury before age 14 and incident onset of psychiatric disorders by 17 years: Prospective cohort study. *The British Journal of Psychiatry*, 212(4), 222–226. <https://doi.org/10.1192/bjp.2017.45>
- Williams, T. F., Scalco, M. D., & Simms, L. J. (2018). The construct validity of general and specific dimensions of personality pathology. *Psychological Medicine*, 48(5), 834–848. <https://doi.org/10.1017/S0033291717002227>
- Wright, A. G. C., Pincus, A., & Hopwood, C. J. (in press). Contemporary integrative interpersonal theory: Integrating structure, dynamics, temporal scale, and levels of analysis. *Journal of Abnormal Psychology*.
- Wright, A. G. C., & Simms, L. J. (2015). A metastructural model of mental disorders and pathological personality traits. *Psychological Medicine*, 45(11), 2309–2319. <https://doi.org/10.1017/S0033291715000252>
- Zhang, B., Sun, T., Cao, M., & Drasgow, F. (2021). Using bifactor models to examine the predictive validity of hierarchical constructs: Pros, cons, and solutions. *Organizational*

Research Methods, 24(3), 530–571. <https://doi.org/10.1177/1094428120915522>

Zimmermann, J., Widiger, T. A., Oeltjen, L., Conway, C. C., & Morey, L. C. (2022). Developing preliminary scales for assessing the HiTOP detachment spectrum. *Assessment*, 29(1), 75–87. <https://doi.org/10.1177/10731911211015313>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wendt, L. P., Jankowsky, K., Schroeders, U., London Personality and Mood Disorder Research Consortium, Nolte, T., Fonagy, P., Montague, P. R., Zimmermann, J., & Olaru, G. (2023). Mapping established psychopathology scales onto the Hierarchical Taxonomy of Psychopathology (HiTOP). *Personality and Mental Health*, 17(2), 117–134. <https://doi.org/10.1002/pmh.1566>

Article 3:

Mindreading Measures Misread? A Multimethod Investigation Into the Validity of Self-Report and Task-Based Approaches

Leon P. Wendt¹, Johannes Zimmermann¹, Carsten Spitzer², & Sascha Müller^{1,2}

¹ Department of Psychology, University of Kassel

² Department of Psychosomatic Medicine and Psychotherapy, Rostock University Medical Center,
University of Rostock

Status:

Published

Supplemental Materials:

R code, data, and other materials

<https://doi.org/10.17605/OSF.IO/Q7EU4>

Citation:

Wendt, L. P., Zimmermann, J., Spitzer, C., & Müller, S. (2024). Mindreading measures misread? A multimethod investigation into the validity of self-report and task-based approaches. *Psychological Assessment*, 36(5), 365-378. <https://doi.org/10.1037/pas0001310>

Mindreading Measures Misread? A Multimethod Investigation Into the Validity of Self-Report and Task-Based Approaches

Leon P. Wendt¹, Johannes Zimmermann¹, Carsten Spitzer², & Sascha Müller^{1,2}

¹ Department of Psychology, University of Kassel

² Department of Psychosomatic Medicine and Psychotherapy, Rostock University Medical Center, University of Rostock

Mindreading ability—also referred to as cognitive empathy or mentalizing—is typically conceptualized as a relatively stable dimension of individual differences in the ability to make accurate inferences about the mental states of others. This construct is primarily assessed using self-report questionnaires and task-based performance measures. However, the validity of these measures has been questioned: according to rival interpretations, mindreading tasks may capture general cognitive ability, whereas mindreading self-reports may capture perceived rather than actual mindreading ability. In this preregistered multimethod study involving 700 participants from the U.S. general population, we tested the validity of mindreading measures by examining the nomological network of self-reports and task-based methods using structural equation modeling. Specifically, we contrasted the empirical associations with theoretical predictions that assume mindreading measures are valid versus invalid. More consistent with rival interpretations, mindreading tasks showed a negligible latent correlation with mindreading self-reports (.05) and a large one with general cognitive ability (.85), whereas mindreading self-reports were specifically associated with perceived performance in mindreading tasks (.29). Also more consistent with rival interpretations, neither mindreading self-reports nor task-based measures showed positive unique associations with psychosocial functioning when controlling for general cognitive ability and general positive self-evaluation. Instead, negative unique associations emerged for both methods, although this effect was not robust for tasks. Overall, the results cast doubt on the validity of commonly used mindreading measures and support their rival interpretations.

Public Significance Statement

This study indicates that psychological tests designed to assess mindreading ability may lack validity in that mindreading tasks may largely capture general cognitive ability rather than specific mindreading ability, while self-report questionnaires on mindreading may primarily capture subjective self-perception.

Supplementary materials: <https://doi.org/10.17605/OSF.IO/Q7EU4>

Keywords: validity, cognitive empathy, mentalizing, self-report, task performance

Mindreading ability is conceptualized as a dimension of relatively stable individual differences, reflecting the ability to accurately recognize and interpret the mental states of others, such as their thoughts, feelings, and motivations. This ability is also described by terms such as cognitive empathy, mentalizing, perspective-taking, empathic accuracy, and theory of

mind (Olderbak & Wilhelm, 2020). A variety of measures have been developed to assess this construct, with self-report questionnaires and performance-based tasks being the predominant methods. There is a widespread assumption among researchers and practitioners that existing self-report and task-based measures of mindreading ability are valid. In other

Leon P. Wendt  <https://orcid.org/0000-0003-2229-2860>
Johannes Zimmermann  <https://orcid.org/0000-0001-6975-2356>
Carsten Spitzer  <https://orcid.org/0000-0002-2711-285X>
Sascha Müller  <https://orcid.org/0000-0002-8663-4543>

The authors have no conflicts of interest to disclose. We report how we determined all methodological and analytical decisions. The study's design, hypotheses, and analysis plan were preregistered before data were collected; see <https://doi.org/10.17605/OSF.IO/VQKXC>. The data and R code required to reproduce the analyses are publicly and permanently available in an Open Science Framework online repository and can be accessed at <https://doi.org/10.17605/OSF.IO/Q7EU4>.

Leon P. Wendt played a lead role in conceptualization, data curation, formal analysis, investigation, methodology, project administration, and writing—original draft. Johannes

Zimmermann played a lead role in supervision, a supporting role in conceptualization, and an equal role in funding acquisition and writing—review and editing. Carsten Spitzer played a supporting role in supervision and writing—review and editing, and an equal role in funding acquisition. Sascha Müller played a supporting role in conceptualization, methodology, and project administration, and an equal role in writing—review and editing.

This preprint was typeset using an adaptation of the template by Wiemik (2020), which can be accessed at <https://doi.org/10.17605/OSF.IO/HSV6A>.

Correspondence concerning this article should be addressed to Leon Wendt, Department of Psychology, University of Kassel, Holländische Str. 36-38, 34127, Kassel, Germany. Email: l.wendt@uni-kassel.de.

words, it is assumed, whether implicitly or explicitly, that their test scores capture mindreading ability without simultaneously reflecting unrelated traits or other contaminating influences. For instance, this assumption underlies recent meta-analyses that have used these tests to infer the genetic and environmental origins of mindreading ability (Abramson et al., 2020), or to elucidate its associations with other constructs, such as emotion regulation (Salazar Kämpf et al., 2023) and mental disorders (Bora, 2021; Johnson et al., 2022). However, some researchers have expressed skepticism regarding the validity of mindreading self-reports and tasks, fueled by evidence that these methods are largely unrelated and thus unlikely to capture a common underlying construct (e.g., Grainger et al., 2023; Müller et al., 2023; Murphy & Lillienfeld, 2019; Realo et al., 2003; Sunahara et al., 2022). In this multimethod study involving a large sample of the general U.S. population ($N = 700$), we use a nomological network approach to examine competing test score interpretations of mindreading self-reports and tasks (Cronbach & Meehl, 1955).

Task-based assessments of mindreading ability typically require participants to infer the mental states of fictional characters from materials such as videos, photographs, or written narratives. Critiques of specific tasks have raised concerns about potential problems such as inadequate representation of the full range of mental states (Oakley et al., 2016), misalignment with the specific cognitive processes involved in mindreading (e.g., Quesque & Rossetti, 2020), ceiling effects (Dodell-Feder et al., 2013), and reliance on expert consensus rather than objective, verifiable responses (Kittel et al., 2022). Broader concerns have been raised about the ecological validity of existing tasks in their ability to reflect the intricacies of real-world mindreading (Osborne-Crowley, 2020), as well as the potentially large overlap with measures of broader abilities, such as general cognitive ability and verbal ability (Coyle et al., 2018; Navarro, 2022). The latter two considerations can be seen as interrelated; if task demands do not require the use of specific mindreading skills, then broader cognitive abilities may account for a greater proportion of interindividual differences in task performance. Therefore, we put forward the rival hypothesis that commonly used mindreading tasks may reflect general cognitive ability rather than specific mindreading ability.

In self-reports of mindreading ability, individuals typically express their level of agreement with statements such as “I am good at reading other people’s minds.” However, some scholars assert that individuals may lack sufficient metacognitive insight into their mindreading ability, implying that self-reports may reflect subjective self-perceptions rather than objective competence (e.g., Realo et al., 2003). One hypothetical reason for this could be the paucity of external feedback available to individuals regarding their mindreading performance in daily life (Ickes, 1993). Following these considerations, we propose the rival hypothesis that self-reports are

more indicative of perceived mindreading ability than of actual ability, an interpretation that has previously been adopted by the *Certainty About Mental States Questionnaire* (CAMSQ; Müller et al., 2023). A theoretical framework through which perceived mindreading ability can be further characterized is provided by mentalizing theory, which posits that mental states are inherently elusive and that well-adjusted individuals are consequently those who recognize these complexities and maintain a modest appraisal of their mindreading ability (e.g., Bateman & Fonagy, 2019). Viewed through this lens, individuals who exhibit overconfidence in their mindreading ability may actually be more prone to making inaccurate judgments in interpersonal contexts (e.g., Sharp & Vanwoerden, 2015).

In this study, we employ several tests designed to test the plausibility of validity hypotheses against rival hypotheses and to adjudicate between their competing claims. Given the lack of a consensus gold standard for assessing mindreading ability, we use the nomological network approach by testing whether theoretical predictions about the construct are adequately reflected in the empirical patterns of the measures (Cronbach & Meehl, 1955). While predictions are commonly articulated only for the validity hypothesis, the utility of additionally considering plausible rival hypotheses was emphasized by Messick (1975) and is also mentioned in the current *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). This approach has the potential to provide more compelling validity evidence by demonstrating that a particular interpretation of a test score is not only consistent with the data, but *more* consistent with the data than plausible alternative interpretations.

Below, we describe the rationale for each validity test that assists in this endeavor, and Table 1 further specifies the statistical estimands as well as the predictions implied by the competing hypotheses. The first three validity tests are based on convergent and discriminant evidence, operating on the principle that measures of similar constructs should exhibit higher correlations than those of dissimilar constructs. The first validity test evaluates the heteromethod correlation between mindreading self-reports and tasks; here, a higher correlation would be more supportive of the validity hypotheses than the rival hypotheses. For the second validity test, we investigate the relationship between mindreading tasks and general cognitive ability tasks, expecting a correlation that should not be too high to remain consistent with the validity hypotheses. For the third validity test, we explore whether the correlation of mindreading self-reports with perceived performance in mindreading tasks exceeds that with actual performance in mindreading tasks, as expected by the rival hypotheses.

The fourth validity test uses evidence based on test–criterion relationships, where patterns of association with a theoretically relevant criterion variable are examined. For this purpose, we selected the construct of psychosocial functioning, which has been widely studied in relation to mindreading

measures (see, e.g., Grainger et al., 2023; He & Côté, 2023; Pinkham et al., 2018). Mindreading ability is considered a causal antecedent of optimal psychosocial functioning, enabling adaptation to the extraordinarily complex and dynamic social environments that characterize human existence (Luyten et al., 2020). Therefore, if mindreading measures are valid, they should predict higher psychosocial functioning, and this statistical association should remain positive when controlling for other constructs that are neither full mediators nor colliders (see, e.g., Rohrer, 2018). Controlling for constructs that may be tapped by mindreading measures is essential to rule out alternative causal mechanisms to the direct causal effect of mindreading ability on psychosocial functioning. On the one hand, this is relevant for the construct of general cognitive ability, as it could exert its own causal effect on psychosocial functioning (Pettersson et al., 2021), which could consequently induce a nonspecific association between mindreading tasks and psychosocial functioning. On the other hand, this consideration also applies to general positive self-evaluation, the tendency to rate oneself favorably across domains, because it likely contributes to the variance in measures of domain-specific self-evaluations, such as those related to mindreading ability. General positive self-evaluation is typically considered an adaptive trait due to its associations with self-esteem and emotional resilience (e.g., Bono & Judge, 2003). In this context, it is important to recognize that such associations could be driven by both bias, the mere

belief that one excels in all areas (Dufner et al., 2019), and substance, the actual ability to excel in all areas (van der Linden et al., 2010). Given these considerations, the fourth validity test is designed to elucidate the unique associations between mindreading measures and psychosocial functioning, while statistically controlling for general cognitive ability and general positive self-evaluation. Whereas the validity hypotheses predict unique positive associations, the rival hypotheses yield contrasting predictions. For mindreading tasks, the rival hypothesis predicts the absence of a unique association and ascribes this to a complete overlap with general cognitive ability. For mindreading self-reports, the rival hypothesis predicts either the absence of a unique association or a negative unique association, in line with the notion that overconfidence in one's mindreading ability could be maladaptive (e.g., Bateman & Fonagy, 2019; Sharp & Vanwoerden, 2015).

Methods

Procedure

Data were collected online using Prolific. A stratified sampling procedure was used to approximate the demographic distribution of age, sex, and ethnicity in the general U.S. population. We aimed for a sample size of at least 700 individuals

Table 1
Validity Tests

Validity Test	Statistical Estimand	Validity Hypotheses	Rival Hypotheses
		Theoretically Predicted Estimate	
#1	Latent correlation between mindreading self-reports and tasks	Mindreading self-reports and tasks reflect mindreading ability without simultaneously reflecting unrelated traits or other contaminating influences	Mindreading tasks reflect general cognitive ability; mindreading self-reports reflect perceived mindreading ability
#2	Latent correlation between mindreading tasks and general cognitive ability tasks	A medium-to-large positive correlation would be predicted if both self-reports and tasks captured mindreading ability	A small or near-zero correlation would be predicted if mindreading self-reports and tasks assessed distinct constructs
#3	Algebraic difference in the latent correlation coefficients between self-reported mindreading ability and perceived versus actual mindreading task performance	A medium positive correlation would be predicted if mindreading ability was facilitated by broader abilities such as general cognitive ability	A large or near-perfect positive correlation would be predicted if mindreading tasks and general cognitive ability tasks largely assessed the same construct
#4	Latent partial regression coefficients of mindreading measures predicting psychosocial functioning, controlling for general positive self-evaluation and general cognitive ability	A higher correlation between self-reported mindreading ability and actual performance in mindreading tasks would be predicted	A higher correlation between self-reported mindreading ability and perceived performance in mindreading tasks would be predicted
		Positive partial regression coefficients would be predicted if mindreading ability was a specific ability that facilitated psychological adjustment	For self-reports, near-zero or negative partial regression coefficients would be predicted as the variance shared with general positive self-evaluation is controlled for; for tasks, near-zero partial regression coefficients would be predicted as the variance shared with general cognitive ability tasks is controlled for

to obtain stable and precise correlation estimates (Kretschmar & Gignac, 2019). Participants were invited to two separate sessions, with a combined average completion time of 69 minutes. While Session 2 included the mindreading self-reports and all tasks, Session 1 included all other questionnaire materials. In Session 1, all measures were presented in random order; in Session 2, self-reports were presented first in random order, followed by tasks in random order. The average interval between sessions was 51 hours. All participants gave informed consent and were compensated at the local minimum wage. We implemented extensive measures to detect and remove careless respondents, including recording page times, the inclusion of five instructed response items and three bogus items, the calculation of an even-odd consistency index, and questions about task diligence¹ (Ward & Meade, 2023). Exclusion criteria were as follows: (a) failing one or more instructed response or bogus items, (b) averaging less than 1.5 seconds per item in any of the sessions (Bowling et al., 2023), (c) admitting insufficient effort (Meade & Craig, 2012), and (d) having an even-odd consistency index of less than .27 (Johnson, 2005), although this last criterion did not result in any exclusions. Of the 846 and 807 individuals who

completed Session 1 and Session 2, respectively, 107 were excluded due to evidence of careless responding.² There were no missing values in the data. The conduct of the study adhered to the ethical guidelines of the German Psychological Society (German Psychological Society, 2018) and complied with local legislation.

Participants

The sample consisted of $N = 700$ participants (52% female, 46% male, and 2% not identifying with either gender). Age ranged from 19 to 85 years ($M = 46.3$, $SD = 15.8$). With regard to ethnicity, the proportions were as follows: White or Caucasian (76%), Black or African American (12%), Asian or Pacific Islander (6%), Hispanic or Latino (4%), multiracial or biracial (1%), and other (less than 0.1%); no participants identified as Native American or Alaska Native. Educational attainment included doctorate (3%), master's (13%), bachelor's (37%), professional degree (12%), high school (33%), and elementary school (2%); no participants reported no schooling. Regarding current occupation, participants reported being

Table 2
Overview of Constructs and Corresponding Measures

Construct	Measure	Material/Procedure (Paraphrased Example Item)
Self-reported mindreading ability	<i>Affective and Cognitive Measure of Empathy – Cognitive Empathy</i> (ACME-CE; Vachon & Lynam, 2016)	Ten statements (e.g., "It's difficult for me to understand people's feelings") evaluated using a 5-point Likert-type scale (e.g., "Agree strongly")
	<i>Behavioral, Emotional, and Social Skills Inventory – Perspective-Taking Skill</i> (BESSI-PT; Soto et al., 2022)	Six activities (e.g., "Understand the emotions of others") rated on a 5-point scale reflecting one's ability (e.g., "Extremely well")
	<i>Certainty About Mental States Questionnaire – Other-Certainty</i> (CAMSQ-OC; Müller et al., 2023)	Ten situations (e.g., "I can tell if someone is trustworthy"), for which the frequency of successful mindreading inferences is indicated on a 7-point frequency scale (e.g., "Always")
	<i>Empathy Quotient – Cognitive Empathy</i> (EQ-CE; Baron-Cohen & Wheelwright, 2004)	Ten statements (e.g., "I quickly figure out what others might like to discuss") assessed on a 5-point Likert-type scale
	<i>Four-Item Mentalising Index</i> (FIMI; Clutterbuck et al., 2021)	Four statements (e.g., "I typically understand how others think, even when it's different from my viewpoint") evaluated on a 4-point Likert-type scale
	<i>Mentalization Scale – Other-Related Mentalization</i> (MentS-O; Dimitrijević et al., 2018)	Nine statements (e.g., "I can comprehend how other people feel") rated on a 5-point Likert-type scale
	<i>Reflective Functioning Questionnaire – Extended – Mentalizing Others</i> (RFQ18-MO; Rogoff et al., 2021)	Nine statements (e.g., "I can easily tell what others are feeling or thinking") rated on a 7-point Likert-type scale
Self-Estimated Mindreading Ability (SE MRA)	One rating of one's own mindreading ability, including a brief definition of mindreading ability, using a 100-point slider scale with six visual anchor points (e.g., "Strongly above average")	
Task-based mindreading ability	<i>Geneva Emotion Knowledge Test – Blends</i> (GEMOK-Blends; Schlegel & Scherer, 2018)	Ten written short stories, each requiring identification of fictional characters' emotions, from five response options
	<i>Movie for the Assessment of Social Cognition – Revised</i> (MASC-R; Dziobek et al., 2006) ^a	Fifteen cinematic scenes from a short film, requiring identification of the mental states of fictional characters from four options
	<i>Personality Pairs Task</i> (PPT; Conway et al., 2020) ^b	Thirteen pairs of person-descriptive statements from the HEXACO-100 (Lee & Ashton, 2018), requiring identification of the likelihood that both statements are true for the same person, using a 100-point continuous slider scale with poles "very unlikely" and "very likely" marking opposite ends
	<i>Reading the Mind in the Eyes Test–Short Form</i> (RMET-S; Olderbak et al., 2015)	Ten black and white photographs of eye regions, requiring identification of mental states of individuals from four options

Note. ^a This is a new and yet unpublished version of the MASC with updated video material. Only the first 15 scenes, which make up the first third of the MASC-R, were used. ^b An adapted version of the PPT was used, with modified items and scoring procedure described in Note S1.

Table 2 (continued)

Construct	Measure	Material/Procedure (Paraphrased Item Example)
General cognitive ability	<i>International Cognitive Ability Resource – Matrix Reasoning Task</i> (ICAR-MR; Condon & Revelle, 2014) ^c	Six geometric puzzles, requiring identification of the missing shape from six options
	ICAR–Verbal Reasoning Task (ICAR-VR; Condon & Revelle, 2014) ^c	Six logic problems, requiring identification of the correct answer from seven options
	Vocabulary Test (VOC; Open-Source Psychometrics Project, 2020) as used in Müller et al. (2023)	Twelve lists of English words, each comprising five words, requiring identification of the two synonymous words
Perceived performance: mindreading tasks	Perceived performance in GEMOK-Blends (PP GEMOK-Blends)	A single item presented after completion of the task, assessing perceived performance, with response options ranging from 0 to the maximum possible number of correct items
	Perceived performance in MASC-R (PP MASC-R)	–
	Perceived performance in PPT (PP PPT)	–
Perceived performance: general cognitive ability tasks	Perceived performance in RMET-S (PP RMET-S)	–
	Perceived performance in ICAR-MR (PP ICAR-MR)	–
	Perceived performance in ICAR-VR (PP ICAR-VR)	–
General positive self-evaluation	Perceived performance in VOC (PP VOC)	–
	<i>Behavioral, Emotional, and Social Skills Inventory–45</i> (BESSI-45; Soto et al., 2022) – Emotional Resilience Skills (BESSI-45-ER)	Nine activities in the domain of emotional resilience (e.g., "Calm down when I'm anxious ") rated on a 5-point scale reflecting one's ability (e.g., "Extremely well")
	BESSI-45 Innovation Skills (BESSI-45-IN)	Nine activities in the domain of innovation (e.g., "Make a work of art ")
	BESSI-45 Self-Management Skills (BESSI-45-SM)	Nine activities in the domain of self-management (e.g., "Keep everything clean and organized")
Psychosocial functioning	BESSI-45 Social Engagement Skills (BESSI-45-SE)	Nine activities in the domain of social engagement (e.g., "Guide a group of people")
	<i>Lubben Social Network Scale-Revised – Friendships</i> (LSNS-R; Lubben et al., 2006)	Six questions about the frequency of contact with friends, rated on a 6-point frequency scale
	<i>Inventary of Interpersonal Problems–32</i> (IIP-32; Horowitz et al., 2000)	Thirty-two interpersonal behaviors, evaluated in terms of difficulty to perform, or whether they are performed excessively, on a 5-point scale
	<i>Negative Consequences of Personality</i> (NCP; Leising & Zimmermann, 2011) – Internalizing Problems (NCP-INT)	Eight life adversities in the domain of internalizing problems (e.g., social isolation), rated for whether they occurred or are likely to occur as a consequence of the subjects personality, on a 4-point scale
	NCP – Externalizing Problems (NCP-EXT)	Nine life adversities in the domain of externalizing problems (e.g., substance use)
	NCP – Occupational and Financial Problems (NCP-OCC)	Five life adversities in the domain of occupational and financial problems (e.g., poor housing situation)
	<i>Personality Disorder Severity–ICD–11 Scale</i> (PDS-ICD-11; Bach et al., 2021)	Fourteen areas of self- and interpersonal functioning, each with characteristic descriptions for different levels and qualities of functioning
	<i>Symptom Checklist K–9</i> (SCL-K9; Petrowski et al., 2019)	Nine symptoms of mental disorder, rated for the distress they caused during the past week on a 4-point scale
<i>Short Social Functioning Questionnaire</i> (SSFQ; Tyrer et al., 2021)	Five statements describing behaviors or experiential patterns, most of these rated regarding the frequency with which they are experienced on a 4-point scale	
<i>WHO Disability Assessment Schedule 2.0 – Getting Along scale</i> (WHODAS 2.0; Üstün et al., 2010)	Five social activities, rated for the difficulty engaging in them during the past 30 days on a 5-point scale	

Note. ^c For each ICAR task, a subset of 6 items was administered.

employed for wages (48%), self-employed (18%), unemployed (13%), retired (13%), studying (3%), on a career break (2%), or unable to work (2%). English language proficiency included 91% native speakers, with the remaining 9% reporting fluency.

Measures

To assess each construct with multiple measures, we administered a battery of psychological tests: eight mindreading self-report scales, four mindreading tasks, three general cognitive ability tasks, seven assessments of perceived task performance, nine indicators of psychosocial functioning, and four scales assessing self-evaluations across a range of behavioral, emotional, and social skills (Soto et al., 2022). The

selection of specific instruments was guided by a comprehensive literature review that assessed their importance in the existing literature through criteria such as citation frequency, especially in the last five years, and established psychometric properties in terms of internal structure and consistency. For more recently introduced measures, the perceived potential for future impact, as judged by the methodological rigor of test development, was a key consideration. The mindreading tasks were specifically selected to tap different approaches, including video, picture, and text materials. Given the time-intensive nature of the tasks, we opted for abbreviated versions. Table 2 provides an overview of the measures, and Table S1 provides model-based internal consistency estimates for all test scores.

Statistical Analysis

The data and R code required to reproduce the analyses are publicly and permanently available in an Open Science Framework online repository (<https://doi.org/10.17605/OSF.IO/Q7EU4>). We adhered to the preregistration (<https://doi.org/10.17605/OSF.IO/VQKCX>) and document minor deviations in Note S2. The analyses were carried out using R Version 4.2.2 (R Core Team, 2022) and *Mplus* Version 8.4 (Muthén & Muthén, 2017), with the R packages *lavaan* Version 0.6.14 (Rosseel, 2012) and *MplusAutomation* Version 1.1.0 (Hallquist & Wiley, 2018). Latent variable models were estimated by maximum likelihood using a robust test statistic and robust standard errors (MLR). To ensure semantic alignment, item coding and test scoring were conducted in a manner consistent with the nomenclature used to represent the respective constructs throughout this article. For example, higher scores on the Inventory of Interpersonal Problems–32 (IIP-32; Horowitz et al., 2000), which are typically scored as indicative of more pronounced interpersonal difficulties and thus greater impairments in psychosocial functioning, were reverse-coded so that higher values denoted higher levels of psychosocial functioning.

Internal Structure

In preliminary analyses, we determined the internal structure of mindreading ability measures for each of the two methods using separate models. This procedure was undertaken to confirm the assumption of essential unidimensionality and to investigate additional factors beyond the general factor, thus allowing a more nuanced interpretation of the measures (e.g., Reise et al., 2013). For the self-report measures, we used an extended exploratory structural equation modeling approach (ESEM; Marsh et al., 2020) that included all 61 items of the eight scales, as shown in Figure S1. Confirmatory factors included an acquiescence factor to account for acquiescent response style and seven questionnaire-specific factors to capture the unique characteristics of each scale (e.g., instructions, response formats). Consequently, these confirmatory factors were designed to isolate variance attributable to the method. Exploratory factors, rotated via orthogonal bigeomin rotation, included a general factor that was intended to capture variance shared across all items and additional factors that were meant to capture other substantive sources of variance. Intercorrelations between factors were set to zero to allow for clear partitioning of variance. Based on recent recommendations (Bader & Moshagen, 2022), the optimal number of exploratory factors was determined using the Tucker-Lewis index (TLI) and the root mean square error of approximation (RMSEA), with additional consideration given to deriving a maximum number of interpretable factors. For the tasks, we used bifactor confirmatory factor analysis (bifactor CFA), as shown in Figure S2. To facilitate model estimation, we adopted a homogeneous parceling

strategy aimed at item-to-construct balance (Little et al., 2002). The model included a general factor to capture variance shared across tasks and task-specific factors to account for variance unique to each task. Due to its scoring procedure, the Personality Pairs Task (PPT) was available only as a total score, which precluded the specification of a task-specific factor.

Validity Tests

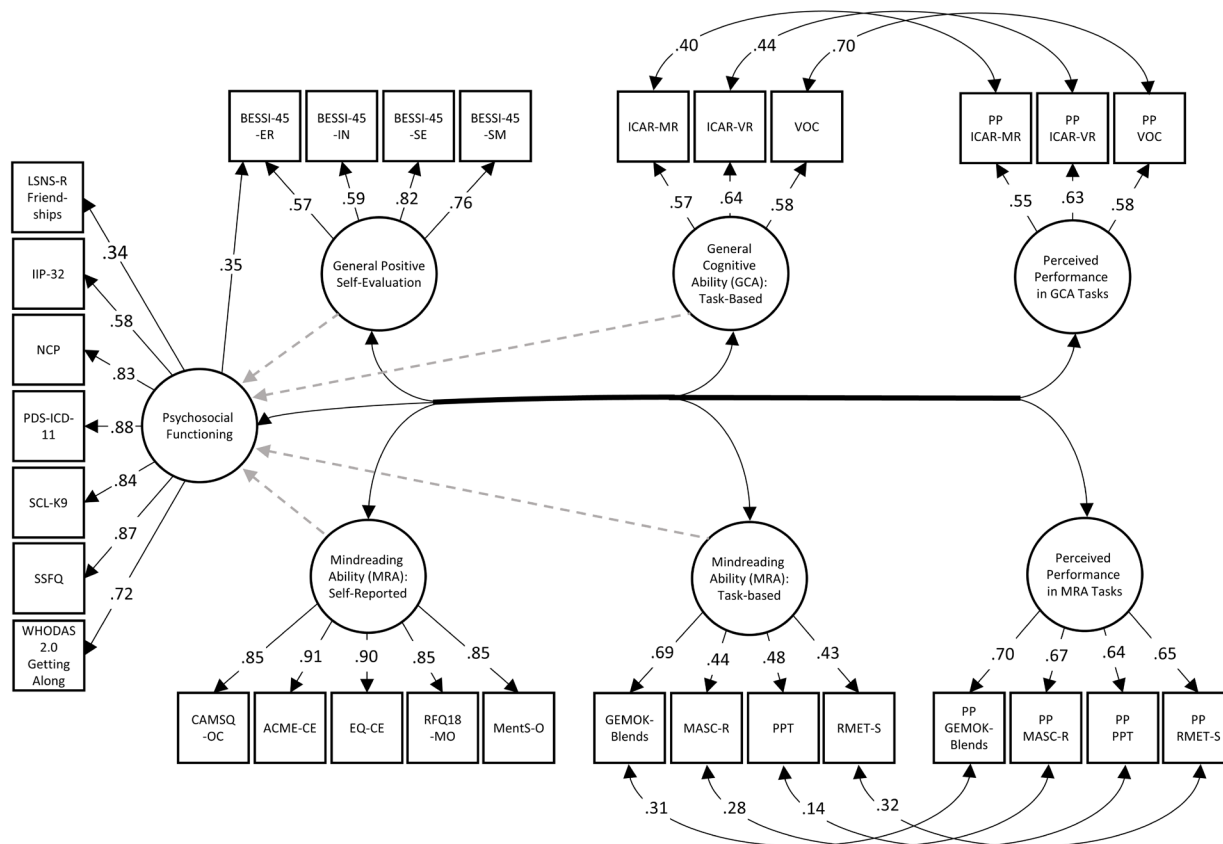
To lay the foundation for validity testing, we specified a structural equation model (SEM) that included all constructs, each instantiated by its corresponding set of measures. In contrast to the more nuanced measurement models used to delineate internal structure, in this step we used the total scores of mindreading measures as indicators to extract a single common factor for each construct. This model largely adhered to a simple structure, with theoretically justified exceptions for correlated residuals between perceived and actual performance in each task and a cross-loading from the psychosocial functioning factor onto BESSI-45 Emotional Resilience Skills. For validity tests one through three, latent correlations were the primary estimands. Algebraic differences in latent correlations were introduced as additional model parameters for the third test. For the fourth validity test, we used two variations of this SEM. The first regressed psychosocial functioning on self-reported mindreading ability, general cognitive ability, and general positive self-evaluation. In the second, the regression slope of self-reported mindreading ability was replaced by a regression slope of task-based mindreading ability, while keeping all other predictors unchanged.

Results

Internal Structure

Using the extended ESEM approach, we sequentially extracted an increasing number of exploratory factors, which were entered alongside the constant set of confirmatory factors, as reported in Table S2. Although fit indices did not indicate a substantial degree of model misspecification for models with two or more content-specific factors, we retained the model with six content-specific factors (TLI = .956, RMSEA = .028), as it provided the largest number of interpretable and well-defined factors. Table S3 presents the standardized factor loading matrix for this model. All but one of the items had standardized loadings greater than .30 on the general factor, with the majority of the loadings exceeding .50. The first content-specific factor was interpreted as SC1 Affective Empathy,

Figure 1
Structural Equation Model Used for Validity Testing



Note. Estimates are standardized. Estimates for structural model parameters, factor means and variances, and indicator intercepts and residual variances are omitted for clarity. Single-headed arrows indicate regression slopes and factor loadings; double-headed arrows indicate correlations. Dotted arrows indicate regression slopes used in Validity Test 4.

with marker items such as “I can empathize with other people,” (paraphrased) which we considered a construct-irrelevant factor given its conceptual distinction from mindreading ability. The other content-specific factors (SC2–SC6) were considered construct-relevant, each denoting specific facets or contexts subsumable under the definition of mindreading ability. These consisted of SC2, labeled Recognizing Mental States that are Actively Masked or Withheld (e.g., “I recognize when people are faking happiness” [paraphrased]), SC3 Perspective-Taking (e.g., “It’s easy for me to see things from someone else’s perspective” [paraphrased]), SC4 Predicting Mental States (e.g., “I can predict how others will respond to a situation” [paraphrased]), SC5 Recognizing Negative Reactions (e.g., “I know when someone is being bored with me” [paraphrased]), and SC6 Severe Inability (e.g., “I am puzzled by the way other people think and feel” [paraphrased]). We next examined internal consistencies of indicator sets based on the extended ESEM, with results documented in Table S4. A unit-weighted total score aggregated across the full set of 61 indicators yielded an internal consistency of $\omega_H = .90$ for

the general factor and $\omega_H \leq .03$ for all other factors, indicating that the general factor was the major contributor to the observed variance. A notable proportion of the variance in the unit-weighted scores for BESSI-PT (27%) and FIMI (12%) was accounted for by affective empathy, a factor that was considered construct irrelevant; therefore, these two scales were excluded from the subsequent SEM.

The proposed two-factor CFA model for task-based measures of mindreading ability showed good fit ($TLI = .995$, $RMSEA = .011$) and yielded a plausible factor loading pattern, as shown in Table S5. Standardized factor loadings ranged from .26 to .52 for the general factor and from .21 to .39 for the task-specific factors, indicating a similar proportion of shared versus unique variance for each mindreading task. Internal consistency for the general factor was $\omega_H = .60$, indicating relatively low reliability.

Validity Tests

The SEM presented in Figure 1 yielded fit indices of $TLI = .924$, $RMSEA = .050$, and $SRMR = .058$, along with plausible

Table 3
Latent Correlations: Point Estimates and 95% Confidence Intervals

Factors	MRA Self-Reported	MRA Tasks	GCA Tasks	PP in MRA Tasks	PP in GCA Tasks	GPSE
MRA Tasks	.05 [-.04, .15]					
GCA Tasks	-.11* [-.20, -.01]	.85* [.76, .94]				
Perceived Performance in MRA Tasks	.29* [.20, .38]	.17* [.04, .30]	.12* [.01, .23]			
Perceived Performance in GCA Tasks	.11 [.01, .21]	.41* [.27, .56]	.65* [.56, .74]	.65* [.55, .75]		
General Positive Self-Evaluation	.46* [.38, .54]	-.15* [-.26, -.04]	-.03 [-.14, .08]	.27* [.17, .36]	.22* [.12, .33]	
Psychosocial Functioning	.13* [.03, .22]	-.02 [-.12, .07]	.03 [-.07, .13]	.10 [.01, .20]	.11* [.10, .21]	.65* [.58, .71]

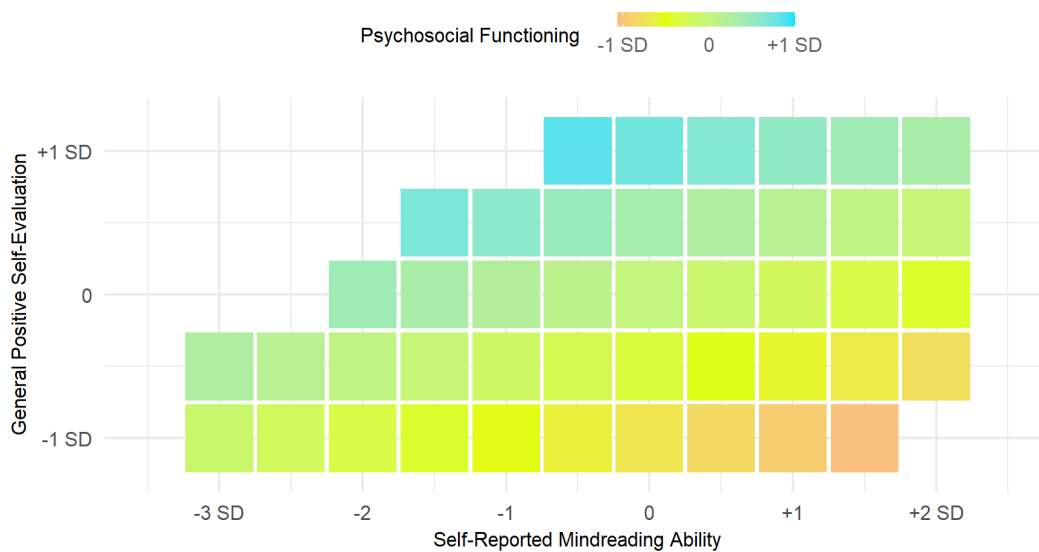
Note. Estimates based on the structural equation model. MRA = Mindreading ability; GCA = General cognitive ability; PP = Perceived performance, GPSE = General positive self-evaluation.

* $p < .05$

parameter estimates, indicating that the model was reasonably consistent with the data. The latent correlations used in Validity Tests 1–3 are reported in Table 3, while the latent partial regression coefficients for Validity Test 4 are reported in Table S6. Validity Test 1 yielded a nonsignificant latent correlation, suggesting a lack of heteromethod convergence between self-reported and task-based mindreading ability, a finding more consistent with the rival hypotheses. Validity Test 2 indicated a large latent correlation between mindreading and general cognitive ability tasks, suggesting substantial convergence in task performance between the two domains, which is more consistent with the rival hypotheses. Validity Test 3 showed that the latent correlation between self-reported mindreading ability and perceived performance in mindreading tasks was significantly higher than that with actual performance in such tasks ($p < .001$), which is also more consistent with the rival hypotheses.

For Validity Test 4, higher self-reported mindreading ability was uniquely associated with poorer psychosocial functioning, $\beta = -0.21$, 95% CI [-0.29, -0.12]. This finding needs to be contextualized with the beta coefficients of the other predictors in the model, particularly the unique positive association between general positive self-evaluation and psychosocial functioning, $\beta = 0.74$, 95% CI [0.66, 0.81]. As shown in Figure 2, the simultaneous consideration of self-reported mindreading ability and general positive self-evaluation in the prediction of psychosocial functioning yielded distinct adaptive and maladaptive profiles. The adaptive profile manifested as high general positive self-evaluation coupled with modest self-reported mindreading ability. Conversely, the maladaptive profile manifested as low general positive self-evaluation and high self-reported mindreading ability. Regarding mindreading tasks, there was a unique negative association with

Figure 2
Prediction of Psychosocial Functioning by Self-Reported Mindreading Ability and General Positive Self-Evaluation



Note. Predictions are based on the structural equation model used in Validity Test 4, with tiles representing the observed profiles of the predictor variables. All variables are standardized. Higher psychosocial functioning is shown in blue; lower psychosocial functioning is shown in orange.

psychosocial functioning, where higher performance in mindreading tasks predicted lower psychosocial functioning, $\beta = -0.65$, 95% CI [-1.28, -0.02]. This effect must be interpreted in the context of better performance in general cognitive ability tasks predicting higher psychosocial functioning, $\beta = 0.71$, 95% CI [0.08, 1.33]. Both regression slopes were statistically significant, although their confidence intervals were particularly wide due to multicollinearity. While the results for Validity Test 4 were generally more consistent with the rival hypotheses, they were not entirely consistent with the rival hypothesis for mindreading tasks, which predicted a complete overlap with general cognitive ability and thus expected a null effect for the partial regression coefficient.

The results for each individual measure of mindreading ability largely mirrored those found at the construct level, as detailed in Tables S7–S11. Three exceptions should be mentioned: first, small positive latent correlations were found between the MASC-R and mindreading self-reports, in contrast to the near-zero correlation observed at the construct level. Second, latent partial regression coefficients of mindreading tasks in predicting psychosocial functioning were predominantly nonsignificant, in contrast to the significant effect at the construct level. Third, positive latent partial regression slopes were observed for mindreading self-reports in predicting LSNS friendships, indicating that more frequent contact with friends was uniquely predicted by higher levels of self-reported mindreading ability.

Some aspects of the latent correlation patterns that were not specifically addressed by the validity tests are further noteworthy. For example, whereas the correlation between actual and perceived performance in mindreading tasks was small, the corresponding correlation between actual and perceived performance in general cognitive ability tasks was large. This suggests that self-assessment of one's own task performance was considerably more difficult in the context of mindreading tasks than in the context of general cognitive ability tasks. Moreover, the correlation between self-reported mindreading ability and perceived performance in mindreading tasks was significantly higher than that between self-reported mindreading ability and perceived performance in general cognitive ability tasks. This is consistent with the interpretation that self-reports of mindreading ability tap into a domain-specific self-evaluation that specifically represents how individuals evaluate their competence within the mindreading domain.

Discussion

To contribute to the burgeoning critical discourse questioning the validity of self-report and task-based measures of mindreading ability (e.g., Murphy & Lilienfeld, 2019; Osborne-Crowley, 2020), we conducted a comprehensive multimethod investigation using a large, heterogeneous sample from the general U.S. population. In preliminary analyses of internal

structure, we first established essential unidimensionality for both self-report and task-based measures of mindreading ability. We then used structural equation modeling to conduct four validity tests, contrasting rival hypotheses—which propose that mindreading self-reports capture perceived mindreading ability and mindreading tasks assess general cognitive ability—with validity hypotheses advocating for their respective validity. The validity tests suggested that the data were more supportive of the rival hypotheses. This was evident at both the level of constructs and at the level of individual measures, which attests to the robustness and potential generalizability of these findings. While our results may have far-reaching implications, they also require further corroboration through replication, extension, and additional tests of generalizability, as construct validation is a dynamic, evolving process (Messick, 1995).

What is Captured by Mindreading Self-Reports and Tasks?

The present analyses support the rival interpretation that self-reports of mindreading ability primarily reflect perceived mindreading ability, a construct that could also be termed mindreading self-concept or mindreading-specific self-evaluation. This interpretation was bolstered by the negligible correlation between mindreading self-reports and actual performance in mindreading tasks, the negative partial regression slope in the prediction of psychosocial functioning, and the medium association with perceived performance in mindreading tasks, which exceeded the association with perceived performance in general cognitive ability tasks. These results are in line with mentalizing theory, in which overconfidence in one's mindreading ability, referred to as hypermentalizing, is considered maladaptive, whereas humility regarding one's mindreading ability, termed genuine mentalizing, is considered adaptive (e.g., Müller et al., 2023; Sharp & Vanwoerden, 2015). However, our results further suggest that self-reports of mindreading ability may, to some extent, capture general positive self-evaluation, with more positive self-evaluations generally being more adaptive. This confluence was evident in the sign change from a positive bivariate association between self-reported mindreading ability and psychosocial functioning to a negative unique association after controlling for general positive self-evaluation and general cognitive ability. This pattern of results further suggests that the more unique variance part of perceived mindreading ability may be isolated by controlling for variance associated with general positive self-evaluation, as in a suppression effect (Hoyle et al., 2023; Watson et al., 2013). The interpretation that self-report measures primarily reflect perceived mindreading ability suggests that their test scores cannot be safely used to study actual mindreading ability (e.g., Ickes, 1993; Realo et al., 2003). This is not to say that these test scores necessarily have no variance related to actual min-

dreading ability. Rather, it implies that the common interpretation that equates high perceived mindreading ability with high actual mindreading ability may lead to flawed conclusions about the construct.

For task-based measures of mindreading ability, our results suggest that variation in task performance can be explained primarily by individual differences in general cognitive ability. This is consistent with previous research pointing to excessive overlap between these two constructs. Our estimated latent correlation of .85 falls between the .65 reported by Coyle et al. (2018) and the .92 observed by Navarro (2022), although it is worth noting that these studies used slightly different tasks. We interpret these large correlations as evidence that mindreading tasks may not be capturing an ability that is distinct from general cognitive ability, a problem that has also been encountered in attempts to establish emotional intelligence as a distinct ability (e.g., MacCann et al., 2014). However, our results deviate somewhat from the notion that mindreading tasks are completely redundant with general cognitive ability tasks, as evidenced by the nonperfect association with general cognitive ability and the tentative evidence of a unique negative association with psychosocial functioning. Thus, mindreading tasks may contain a limited amount of unique variance that could potentially tap into a more specific ability dimension or other person characteristics. It should be noted, though, that the confidence interval for the unique association with psychosocial functioning was so wide that it is currently unclear whether this effect is meaningful at all. Moreover, this effect runs counter to both theoretical predictions and the empirical findings of He and Côté (2023) and Pinkham et al. (2018), who found positive unique associations between mindreading tasks and psychosocial functioning. However, it is important to emphasize that neither of these two other studies used a latent variable approach and only controlled for a single measure of general cognitive ability. As a result, these studies ran the risk of producing spurious unique effects due to residual confounding, which refers to a situation in which statistical control is compromised by the unreliability of the measure being controlled for (Westfall & Yarkoni, 2016).

In contrast to our approach, which focused on the commonalities across mindreading tasks, some scholars suggest that the valid variance relevant to mindreading ability may reside in the unique features of the tasks (see, e.g., Navarro, 2022). According to this perspective, a wide range of tasks may be too heterogeneous, leading to the extraction of a common factor that fails to capture the part of the variance that specifically reflects mindreading ability (e.g., Quesque & Rossetti, 2020). However, such an interpretation is not parsimonious; and focusing on shared variance through a latent variable approach is typically effective in condensing the meaningful variance in psychological assessments (e.g., Epstein, 1979; Wang & Navarro-Martinez, 2023). While specific cognitive abilities may occasionally provide incremental predictive value beyond broader cognitive domains, they more

often do not (Ree & Carretta, 2022). More critically, the assertion that the valid variance could be contained in the unique variance did not find empirical support in our data, as similar patterns of associations were observed for each of the tasks individually. Given these considerations, we tend to think that unique aspects of tasks often represent idiosyncrasies that are statistical nuisances rather than of substantive interest (e.g., Sellbom & Tellegen, 2019).

Implications and Recommendations

First and foremost, our study challenges the assumption of established validity in mindreading measures and provides evidence for alternative interpretations. Validity is a nuanced concept that can refer to either the ratio of valid to invalid variance in test scores (Schimmack, 2021) or the fidelity with which these scores reproduce the true nomological network of the intended construct (e.g., Westen & Rosenthal, 2003). Overlooking contamination in mindreading measures could risk confusing the nomological networks of mindreading ability with those of unrelated traits or other contaminating influences. As a result, previous research that relies on validity assumptions may have reached potentially inaccurate or misleading conclusions. When evaluating studies that use mindreading measures, researchers should explicitly consider whether empirical results could be alternatively explained by plausible rival interpretations. However, navigating this terrain can be complex due to the often imprecise nature of psychological theories and the possibility that different interpretations may yield similar predictions for the nomological network (e.g., Eronen & Bringmann, 2021).

Second, the ongoing challenge is to optimize the assessment of mindreading ability. A key consideration in refining tasks is to ensure that they elicit the specific processes involved in mindreading, rather than capturing broader cognitive functions. This may involve the development of more naturalistic or interactive tasks, along the lines suggested by Osborne-Crowley (2020). Existing self-report measures directly and overtly ask individuals to self-assess their mindreading ability (e.g., "I am a good mind reader"), whereas an alternative strategy might avoid such items and instead aim to identify a set of behaviors that could serve as proxies. Another alternative might be to solicit informant or expert ratings, with preliminary evidence suggesting that aggregated scores from multiple informants can yield acceptable interrater reliability and potentially improved validity (Elfenbein et al., 2015).

Third, recognizing that the pursuit of perfectly valid measures of mindreading ability may be an unrealistic goal, researchers could employ strategies to mitigate the shortcomings of existing measures. One approach is to identify contaminating constructs in mindreading measures and statistically remove their influence while preserving valid variance (Hoyle et al., 2023). However, this approach has limitations: partialing out a variable that itself contains measure-

ment error may result in a residualized variable that still carries some of the measurement error, as discussed previously (Westfall & Yarkoni, 2016). Furthermore, the impact of partialing on the validity of test scores can vary substantially under different scenarios. If another construct, such as general cognitive ability, accounts for the variance in a target construct, such as mindreading ability, there could be several explanations for this relationship: a substantive relationship between the constructs (Scenario A), measurement contamination (Scenario B), or a combination of the two. If no partialing is done, Scenario A is unproblematic because the test score would faithfully reproduce the true nomological network of mindreading ability; however, in Scenario B, the test score would produce the nomological network of general cognitive ability, which may or may not coincide with the one for mindreading ability. Now, when partialing is used, the statistical overlap between constructs is eliminated, with different implications: it correctly removes the invalid variance in Scenario B, thereby creating a more valid variable, while inadvertently removing some of the valid variance in Scenario A. In Scenario A, the use of partialing could potentially compromise the integrity of the test score, resulting in a less favorable ratio of valid to invalid variance (Hoyle et al., 2023).

Fourth, notwithstanding the proliferation of mindreading measures in recent decades, many validation studies lack the methodological depth and rigor necessary to reliably detect potential sources of invalidity. They often resort to generic tests of validity or reliability and fail to articulate theoretical predictions regarding rival interpretations. Moreover, there is a persistent focus on monomethod evidence in the current validation landscape—often confined to tests of internal consistency, convergent validity, or structural validity. While such tests are indispensable, their potential to disambiguate competing interpretations is often limited, thereby diminishing their contribution to a nuanced understanding of psychological tests (Schimmack, 2021). Preregistration is another important aspect of construct validation studies because it requires test developers to articulate theoretical predictions in advance, thereby increasing the likelihood that validity problems will be detected and decreasing the likelihood that a sound theory will be unjustly revised simply to maintain the validity assumption of a flawed measure. In addition, replication serves as another critical element of test validation, particularly when results deviate from initial expectations. For example, meta-analytic evidence has generally demonstrated limited convergence between self-report and task-based mindreading measures (Murphy & Lilienfeld, 2019). Thus, it was surprising when Clutterbuck et al. (2021) reported a heteromethod correlation of .35 between the FIMI and RMET-S and claimed this as evidence of the superior validity of the FIMI. This claim was challenged in a subsequent commentary by Murphy et al. (2022), who suggested that the FIMI was “merely a relabeled brief version of widely used existing scales” (p. 403). Notably, in the present study, a low correla-

tion of .01 between the FIMI and RMET-S was observed, aligning more closely with the critical perspective of Murphy et al. (2022). While the exact reasons for the failure to replicate the effect reported by Clutterbuck et al. (2021) remain unclear, a plausible explanation could be contamination by careless respondents, a problem known to lead to spurious associations between self-reports and tasks (e.g., Huang & DeSimone, 2021).

Fifth, ontological considerations add an additional layer of complexity to the evaluation of the validity of mindreading measures. According to Borsboom et al. (2004), a measure can achieve validity only if the construct it seeks to quantify exists as conceptualized; otherwise, reevaluation or potential obsolescence of the construct may be necessary. Following this line of reasoning, one interpretation of our findings suggests that mindreading ability may simply be an epiphenomenon of general cognitive ability, resulting in empirically mostly indistinct constructs. From this perspective, general cognitive ability tasks would be sufficient to assess mindreading ability, rendering specialized mindreading measures redundant.

Sixth, if the construct of mindreading ability had low variability within the general population, this may provide another plausible explanation for the validity problems encountered. Mindreading ability may not exist on a continuum, but rather as a taxonic latent structure in which normative participants exhibit broadly comparable levels of ability, while a smaller subgroup—possibly those with neurodevelopmental disorders such as autism spectrum disorder (Happé, 2015)—shows marked impairment. Empirical evidence supports taxonic latent structures in autistic traits (e.g., Haslam et al., 2020). Under the assumption of low variability, the administration of mindreading measures to normative samples may per se lead to low validity. This is because even small amounts of measurement error would disproportionately affect the ratio of valid to invalid variance. In this regard, it is important to emphasize that validity judgments are by definition population-specific, so it remains an open question to what extent our findings would generalize to other populations, such as clinical populations.

Seventh, reconceptualizing mindreading ability through a functionalist lens might shed light on certain aspects of our findings. A functionalist reconceptualization could define mindreading ability in terms of its ecological utility in everyday life, such as the ability to achieve favorable outcomes in interpersonal situations in which mindreading is required. Such a reconceptualization would transform mindreading ability from a narrow concept concerned primarily with inferring mental states from static information to a broader perspective encompassing interpersonal communication. Significantly, in real-world scenarios, active communication may serve as the primary mechanism for understanding others, rather than passive mindreading. Viewing mindreading as a broad interpersonal skill that is critical for managing social re-

relationships may explain why tasks that fit a narrower definition of mindreading ability may be inadequate. Moreover, this functionalist definition may also better explain why modest self-perception is more advantageous than overconfidence in mindreading (e.g., Müller et al., 2023), as individuals who maintain modest self-perception could be more likely to refrain from drawing premature conclusions about others and instead prioritize ongoing communication (Bateman & Fonagy, 2019).

Limitations

In this study, our analyses were limited to a specific selection of self-report and task-based measures of mindreading ability. Consequently, our conclusions may not apply to other measures that we did not consider. Future research should therefore repeat the analyses with additional measures, particularly those using different methodologies, as well as long forms of the measures used in this study, to ensure a broader and more comprehensive evaluation. A second limitation of this study is that only four validity tests were conducted. Because validity tests are strongly dependent on theoretical assumptions that may evolve, our conclusions about validity are limited to the current theoretical understanding of the construct. Therefore, it is critical for future research to develop and evaluate new validity tests in line with future scientific advances. A third limitation of our study is the potential lack of motivation among participants to fully engage in the mindreading tasks. Although participants on panel providers such as Prolific are aware that careless responses can result in financial penalties and affect future study opportunities, this does not guarantee maximum effort from all individuals. Thus, our study could best be described as a medium-stakes setting, which may have negatively impacted the validity of the tasks (Duckworth et al., 2011), especially if participants' effort consistently varied across tasks (Huang & DeSimone, 2021). Therefore, future research should explore ways to further increase the stakes, perhaps by offering additional incentives for each correct answer. Finally, the fourth validity test was limited by our reliance on self-report measures to assess psychosocial functioning. While these measures are typically viewed favorably in terms of their validity, as evidenced, for example, by their convergence with clinician ratings (Buer Christensen et al., 2020), they are of course not free from biases specific to the self-report method (Podsakoff et al., 2024). To improve the evidence for criterion-oriented validity, future studies should include alternative assessment methods, such as informant reports or behavioral observations of psychosocial functioning. This approach would allow for a more comprehensive assessment of this test–criterion relationship by reducing method-specific biases.

In sum, while this study provides empirical evidence that is more consistent with rival interpretations of mindreading measures, it cannot provide a definitive judgment regarding their validity, nor does it provide a quantitative estimate of

validity (see, e.g., Westen & Rosenthal, 2003). Our findings and conclusions require further corroboration by independent studies designed to establish replicability, explore the limits of generalizability, and conduct additional validity tests.

Footnotes

¹ Diligence in completing each task was assessed using three items adapted from Meade and Craig (2012). A paraphrased example of these items is “I worked on this task as hard as I could.” Items were rated on a 5-point scale ranging from “does not apply at all” to “fully applies.”

² Of the excluded participants, 35 failed more than one criterion. Seventy-two participants failed only one criterion. Of these, 52 were excluded for failing an instructed response item or a bogus item, 14 for responding too quickly, and six for reporting low effort in completing a task.

References

- Abramson, L., Uzefovsky, F., Toccaceli, V., & Knafo-Noam, A. (2020). The genetic and environmental origins of emotional and cognitive empathy: review and meta-analyses of twin studies. *Neuroscience & Biobehavioral Reviews*, *114*, 113–133. <https://doi.org/10.1016/j.neubiorev.2020.03.023>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.) (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Bach, B., Brown, T. A., Mulder, R. T., Newton-Howes, G., Simonsen, E., & Sellbom, M. (2021). Development and initial evaluation of the ICD-11 personality disorder severity scale: PDS-ICD-11. *Personality and Mental Health*, *15*(3), 223–236. <https://doi.org/10.1002/pmh.1510>
- Bader, M., & Moshagen, M. (2022). Assessing the fitting propensity of factor models. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000529>
- Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, *34*(2), 163–175. <https://doi.org/10.1023/B:JADD.0000022607.19833.00>
- Bateman, A. W., & Fonagy, P. (Eds.). (2019). *Handbook of mentalizing in mental health practice*. American Psychiatric Publishing.
- Bono, J. E., & Judge, T. A. (2003). Core self-evaluations: A review of the trait and its role in job satisfaction and job performance. *European Journal of Personality*, *17*, S5–S18. <https://doi.org/10.1002/per.481>
- Bora, E. (2021). A meta-analysis of theory of mind and ‘mentalizing’ in borderline personality disorder: A true neuro-social-cognitive or meta-social-cognitive impairment? *Psychological Medicine*, *51*(15), 2541–2551. <https://doi.org/10.1017/S0033291721003718>
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>

- Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2023). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods, 26*(2), 323–352. <https://doi.org/10.1177/10944281211056520>
- Buer Christensen, T., Eikenaes, I., Hummelen, B., Pedersen, G., Nysæter, T.-E., Bender, D. S., Skodol, A. E., & Selvik, S. G. (2020). Level of personality functioning as a predictor of psychosocial functioning—Concurrent validity of criterion A. *Personality Disorders: Theory, Research, and Treatment, 11*(2), 79–90. <https://doi.org/10.1037/per0000352>
- Clutterbuck, R. A., Callan, M. J., Taylor, E. C., Livingston, L. A., & Shah, P. (2021). Development and validation of the Four-Item Mentalising Index. *Psychological Assessment, 33*(7), 629–636. <https://doi.org/10.1037/pas0001004>
- Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence, 43*, 52–64. <https://doi.org/10.1016/j.intell.2014.01.004>
- Conway, J. R., Coll, M. P., Cuve, H. C., Koletsi, S., Bronitt, N., Catmur, C., & Bird, G. (2020). Understanding how minds vary relates to skill in inferring mental states, personality, and intelligence. *Journal of Experimental Psychology: General, 149*(6), 1032–1047. <https://doi.org/10.1037/xge0000704>
- Coyle, T. R., Elpers, K. E., Gonzalez, M. C., Freeman, J., & Baggio, J. A. (2018). General intelligence (g), ACT scores, and theory of mind:(ACT) g predicts limited variance among theory of mind tests. *Intelligence, 71*, 85–91. <https://doi.org/10.1016/j.intell.2018.10.006>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. <https://doi.org/10.1037/h0040957>
- Dimitrijević, A., Hanak, N., Dimitrijević, A. A., & Marjanović, Z. J. (2018). The Mentalization Scale (Ments): A self-report measure for the assessment of mentalizing capacity. *Journal of Personality Assessment, 100*(3), 268–280. <https://doi.org/10.1080/00223891.2017.1310730>
- Dodell-Feder, D., Lincoln, S. H., Coulson, J. P., & Hooker, C. I. (2013). Using fiction to assess mental state understanding: a new task for assessing theory of mind in adults. *PLoS ONE, 8*(11): e81279. <https://doi.org/10.1371/journal.pone.0081279>
- Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences, 108*(19), 7716–7720. <https://doi.org/10.1073/pnas.1018601108>
- Dufner, M., Gebauer, J. E., Sedikides, C., & Denissen, J. J. (2019). Self-enhancement and psychological adjustment: A meta-analytic review. *Personality and Social Psychology Review, 23*(1), 48–72. <https://doi.org/10.1177/1088868318756467>
- Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., ... & Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders, 36*(5), 623–636. <https://doi.org/10.1007/s10803-006-0107-0>
- Elfenbein, H. A., Barsade, S. G., & Eisenkraft, N. (2015). The social perception of emotional abilities: Expanding what we know about observer ratings of emotional intelligence. *Emotion, 15*(1), 17–34. <https://doi.org/10.1037/a0038436>
- Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*(7), 1097–1126. <https://doi.org/10.1037/0022-3514.37.7.1097>
- Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science, 16*(4), 779–788. <https://doi.org/10.1177/1745691620970586>
- German Psychological Society (Ed.). (2018). *Ethisches Handeln in der psychologischen Forschung: Empfehlungen der Deutschen Gesellschaft für Psychologie für Forschende und Ethikkommissionen [Acting ethically in psychological research: Recommendations of the German Psychological Society for researchers and ethics committees]* (1st ed.). Hogrefe. <https://doi.org/10.1026/02802-000>
- Grainger, S. A., McKay, K. T., Riches, J. C., Chander, R. J., Cleary, R., Mather, K. A., ... & Henry, J. D. (2023). Measuring Empathy Across the Adult Lifespan: A Comparison of Three Assessment Types. *Assessment*. Advance online publication. <https://doi.org/10.1177/10731911221127902>
- Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling, 25*, 621–638. <https://doi.org/10.1080/10705511.2017.1402334>
- Happé, F. (2015). Autism as a neurodevelopmental disorder of mind-reading. *Journal of the British Academy, 3*, 197–209. <https://doi.org/10.5871/jba/003.197>
- Haslam, N., McGrath, M. J., Viechtbauer, W., & Kuppens, P. (2020). Dimensions over categories: A meta-analysis of taxometric research. *Psychological Medicine, 50*(9), 1418–1432. <https://doi.org/10.1017/S003329172000183X>
- He, J. C., & Côté, S. (2023). Are Empathic People Better Adjusted? A Test of Competing Models of Empathic Accuracy and Intrapersonal and Interpersonal Facets of Adjustment Using Self-and Peer Reports. *Psychological Science*. Advance online publication. <https://doi.org/10.1177/09567976231185127>
- Horowitz, L. M., Aiden, L. E., Wiggins, J. S., & Pincus, A. L. (2000). *Inventory of Interpersonal Problems manual*. Odessa, FL: The Psychological Corporation
- Hoyle, R. H., Lynam, D. R., Miller, J. D., & Pek, J. (2023). The questionable practice of partialing to refine scores on and inferences about measures of psychological constructs. *Annual Review of Clinical Psychology, 19*, 155–176. <https://doi.org/10.1146/annurev-clinpsy-071720-015436>
- Huang, J. L., & DeSimone, J. A. (2021). Insufficient effort responding as a potential confound between survey measures and objective tests. *Journal of Business and Psychology, 36*, 807–828. <https://doi.org/10.1007/s10869-020-09707-2>
- Ickes, W. (1993). Empathic accuracy. *Journal of Personality, 61*(4), 587–610. <https://doi.org/10.1111/j.1467-6494.1993.tb00783.x>

- Johnson, B. N., Kivity, Y., Rosenstein, L. K., LeBreton, J. M., & Levy, K. N. (2022). The association between mentalizing and psychopathology: A meta-analysis of the reading the mind in the eyes task across psychiatric disorders. *Clinical Psychology: Science and Practice, 29*(4), 423–439. <https://doi.org/10.1037/cps0000105>
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*(1), 103–129. <https://doi.org/10.1016/j.jrp.2004.09.009>
- Kittel, A. F. D., Olderbak, S., & Wilhelm, O. (2022). Sty in the mind's eye: A meta-analytic investigation of the nomological network and internal consistency of the "Reading the Mind in the Eyes" test. *Assessment, 29*(5), 872–895. <https://doi.org/10.1177/1073191121996469>
- Kretzschmar, A., & Gignac, G. E. (2019). At what sample size do latent variable correlations stabilize? *Journal of Research in Personality, 80*, 17–22. <https://doi.org/10.1016/j.jrp.2019.03.007>
- Lee, K., & Ashton, M. C. (2018). Psychometric Properties of the HEXACO-100. *Assessment, 25*(5), 543–556. <https://doi.org/10.1177/1073191116659134>
- Leising, D., & Zimmermann, J. (2011). An integrative conceptual framework for assessing personality and personality pathology. *Review of General Psychology, 15*(4), 317–330. <https://doi.org/10.1037/a0025070>
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*(2), 151–173. https://doi.org/10.1207/S15328007SEM0902_1
- Lubben, J., Blozik, E., Gillmann, G., Iliffe, S., von Renteln Kruse, W., Beck, J. C., & Stuck, A. E. (2006). Performance of an abbreviated version of the Lubben Social Network Scale among three European community-dwelling older adult populations. *The Gerontologist, 46*(4), 503–513. <https://doi.org/10.1093/geront/46.4.503>
- Luyten, P., Campbell, C., Allison, E., & Fonagy, P. (2020). The mentalizing approach to psychopathology: State of the art and future directions. *Annual Review of Clinical Psychology, 16*, 297–325. <https://doi.org/10.1146/annurev-clinpsy-071919-015355>
- MacCann, C., Joseph, D. L., Newman, D. A., & Roberts, R. D. (2014). Emotional intelligence is a second-stratum factor of intelligence: evidence from hierarchical and bifactor models. *Emotion, 14*(2), 358–374. <https://doi.org/10.1037/a0034755>
- Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2020). Confirmatory factor analysis (CFA), exploratory structural equation modeling (ESEM), and set-ESEM: Optimal balance between goodness of fit and parsimony. *Multivariate Behavioral Research, 55*(1), 102–119. <https://doi.org/10.1080/00273171.2019.1602503>
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. <https://doi.org/10.1037/a0028085>
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955–966. <https://doi.org/10.1037/0003-066X.30.10.955>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Müller, S., Wendt, L. P., & Zimmermann, J. (2023). Development and validation of the certainty about Mental States Questionnaire (CAMSQ): A self-report measure of mentalizing oneself and others. *Assessment, 30*(3), 651–674. <https://doi.org/10.1177/10731911211061280>
- Murphy, B. A., Hall, J. A., & Duong, F. (2022). It looks like construct validity, but look again: Comment on Clutterbuck et al. (2021) and recommendations for test developers in the broad "empathy" domain. *Psychological Assessment, 34*(4), 397–404. <https://doi.org/10.1037/pas0001063>
- Murphy, B. A., & Lilienfeld, S. O. (2019). Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. *Psychological Assessment, 31*(8), 1062–1072. <https://doi.org/10.1037/pas0000732>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf
- Navarro, E. (2022). What is theory of mind? A psychometric study of theory of mind and intelligence. *Cognitive Psychology, 136*, 101495. <https://doi.org/10.1016/j.cogpsych.2022.101495>
- Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of Abnormal Psychology, 125*(6), 818–823. <https://doi.org/10.1037/abn0000182>
- Olderbak, S., & Wilhelm, O. (2020). Overarching principles for the organization of socioemotional constructs. *Current Directions in Psychological Science, 29*(1), 63–70. <https://doi.org/10.1177/0963721419884317>
- Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brennehan, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: Toward a brief form for research and applied settings. *Frontiers in Psychology, 6*, Article 1503. <https://doi.org/10.3389/fpsyg.2015.01503>
- Open-Source Psychometrics Project. (2020, October 1). *Vocabulary IQ test*. <https://openpsychometrics.org/tests/VIQT/>
- Osborne-Crowley, K. (2020). Social cognition in the real world: Reconnecting the study of social cognition with social reality. *Review of General Psychology, 24*(2), 144–158. <https://doi.org/10.1177/1089268020906483>
- Petrowski, K., Schmalbach, B., Kliem, S., Hinz, A., & Brähler, E. (2019). Symptom-Checklist-K-9: Norm values and factorial structure in a representative German sample. *PLoS One, 14*(4), e0213490. <https://doi.org/10.1371/journal.pone.0213490>
- Pettersson, E., Lichtenstein, P., Larsson, H., D'Onofrio, B. M., Lahey, B. B., & Latvala, A. (2021). Associations of resting heart rate and intelligence with general and specific psychopathology: a prospective population study of 899,398 Swedish men. *Clinical Psychological Science, 9*(3), 524–532. <https://doi.org/10.1177/2167702620961081>

- Pinkham, A. E., Harvey, P. D., & Penn, D. L. (2018). Social cognition psychometric evaluation: results of the final validation study. *Schizophrenia Bulletin*, 44(4), 737–748. <https://doi.org/10.1093/schbul/sbx117>
- Podsakoff, P. M., Podsakoff, N. P., Williams, L. J., Huang, C., & Yang, J. (2024). Common Method Bias: It's Bad, It's Complex, It's Widespread, and It's Not Easy to Fix. *Annual Review of Organizational Psychology and Organizational Behavior*. Online advance publication. <https://doi.org/10.1146/annurev-orgpsych-110721-040030>
- Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, 15(2), 384–396. <https://doi.org/10.1177/1745691619896607>
- R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing.
- Realo, A., Allik, J., Nolvak, A., Valk, R., Ruus, T., Schmidt, M., & Eilola, T. (2003). Mind-reading ability: Beliefs and performance. *Journal of Research in Personality*, 37(5), 420–445. [https://doi.org/10.1016/S0092-6566\(03\)00021-7](https://doi.org/10.1016/S0092-6566(03)00021-7)
- Ree, M. J., & Carretta, T. R. (2022). Thirty years of research on general and specific abilities: Still not much more than g. *Intelligence*, 91, 101617. <https://doi.org/10.1016/j.intell.2021.101617>
- Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, 95(2), 129–140. <https://doi.org/10.1080/00223891.2012.725437>
- Rogoff, S., Moulton-Perkins, A., Warren, F., Nolte, T., & Fonagy, P. (2021). 'Rich' and 'poor' in mentalizing: Do expert mentalizers exist? *PloS One*, 16(10), e0259030. <https://doi.org/10.1371/journal.pone.0259030>
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27–42. <https://doi.org/10.1177/2515245917745629>
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Salazar Kämpf, M., Adam, L., Rohr, M. K., Exner, C., & Wieck, C. (2023). A Meta-Analysis of the Relationship Between Emotion Regulation and Social Affect and Cognition. *Clinical Psychological Science*. Advance online publication. <https://doi.org/10.1177/21677026221149953>
- Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology*, 5. <https://doi.org/10.15626/MP.2019.1645>
- Schlegel, K., & Scherer, K. R. (2018). The nomological network of emotion knowledge and emotion understanding in adults: Evidence from two new performance-based tests. *Cognition and Emotion*, 32(8), 1514–1530. <https://doi.org/10.1080/02699931.2017.1414687>
- Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, 31(12), 1428–1441. <https://doi.org/10.1037/pas0000623>
- Sharp, C., & Vanwoerden, S. (2015). Hypermentalizing in borderline personality disorder: A model and data. *Journal of Infant, Child & Adolescent Psychotherapy*, 14(1), 33–45. <https://doi.org/10.1080/15289168.2015.1004890>
- Soto, C. J., Napolitano, C. M., Sewell, M. N., Yoon, H. J., & Roberts, B. W. (2022). An integrative framework for conceptualizing and assessing social, emotional, and behavioral skills: The BESSI. *Journal of Personality and Social Psychology*, 123(1), 192–222. <https://doi.org/10.1037/pspp0000401>
- Sunahara, C. S., Rosenfield, D., Alvi, T., Wallmark, Z., Lee, J., Fulford, D., & Tabak, B. A. (2022). Revisiting the association between self-reported empathy and behavioral assessments of social cognition. *Journal of Experimental Psychology: General*, 151(12), 3304–3322. <https://doi.org/10.1037/xge0001226>
- Tyrer, P., Yang, M., Tyrer, H., & Crawford, M. (2021). Is social function a good proxy measure of personality disorder? *Personality and Mental Health*, 15(4), 261–272. <https://doi.org/10.1002/pmh.1513>
- Üstün, T. B., Kostanjsek, N., Chatterji, S., Rehm, J., & World Health Organization (2010). *Measuring health and disability: manual for WHO Disability Assessment Schedule (WHODAS 2.0)*. <https://apps.who.int/iris/handle/10665/43974>
- Vachon, D. D., & Lynam, D. R. (2016). Fixing the problem with empathy: Development and validation of the affective and cognitive measure of empathy. *Assessment*, 23(2), 135–149. <https://doi.org/10.1177/1073191114567941>
- van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five inter-correlations and a criterion-related validity study. *Journal of Research in Personality*, 44(3), 315–327. <https://doi.org/10.1016/j.jrp.2010.03.003>
- Wang, X., & Navarro-Martinez, D. (2023). Increasing the external validity of social preference games by reducing measurement error. *Games and Economic Behavior*, 141, 261–285. <https://doi.org/10.1016/j.geb.2023.06.006>
- Ward, M. K., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, 74, 577–596. <https://doi.org/10.1146/annurev-psych-040422-045007>
- Watson, D., Clark, L. A., Chmielewski, M., & Kotov, R. (2013). The value of suppressor effects in explicating the construct validity of symptom measures. *Psychological Assessment*, 25(3), 929–941. <https://doi.org/10.1037/a0032781>
- Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, 84(3), 608–618. <https://doi.org/10.1037/0022-3514.84.3.608>
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS ONE*, 11(3): e0152719. <https://doi.org/10.1371/journal.pone.0152719>

How to cite this article:

Wendt, L. P., Zimmermann, J., Spitzer, C., & Müller, S. (2024). Mindreading measures misread? A multimethod investigation into the validity of self-report and task-based approaches. *Psychological Assessment*, 36(5), 365–378. <https://doi.org/10.1037/pas0001310>

Anlage 1. Erklärung gemäß § 8 der Allgemeinen Bestimmungen für Promotionen der Universität Kassel vom 14.07.2021.

1. Bei der eingereichten Dissertation zu dem Thema „Exploring and Validating Construct Interpretations of Psychological Measurements“ handelt es sich um meine eigenständig erbrachte Leistung.
2. Anderer als der von mir angegebenen Quellen und Hilfsmittel habe ich mich nicht bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen veröffentlichten oder unveröffentlichten Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Dissertation oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die abgegebenen digitalen Versionen stimmen mit den abgegebenen schriftlichen Versionen überein.
5. Ich habe mich keiner unzulässigen Hilfe Dritter bedient und insbesondere die Hilfe einer kommerziellen Promotionsberatung nicht in Anspruch genommen.
6. Die Mitwirkung von Koautoren habe ich durch eine von diesen unterschriebene Erklärung dokumentiert. Eine Übersicht, in der die einzelnen Beiträge nach Ko-Autoren und deren Anteil aufgeführt sind, füge ich anbei (siehe Anlage 2).
7. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

Datum

Unterschrift

Anlage 2. Erklärung über den Eigenanteil an den veröffentlichten oder zur Veröffentlichung vorgesehenen wissenschaftlichen Schriften innerhalb meiner Dissertationsschrift

[Statement on my contribution to the published or to be published research articles that are included in my cumulative dissertation]

Universität Kassel, Fachbereich Humanwissenschaften Erklärung zur kumulativen Dissertation im Promotionsfach Psychologie gemäß Ergänzung zu § 5a Abs. 4 Satz 1 der Allgemeinen Bestimmungen für Promotionen an der Universität Kassel vom 13. Juni 2011

(1) Antragssteller [Applicant]:

Wendt, Leon Patrick

Institut für Psychologie, Universität Kassel

Thema der Dissertation [Dissertation Topic]:

Exploring and Validating Construct Interpretations of Psychological Measurements

(2) Nummerierte Aufstellung der eingereichten Schriften [List of research articles]:

1. Wendt, L. P., Wright, A. G., Pilkonis, P. A., Woods, W. C., Denissen, J. J., Kühnel, A., & Zimmermann, J. (2020). Indicators of affect dynamics: structure, reliability, and personality correlates. *European Journal of Personality*, 34(6), 1060-1072. <https://doi.org/10.1002/per.2277>
2. Wendt, L. P., Jankowsky, K., Schroeders, U., London Personality and Mood Disorder Research Consortium, Nolte, T., Fonagy, P., Montague, P. R., Zimmermann J., and Olaru, G. (2023). Mapping established psychopathology scales onto the Hierarchical Taxonomy of Psychopathology (HiTOP). *Personality and Mental Health*, 17(2), 117-134. <https://doi.org/10.1002/pmh.1566>
3. Wendt, L. P., Zimmermann, J., Spitzer, C., & Müller, S. (2024). Mindreading measures misread? A multimethod investigation into the validity of self-report and task-based approaches. *Psychological Assessment*, 36(5), 365-378. <https://doi.org/10.1037/pas0001310>

(3) Darlegung des eigenen Anteils an diesen Schriften [Statement of own contribution]:

Zu Nr. 1: Ich bin der Erstautor. Literaturrecherche, Datenaufbereitung, Datenauswertung und Ergebnisvisualisierung wurden vollständig von mir durchgeführt. Die Erstellung des Manuskripts und die Ergebnisdiskussion wurden überwiegend von mir und in Teilen von Johannes Zimmermann, Aidan G. Wright und Jaap J. Denissen durchgeführt. Die Konzeption der Studie wurde zum Teil von mir und zum Teil von Johannes Zimmermann und Aidan G. Wright entwickelt. Bei der Studie handelt es sich um eine Sekundäranalyse bereits erhobener Daten. Die Datenerhebung wurde von Jaap J. Denissen, Anja Kühnel, Paul A. Pilkonis, Aidan G. Wright und William C. Woods durchgeführt. Alle Autoren waren an der Kommentierung der Manuskriptentwürfe beteiligt.

[Regarding no. 1: I am the first author. Literature review, data preparation, data analysis and visualisation of results were performed entirely by me. The preparation of the manuscript and discussion of the results was predominantly performed by me and in part by Johannes Zimmermann, Aidan G. Wright and Jaap J. Denissen. The concept of the study was developed in part by me and in part by Johannes Zimmermann and Aidan G. Wright. The study is a secondary analysis of previously collected data. The data collection was performed by Jaap J. Denissen, Anja Kühnel, Paul A. Pilkonis, Aidan G. Wright and William C. Woods. All authors were involved in commenting on manuscript drafts.]

Zu Nr. 2: Ich bin der Erstautor. Literaturrecherche, Datenaufbereitung und Ergebnisvisualisierung wurden vollständig von mir durchgeführt. Die Datenauswertung wurde überwiegend von mir und in Teilen von Kristin Jankowsky und Gabriel Olaru durchgeführt. Die Erstellung des Manuskripts und die Ergebnisdiskussion wurden überwiegend von mir und in Teilen von Johannes Zimmermann, Gabriel Olaru, Kristin Jankowsky, Ulrich Schroeders, Tobias Nolte und Peter Fonagy durchgeführt. Die Konzeption der Studie wurde überwiegend von mir und in Teilen von Gabriel Olaru und Johannes Zimmermann entwickelt. Bei der Studie handelt es sich um eine Sekundäranalyse bereits erhobener Daten. Die Datenerhebung wurde von Tobias Nolte, Peter Fonagy und P. Read Montague durchgeführt. Alle Autoren waren an der Kommentierung der Manuskriptentwürfe beteiligt.

[**Regarding no. 2:** I am the first author. Literature review, data preparation and visualisation of results were performed entirely by me. The data analysis was performed predominantly by me and in part by Kristin Jankowsky and Gabriel Olaru. The preparation of the manuscript and discussion of the results was performed predominantly by me and in part by Johannes Zimmermann, Gabriel Olaru, Kristin Jankowsky, Ulrich Schroeders, Tobias Nolte, and Peter Fonagy. The concept of the study was developed predominantly by me and in part by Gabriel Olaru and Johannes Zimmermann. The study is a secondary analysis of previously collected data. The data collection was performed by Tobias Nolte, Peter Fonagy and P. Read Montague. All authors were involved in commenting on manuscript drafts.]

Zu Nr. 3: Ich bin der Erstautor. Datenaufbereitung, Datenauswertung und Ergebnisvisualisierung wurden vollständig von mir durchgeführt. Die Konzeption der Studie wurde überwiegend von mir und in Teilen von Sascha Müller und Johannes Zimmermann entwickelt. Die Literaturrecherche wurde überwiegend von mir und in Teilen von Sascha Müller durchgeführt. Die Erstellung des Manuskripts und die Ergebnisdiskussion wurden überwiegend von mir und in Teilen von Sascha Müller, Johannes Zimmermann und Carsten Spitzer durchgeführt. Die Datenerhebung wurde überwiegend von mir und in Teilen von Sascha Müller, Johannes Zimmermann und Carsten Spitzer durchgeführt. Alle Autoren waren an der Kommentierung der Manuskriptentwürfe beteiligt.

[Regarding no. 3: I am the first author. Data preparation, data analysis and visualisation of the results were performed entirely by me. The concept of the study was predominantly developed by me and in parts by Sascha Müller and Johannes Zimmermann. The literature review was performed predominantly by me and in part by Sascha Müller. The preparation of the manuscript and discussion of the results was performed predominantly by me and in part by Sascha Müller, Johannes Zimmermann and Carsten Spitzer. The data collection was performed predominantly by me and in parts by Sascha Müller, Johannes Zimmermann and Carsten Spitzer. All authors were involved in commenting on manuscript drafts.]

Datum [Date], Unterschrift des Antragstellers [Signature of the Applicant]

31. Januar 2024

Anlage 3. Dokumentation der genutzten Daten

Bei den in den Artikeln 1 und 2 vorgestellten Studien handelt es sich um Sekundäranalysen, für die bereits vorhandene Datensätze verwendet wurden. Artikel 1 verwendet Daten aus drei unabhängigen Erhebungen, die erste von Jaap J. Denissen und Anja Kühnel (<https://www.psychology.hu-berlin.de/de/prof/perdev/downloadentwper/diarystudy>), die zweite und dritte von Aidan G. Wright, William C. Woods und Paul A. Pilkonis. Die verwendeten Daten aller Stichproben aus Artikel 1 wurden in einem öffentlichen Online-Projektarchiv unter der DOI <https://doi.org/10.17605/OSF.IO/6GHCX> dauerhaft verfügbar gemacht. Die Daten aus Artikel 2, verwaltet von Peter Fonagy, Tobias Nolte und Read. C. Montague, sind auf begründete Anfrage bei den Autoren erhältlich. Bei der in Artikel 3 vorgestellten Studie handelt es sich um eine Primäranalyse von Daten, die zu diesem Zweck erhoben wurden. Die für Artikel 3 verwendeten Daten wurden in einem öffentlichen Online-Projektarchiv unter der DOI <https://doi.org/10.17605/OSF.IO/Q7EU4> dauerhaft verfügbar gemacht.