DOI:
10.1007/s00702-016-1543-4

Document Version
Peer reviewed version

Citation for published version (APA):
Kaurin, A., Egloff, B., Stringaris, A., & Wessa, M. (2016). Only complementary voices tell the truth: a reevaluation of validity in multi-informant approaches of child and adolescent clinical assessments. *Journal of Neural Transmission*. https://doi.org/10.1007/s00702-016-1543-4

Complementary voices tell the truth: A reevaluation of validity in multi-informant approaches of child and adolescent clinical assessments

Aleksandra Kaurin[a,], Boris Egloff[b], Argyris Stringaris[c] & Michèle Wessa[a]

[a] Department of Clinical Psychology and Neuropsychology, Johannes Gutenberg-University, Institute of Psychology, Mainz, Germany

[b] Department of Personality Psychology and Psychological Assessment, Johannes Gutenberg-University, Institute of Psychology, Mainz, Germany

[c] Department of Child and Adolescent Psychiatry, King's College London, Institute of Psychiatry, Psychology and Neuroscience, London, UK

Correspondence concerning this article should be addressed to Aleksandra Kaurin, Department of Clinical Psychology and Neuropsychology, Johannes Gutenberg-University, Institute of Psychology, Johannes Gutenberg-University Mainz, D-55099 Mainz, Germany. E-mail: alkaurin@uni-mainz.de

Abstract

Multi-informant approaches are thought to be key to clinical assessment. Classical theories of psychological measurements assume that only convergence among different informants' reports allows for an estimate of the *true* nature and causes of clinical presentations. However, the integration of multiple accounts is fraught with problems because findings in child and adolescent psychiatry do not conform to the fundamental expectation of convergence. Indeed, reports provided by different sources (self, parents, teachers, peers) share little variance. Moreover, in some cases informant divergence may be meaningful and not error variance.

In this review we give an overview of conceptual and theoretical foundations of valid multi-informant assessment and discuss why our common concepts of validity need revaluation.

keywords: multimethod assessment; cross-informant agreement; construct validity; incremental validity; meaningful divergence

„*The problem is one of opposition between subjective and objective points of view. There is a tendency to seek an objective account of everything before admitting its reality. But often what appears to be a more subjective point of view cannot be accounted for in this way. So either the objective conception of the world is incomplete, or the subjective involves illusions that should be rejected*."

Thomas Nagel, *Subjective and Objective* in Mortal Questions (1979)

Imagine a parent consults a child psychologist because her son John has recently been displaying difficulties concentrating, headaches and irritability. The clinician may hypothesise that John's symptoms are best explained by an anxiety disorder, but how does she collect relevant information to substantiate this diagnosis and to rule out alternative diagnoses?

In order to get a comprehensive picture of John's concerns across many different situations she chooses to ask John and different persons who know him – typically relatives, peers or teachers – to report on his symptoms. The clinician obtains self-reports from John and an informant-report from his mother (method 1 and 2). Moreover, she may use her observations of his behaviour during the mildly stressful clinical assessment (method 3) and interview his teacher about John's behaviour at school (method 4). This method is commonly referred to as a *multi-informant approach* (De Los Reyes, 2013). Likely all perspectives may contribute valid observations about John's concerns. Yet, would they tell a coherent story altogether? Interviewing multiple sources informs the assessment process on a variety of different symptom levels. However, a satisfactory convergence, is rarely attained because the relationship of informants' reports is predominantly characterised by random noise (Burns & Haynes, 2006). Even if identical or parallel – i.e. psychometrically identical -- measures were applied (De Los Reyes, 2011), informants' reports share little variance (see Achenbach, McConaughy, and Howell (1987) for a comprehensive meta-analysis of correspondence between informants in 119 studies): parents' and teachers' reports overlap by approximately 15% for internalizing symptoms (with informants underestimating the presence of respective symptoms) and 30% for externalizing behaviour problems. The convergence of children's and adults' reports, however, circles around 20% for either condition (McConaughy, Stanger, & Achenbach, 1992)

Clearly, diverging accounts have adverse effects on research findings and clinical judgments: First, they result in markedly varied epidemiological estimates leading researchers to over- or underestimate prevalence rates of specific disorders (s. Polanczyk, Willcutt, Salum, Kieling, and Rohde (2014) for an a meta-analytic overview of heterogeneity in prevalence estimates in Attention Deficity Hyperactivity Disorders). Moreover, a valid evaluation of the success of clinical trials is likely to fail (Kolko & Kazdin, 1993). For instance, in 1990 the Infant Health and Development program was initiated in order to reduce health risks that are associated with low birth weight. The evaluation of this intervention was based on reports provided by mothers. These reports, however, were confounded by maternal education, thus their ability to detect and verbally express their child's health issues. It is likely, that the programme had an impact on mothers' sensitivity for the concerns of their children. Ignoring this relationship, however, led to a pattern of results where the experimental group of this randomised controlled trial had worse outcomes than the control group (see Kraemer et al. (2003) for an overview).

Second, unrecognised clinical conditions prevent an early intervention that may inhibit a) the development of a full-blown expression of the disorder or b) its chronicity (Luby, 2012; Offord et al., 1996). Especially with regards to internalizing disorders such as anxiety disorders a large proportion of children and adolescents is considered to remain unidentified (Pine, Helfinstein, Bar-Haim, Nelson, & Fox, 2009). Decreased levels of sensitivity may be traced back to the observation that some children do not express their concerns, thus informants have difficulties inferring the children's concerns (e. g. Weisbrot, Gadow, DeVincent, & Pomeroy, 2005).

Third, low cross-informant agreement raises questions about how to classify mental disorders. For instance, John's recent irritability may have gotten him into trouble with his peers due to his temper outbursts. To his teacher such behaviour may present as a symptom of a conduct disorder. John, however, may report that his excessive worry made him be more easily annoyed by others. How – on a general level – should a condition be classified that one informant reports as externalising and the subject itself as internalising disorder? What becomes evident is that in order to estimate true nomological relations of the constructs assessed, source effects need to be partitioned out from the measures, because associated biases will likely distort their covariance (see Greenbaum, Decrick, Prange, and Friedman (1994) for a comprehensive examination of source effects on the relation of internalising, thought, attention and externalising problems).

This so-called *grand discrepancy* (De Los Reyes, Thomas, Goodman, & Kundey, 2013) presents the clinician with a dilemma: Empirical science assumes that there is such a thing as *truth*. To the clinician in our example John's recent condition has a *true* underlying cause. She applies multiple instruments that are specifically designed to identify this cause (e. g. anxiety disorder). Each of these measures underwent the process of validation – a test of whether the empirical relations between test scores match the relations in the nomological network (Borsboom, 2005). Theory holds that each of the measures properly represents the construct of interest. However, if they differ so radically – which is the correct one? And, if she uses all four measures that means that neither is correct on its own (Campbell & Fiske, 1959). In any case, some part of the theory seems wrong. Yet, there is a decision to take: in order to provide John with a diagnosis that accurately determines the cause and nature of his complaints and reflects the demands of effective therapy the clinician has to meet the needs of clinical pragmatics and sacrifice her theoretical doubts.

Experience and empirical evidence tell us that clinicians are inclined to make diagnostic decisions that are in line with parent provided information, although parent- and child-provided information share little variance (DiBartolo, Albano, Barlow, & Heimberg, 1998; Grills & Ollendick, 2003; Luby, 2012; Youngstrom et al., 2004). Yet, there has been no scientific consensus on algorithms that appropriately reconcile diverging reports (De Los Reyes et al., 2013; Offord et al., 1996). Consequently, the question of how to derive valid estimates of child characteristics on the basis of collateral information has been left unresolved. As a first step towards a solution of this challenging status quo we give an overview of a) conceptual and theoretical foundations of valid multi-informant assessment and b) discuss why our common concepts of validity need revaluation. Here, we focus on child and adolescent clinical assessments in particular, because multi-informant approaches are of fundamental importance in this population.

The problem of truth.

The fact that psychological constructs are of hypothetical nature implies that they are never directly observable. Similarly, for no form of child and adolescent psychopathology a mechanism has been uncovered that allows an accurate diagnostic test. With the use of a wide range of instruments (interviews, questionnaires, standardised tests, behavioural observation and biophysiological measures) we translate the hypothesised attributes into recognisable and observable indicators (Cronbach & Meehl, 1955). Our development of these instruments is optimally driven by two theoretical prerequisites: (1) the existence of the construct of interest and (2) hypotheses about how variations in the construct causally produce variations in the outcomes, that we measure. We cannot measure a trait that does not exist (Borsboom, 2005). Also, if it exists, yet does not produce causal variations in our criterion, we may measure something completely different or nothing at all (see Block (1995) for an overview of the *Jingle-Jangle-Jungle fallacy*).

Measurement instruments can be broadly defined as vehicles "(…) *that uncover psychological attributes and procedures of objects and transform these attributes into symbols that can be processed* (…)" (Schmitt, 2006). Yet, by definition, these symbols are imperfect. Psychological measurement theories put forward that each person has a *true score* on the attribute assessed. Evidently, the average observed score of a person is only an approximation of the latent, hypothesised construct. Beyond variance that is entirely attributable to the trait of interest (Judd, Smith & Kidder, 1991; Schmidt & Hunter, 1999), this reflection, however, is assumed to contain another component: In Classical Test Theory (CTT; Lord, Novick, & Birnbaum, 1968) any discrepancy between the hypothetical *true* score and an observed estimate is explained by measurement error, a random source of variance (see Sutcliffe (1965) for the *platonic true score interpretation*). Other than the estimate of the true score, the error term varies unsystematically and becomes virtually zero when the number of measurements tends to infinity. In accordance with this equation from CTT, maximizing the number of measurements implies *approximating the truth*. More informants, in this case, increase the a) reliability and b) validity of our measurement (Roberts & Caspi, 2001).

Assessments in child and adolescent psychiatric contexts are adapted to this logic by combining multiple informants' reports. However, little convergence among informants' reports poses large challenges to the validity of multi-informant assessments. Two different explanations may explain small proportions of convergence: First, if informants' reports share approximately 20-30% *common variance*, this proportion – according to CTT – is traceable to the latent trait assessed (see Figure 1 A), because the overlap of different methods depends on how much trait specific variance each captures in relation to error variance. Then, 70-80% mirror error variance. The second approach is more fundamental: The conceptualisation of the true score as the expected value of observed scores is based on principles of the theory of errors. Generally, this theory states that repeated measurements of the exact same, constant entity lead to different results, because every measurement is characterised by error variance (Edgeworth, 1888). This principle, however, was mostly applied in astronomy and yields a major fallacy, while being transferred to psychological assessment contexts. In this case, observed scores are collected at the level of the individual and – other than flipping a coin– do not belong to a set of repeated measurements with the same instrument. Even under circumstances of repeated measures, psychometry will not satisfy the need for a fixed true score: Each measurement itself has an impact on the traits assessed, because humans – unlike celestial bodies – learn and memorise their previous responses or tire out. From this perspective, a true score cannot ever be attained at the individual level unless the subject "*were repeatedly tested in a long run of testing occasions with intermediate brainwashing and time travel*" (Borsboom, 2005; p. 45).

Against this backdrop, we may either conclude that (1) our methods are predominantly characterised by random noise (Campbell & Fiske, 1959) or (2) that CTT may not prove to be an adequate treatment of psychological test scores (Borsboom, 2005). This implies that neither method appropriately and *validly* mirrors the construct of interest. Similarly, it is possible that *at least* one method may not capture the trait assessed (Campbell & Fiske, 1959). In both cases, the capacity of each account to indicate *construct validity* is highly decreased because nomological relations of the constructs of interest are distorted by variance caused by distinct sources (Dirks, Boyle, & Georgiades, 2011; Dirks, De Los Reyes, Briggs-Gowan, Cella, & Wakschlag, 2012; Greenbaum et al., 1994). Beyond that, it is difficult to test incremental validity. That would be given when the predictability of a specific criterion is increased beyond that provided by an established method (e.g. parent-report).

However, the idea that error terms may be of *systematic* – rather than *unsystematic* – nature, further challenges our attempt to summarise individual scores within one equation.

In spite of lacking convergence, individual measures uniquely contribute to the prediction of trait-specific behaviours (Asendorpf, Banse, & Mücke, 2002; Egloff & Schmukle, 2002; Hirschmüller, Egloff, Nestler, & Back, 2013). Interestingly, not only information provided by different informants is characterised by little amounts of shared variance. Also, specific trait estimates based on different methods filled in by one and the same person show very little to no convergence (e.g. implicit and explicit measures of shyness; Asendorpf et al., 2002). This may allow disentangling the 70-80% into meaningful components of inter-informant variation (De Los Reyes, Alfano, & Beidel, 2010; Kraemer et al., 2003). Such perspective puts emphasis on epistemological issues – i.e. their ability to represent reality – of the construct under investigation because the divergence of different accounts may be meaningful because they compensate each other's shortcomings by complementary information. This information – in turn – leads to increased levels of explained trait variance. From this standpoint, traditional definitions of *traits* (Campbell & Fiske, 1959) may not apply, because variance attributable to the construct of interest is uniquely linked to specific determinants of the individual of each informant (e.g. situations in which behaviours are observed).

Truth matters.

With respect to *multi-informant approaches*, research has shown, that the act of reporting on others' or own states or traits may be biased by a variety of distinct sources like age-related limitations to introspection (Luby, Belden, Sullivan, & Spitznagel, 2007) or parental psychopathology (see Müller, Achtergarde, and Furniss (2011) for a comprehensive examination of the *depression-distortion hypothesis*). These factors are assumed to interfere with informants' ratings of the characteristics assessed. As a consequence informants may not share the same understanding of which indicators (i. e. behaviours, states) represent the construct of interest in general. Or, beyond a mutual understanding, informants may differ in their abilities and motivation to extract relevant observations from the wealth of events in everyday life (Cairns & Green, 1979). However, in the absence of a solid theory that explains processes of divergence, this work has led to mostly inconsistent results.

Yet, the fact that the vast majority of child and adolescent mental disorders is never entirely consistent across time, situations or methods (Bögels et al., 2010; Dirks et al., 2012; Kraemer et al., 2003) may help uncover explanatory mechanisms. This notion has been conceptualised as *relative consistency*, systematic behavioural variations determined by a set of situation-specific constraints. Herein may lie the *cause* for low cross-informant agreement as well as and the *solution* for this ambiguity. Variations allow to uncover the mechanisms that generate differential behaviour (Schmitt, 2006) and once uncovered, these mechanisms may help to reconcile or to integrate conflicting accounts.

Literature suggests at least two mechanisms may account for systematic variations across multiple informants: First, relevant behavioural indicators may not be equally available for all informants (Kraemer et al., 2003; Vazire, 2010). Thus, not all informants make inferences based on the same knowledge, yet their perspectives contain equally valid information for the assessment. Second, the particular approach of each informant or method may trigger different responses in the assessee. This issue has been extensively studied under the umbrella of *multidetermination of behaviour* (Shadish, Cook, & Campbell, 2002).

The idea that the individual approach of each informant may prompt different behaviours in the assessee may be best illustrated with our example. John's self-reported sleeplessness (method 1) and irritability might reflect his anxiety, yet both may also result from excessive computer-gaming sessions or hyperactivity. At home, John may progressively shut himself away from his family and this withdrawal is likely to be interpreted as a sign of anxiety or depression by his mother. Beyond that, his mother's report (method 2) may be biased by her motivation to present as a caring parent thereby exaggerating her worries and adding to John's actual symptoms. Contrasted with severe cases the clinician saw earlier that day, her spontaneous behavioural observations (method 3) may underscore John's current impairment. Moreover, because he feels uncomfortable presenting as timid and nervous towards a stranger, he will cover his anxiety. Finally – as outlined above – John's anxiety may present to his teacher as an externalising condition. However, the teacher's impression (method 4) of John's behaviour may be influenced by the sympathy for his student. If John has been an excellent student so far, the teacher may give his recent agitation a sympathetic consideration.

Clearly, each measurement depends on its respective source. Generalisability Theory (GT; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) was established as a theoretical framework to investigate the effects of multidetermination on convergence among information sources (e.g. informants, methods). According to GT, each sample of measurements represents a *universe* of all possible measurements (Cardinet, Tourneur, & Allal, 1976). With the assumption of the universe being infinite, two measurements cannot ever be identical. However, central to GT is the issue to what degree observed scores match average scores obtained under all possible circumstances. Here, variance of a test score is distinguishable into several factors, that were carefully derived from theoretical and practical considerations.

Aggregating across different informants' perspectives – and thereby across time, situations and methods – leads to a clearer reflection of the diagnostically relevant factor by controlling for multiple determinants of human behaviour (Brown, 1910; Spearman, 1910). Yet, how can these meaningful determinants be translated into research practice and clinical assessments? The introduction of Campbell's and Fiske's (1959) multitrait-multimethod matrix was a milestone for the estimation of validity of assessments based on multiple judgments. It allows contrasting variance unique to the perspective of an informant (i.e. perceptual biases due to differential presentation of symptoms across situations, person-situation-interaction) and variance attributable to the latent trait (i. e. consensual views on the basis of correlations among different assessments). Essential to this framework is the use of *converging accounts* as indicators of construct validity. The authors state that correlations among different methods of the same trait (*convergent* validity) should be high. The degree of this coefficient, however, has not been benchmarked. How can this concept be put to the test?

Jöreskog (1969) suggests to partition distinct facets of variance by a *covariance structure modeling approach,* i.e. confirmatory factor analyses (CFA). This analysis allows disentangling *trait*, *source* and *error variance* simultaneously in each individual symptom rating. An assessment is considered to be valid, if trait variance outweighs source variance. Only in this case the measurement is not inflated by variance attributable to the informants and the assessment allows to generalise across informants' individual reports (Eid, Lischetzke, Nussbeck, & Trierweiler, 2003). However, studies that systematically review the ratio of trait and source variance are few and specific patterns of results indicate the inappropriateness of MTMM or GT conceptualisations of *trait variance* for multi-informant assessments. Burns & Haynes (2006) demonstrate that in specific cases, generalisation is possible only across one set of informants: For instance, parent-ratings may consist of 10% trait and 83% source variance, whereas teacher-ratings indicate 56% trait and 28% source variance (Burns, Walsh, & Gomez, 2003; Gomez, Burns, Walsh, & De Moura, 2003). Whether strong source effects reflect situation specificity of child behaviour or measurements that are predominantly influenced by biases may – according to the authors – only be clarified with two separate CFAs: one specifying situations at school (e.g. reports provided by teachers and peers) and another specifying situations at home (e.g. reports provided by mothers and fathers; see Figure 1 B). If the strong source effects in the first analysis result from behaviour that is situation specific, then each CFA should lead to an increase of trait over source variance.

The approach of GT sets out to maximise variance attributable to the latent trait of interest. In some cases, however, it is impossible to model distinct situation specific behaviours (e.g. at school and at home) in one mathematical model, because effects of contextual variations of specific traits cannot be separated from symptom ratings that are highly contaminated by bias (Burns & Hayes, 2005). Thus, a more specific approach is necessary to capture the logic of highly, yet meaningfully, disagreeing reports.

In contrast to MTMM the *Mix and Match approach* (Kraemer et al., 2003) makes use of diverging accounts to increase the validity of the measure. It is not the sheer mass of information that reduces inaccuracy, because an infinite number of correlated (collinear) accounts cannot correct for shortcomings of each other's reports. Such a mathematical model implies that informant-reports are *never interchangeably useable*.

The authors hold that fusing diverging, independent perspectives on one individual helps to capture the whole diversity of possible indicators of the construct, thereby offsetting biases of each individual informant. Informants' reports are suggested to emerge from a function of three *orthogonal* dimensions and a random error term: In line with GT, in addition to variance explained by an unsystematic error term, unshared variance between informants may be further divided into (1) information that is unique to that informant's *perspective* (e.g. self vs. other) and (2) information that is unique to environmental circumstances, i.e. the *context* under which symptoms may be displayed (e.g. school vs. home). Consequently, a lack of convergence may be explained with the fact that one informant may have observed valid information that others do not have, which leads to less congruent accounts. Conceptualised on the grounds of *linear algebra*, the clinician may pinpoint the location of John's most approximate score if she maximised the number of non-collinear informants. Particularly, if the clinician assumed the *trait*, *context* and *perspective* to be valid dimensions of an informant's report, she will need at least three independent (orthogonally interrelated) sources to triangulate John's most approximate score on the attribute assessed.

According to this understanding, the clinician in our example can consider herself lucky if the three applied methods are incongruent and contribute unique and essential evidence to the picture, and the picture gets sharper the less correlated the perspectives are (see Figure 1 C). Only in this case, divergence among informants' reports is meaningful. Against this backdrop, the idea of CTT and GT begins to unravel because *truth* cannot accurately result from aggregation across multiple measurements. From the perspective of clinical activities, this may sound paradoxical. Yet, in terms of research, it leads to an increase of trait-specific variance by partition of variance underlying different informants' reports. By doing so, the aim of the clinical assessment (e.g. diagnostic decision, treatment response) gains in predictability. In clinical reality, however, the clinician is still lacking a set of operations that allow her to translate this evidence into a real-life, clear-cut outcome.

<< insert Figure 1 here >>

So, truth lies in the eye of the beholder?

The Mix and Match approach demonstrates that different reports may tell different, but complementary parts of the story (Klonsky & Oltmanns, 2002). Yet, how does the clinician know that the divergence is meaningful and not simply due to error?

The Self-Other Knowledge Asymmetry model (SOKA; Vazire, 2010) provides a framework of moderators to trial the differential predictive value of reports made by informants relative to those by the subject him/herself. In contrast to previously reported research, this perspective puts emphasis on the question about what specific kinds of attributes of the characteristics assessed are more precisely reported by others compared to the subject. Our clinician may significantly benefit from this approach as she may interview John, his mother and his teacher on differential aspects of his characteristics.

Based on Funder's (1995) *realistic accuracy* model an accurate estimate of the trait assessed is achieved, if four factors are consecutively realised during an assessment. First, John has to express behaviourally *relevant* indicators of the construct of interest. If we assumed he had an anxiety disorder, these could be avoidance, withdrawal and heightened vigilance. Second, these behaviours need to be *available* to his mother, teacher or the clinician. Third, any informant needs to *detect* these relevant indicators. Finally, these indicators need to be validly *utilised* by each informant. All four factors are multiplicatively related, stating that if one of them is missing (i. e. equals zero), an accurate informant rating cannot be reached (Funder, 1995, 2012). Interindividual differences of informants' judgments are assumed to be pronounced within the *availability* and *detection* components. In particular, Vazire (2010) makes two predictions: First, highly *observable* behaviours (e.g. extraversion-related talkativeness) are partly better picked up by informants, whereas traits low in observability (e.g. anxiety) are more comprehensively reported by the subject itself. Second, self- and informant ratings may have differential predictive value for traits high in *evaluativeness* – socially (un)desirable traits whose judgment poses a threat to the self-esteem of the assessee (e.g. intelligence).

In accordance with the predictions derived from the SOKA model, self-reports most accurately predicted neuroticism and in comparison informant-reports more accurately predicted extraversion and traits that were related to the intellectual abilities of the assessee (Vazire, 2010).

The evidence from this study mirrors findings in child and adolescent psychopathology research: Internalizing conditions (e.g. anxiety, depression) are assumed to be accurately reported by the child or adolescent itself (Silverman & Ollendick, 2005). Evidently, the self has a highly advantaged approach to relevant information in this case because these conditions are largely characterised by cognitive and affective processes that project little into overt behaviours. With regards to externalizing conditions, parent reports of oppositional symptoms uniquely contribute to the ODD diagnosis in addition to child-reports (Angold & Costello, 2000). Moreover, in the assessment of ADHD (combined hyperactive/impulsive subtype) the joint use of teacher- and parent-reports exceeds variance explained by parent-report alone, but the assessment of either subtype on its own did not profit from combining teacher- and parent-report (Owens & Hoza, 2003). However, in line with the suggestion made by Burns and Haynes (2006) the validity of teacher reports increases if only behaviours shown in the classroom were considered (Smith, Pelham Jr, Gnagy, Molina, & Evans, 2000).

Also, for traits high in *evaluativeness* such as social skills both teacher- and peer-ratings demonstrated incremental value in a sample of third- to five-graders (Kwon, Kim, & Sheridan, 2012).

A framework towards the integration of meaningful divergence.

Another – perhaps more radical – perspective on the divergence of different measures of the same construct is provided by dual-process theories of human behaviour and cognition.

These theories suggest, that specific behaviours may be described as a function of two distinct mechanisms (e.g. Kahneman, 2003)

To illustrate, Back, Schmukle, and Egloff (2009) introduced the *Behavioural Process Model of Personality* (BPMP). This model extents the Reflective-Impulsive Model of decision making (see Strack and Deutsch (2004) for an overview) to the domain of personality. According to the BPMP, stable individual differences in social behaviour can be understood as the result of the typical functioning (across time and multiple situations) of reflective processes (how people typically perceive and categorise situations, which behavioural options they prefer, and how they deliberately realise these preferences) and impulsive processes (how situational cues are automatically processed, and what kinds of actions are automatically performed), which jointly trigger social behaviour.

These stable individual differences in information-processing also affect individuals' beliefs about themselves (i.e. their self-concepts). Presumably, individual differences in the typical operation of reflective processes can be translated into differences in propositional representations of the self (i.e., the explicit self-concept of personality), which are measured with standard direct measures (e.g., questionnaires). The typical functioning of impulsive processes, by contrast, leads to chronic links between semantic network elements, and thus, differences in associative representations of the self (i.e., the implicit self-concept of personality), which are assessed with indirect measures (e.g., Implicit Association tests for assessing personality).

Our example again serves to illustrate how reflective and impulsive processes distinctively manifest within one person. The clinician asks John to fill in a questionnaire about his experienced levels of anxiety. Also, she indirectly assesses his anxiety with an implicit test where he is asked to sort words of anxious and non-anxious content to categories of the *self* or *other* respectively. Because John wants to remain his image as someone who is confident or because he may trace back his symptoms to a physiological cause or simply because he feels uncomfortable talking about his concerns he may (deliberately) underscore his recent levels of anxiety in his self-report. The implicit test, however, allows to control for faking tendencies or response biases due to low levels of face validity. Also, this approach uncovers automatic and non-conscious aspects of John's implicit self-concept that he cannot be aware of. These non-conscious aspects may include processes of *evaluative conditioning*. Here emotional contents of words or objects are semantically associated with another stimulus. In our example words like afraid, nervous, anxious, uncertain or fearful may be tied to John's implicit self-representations thus leading to quicker reaction times in the sorting task, when anxious words need to be paired with the self vs. other. As a consequence, he may provide the clinician with two estimates of his anxiety that do not overlap at all.

Following this line of reasoning, individual differences in the explicit and implicit self-concept, as measured by direct and indirect tests of personality, are condensations of typical differences in reflective and impulsive processes that predict social behaviour. Both may be conceptualised as functional subfacets of the constructs of interest. It then follows that implicit and explicit measures of e.g. anxiety may be only slightly correlated (even when corrected for unreliability of measurement) because both operate at distinct levels of perception, thus differ in their explicability. Moreover, each measure predicts unique variance in behaviour (see Figure 1 D). For example, Asendorpf et al. (2002) showed that an IAT for measuring shyness uniquely predicted spontaneous shyness behaviours whereas self-reported shyness uniquely predicted controlled aspects of shyness behaviours (so-called double dissociation). Similar findings were obtained by (Egloff & Schmukle, 2002) in the domain of anxiety and by Back et al. (2009) for the 'Big Five' personality traits (see also Hirschmüller et al., 2013). Thus, the divergence of two measures constitutes no problem at all – to the contrary, the divergence is meaningful and allows for incremental and unique predictions of behaviour.

Discussion

In view of the fact that informants' reports are characterised by little agreement, we set out to review concepts of validity in multi-informant assessment contexts. Our aim was to exemplify why these concepts impose limits for collateral data integration and to present a framework that allows combining diverging assessment information for a valid comprehensive clinical judgment.

We demonstrated that in contrast to general assumptions made by Classical Test Theory (Lord & Novick, 1968), Generalisability Theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972) and the Multitrait-Multimethod approach (Campbell & Fiske, 1959) trait variance and trait indicative behaviours can be incrementally predicted by different reports that share little to no variance (Mix and Match approach, Kraemer et al., 2003; Self-Other Knowledge Asymmetry model, Vazire, 2010; Behavioural Process Model of Personality, Back, Schmukle & Egloff 2009). At least two aspects in this discussion of validity, however, warrant further attention:

First, the meaningful combination of informants' reports leads to increases of trait variance up to levels of 50% in Kraemer et al. (2003). But, a benchmark that defines the maximally possible amount of explained trait variance has not yet been established. With that said, one could only speculate about the nature of the remaining 50%. With regards to the multidetermination of human behaviour, trait indicators were reported to have small effect sizes in the prediction of behaviour (Ahadi & Diener, 1989). Similarly, given the high contextual variability of clinical conditions (e.g. Bögels et al., 2010) we may assume that much higher levels of explained trait variance cannot be reached. However, because Kraemer et al. (2003) did not control for the unreliability of each measure applied and not all informants were provided with questionnaires that had 1) the same psychometric properties, 2) similar contents and 3) constant time frames of symptom reports, it is likely that in this particular study the unexplained variance mirrors methodological artefacts to great extents.

Second, with regards to the BPMP it is possible that not all indicative behaviours are captured by established measures of clinical and research practice. This question of content validity, however, is difficult to answer, because research in this domain exhibits a strong single-method approach. When it comes to the validation of new instruments researchers repeatedly chose to establish how much variance is shared with a gold-standard measure of the same construct. The *tautology* of this approach becomes highly evident, when the items of both methods are semantically similar (or even the same). Such an approach sheds light on very specific aspects of the trait assessed. As a result, little evidence is unveiled that may inform construct validity and conclusions are restricted to this operationalization, because very specific aspects of the construct assessed are illuminated (Burns & Hayes, 2005). From this perspective, high levels of clinical, pathophysiological and behavioural *heterogeneity* may be a result of little construct validity (see Corvin et al. (2013) for a discussion of heterogeneity in schizophrenia). This aspect emphasises the importance of *divergence* on a more general level: Evidently, the agreement between John's mother and his teacher about his anxiety alone is not sufficient for a valid assessment. Importantly, their reports need to discriminate between the trait assessed and other factors. Yet, this step in the process of validation is much more difficult to achieve. The divergence of two methods indicates their discriminant validity only to the extent that the attributes under investigation are *truly* unrelated. In the absence of valid measures, a solid theory that specifies nomological relations among different constructs is therefore indispensible (Schmitt, 2006). With regards to the descriptive approach applied in clinical research, this line, however, is blurred. The clinician from our example relies on a lot of questions about phenomena that are related to an anxiety disorder. But these phenomena may also have a range of other causes (Pickles & Angold, 2003; Block, 1995). For instance, irritability is represented in six different psychiatric childhood disorders – both, internalising and externalising (Stringaris, 2015). The overlap of symptoms across different conditions may present as *diagnostic overshadowing bias* to clinical reality. Also, anxiety disorders are likely to be missed by clinicians in children with Autism Spectrum Disorders, because both conditions are characterised by irritability, fear and avoidance (Mason & Scior, 2004). Similarly, in research designs that explore the incremental value of an additional measurement, the problem of *criterion contamination* arises (Garb, 2005). A criterion is labeled as contaminated if predictors and criteria are not independent of each other. For instance, if we aim at predicting the clinical diagnosis from clinical files and parent reports, contamination occurs if the clinician based her judgment on this information.

Promising findings about the complementary use of multi-informant assessment in child and adolescent psychiatry illuminate an encouraging research direction in this field. Future studies, however, need to carefully control for methodological confounds in order to validly estimate the incremental value of each informants' report.

Conclusion

In classical theories of psychological measurements only convergence among different informants' reports indicates an approximation of the *true* nature and causes of mental health concerns. However, behavioural problems present themselves in different ways across different situations. As a consequence, divergence among informants' reports is considered to be meaningful, if each perspective uniquely explains trait-related variance or contributes to the prediction of behaviour. Different informants tell different, yet complementary parts of one true story and it remains an important task of clinical practice and research to develop sophisticated algorithms that allow a meaningful integration of diverging information.

Figure Caption

Figure 1. Heuristic illustrations of different concepts of validity proposed by Classical Test Theory (A), Generalisability Theory (B), the Mix and Match Approach (C) and the Behavioural Process Model of Personality (D).

Note. X and Y: informants/methods; X_1/Y_1 *and* X_2/Y_2 *multiple assessments across same sources; Z = construct assessed;* Z_A *and* Z_B *= functional subfacets of the constructs assessed; dashed lines denote trait-variance exclusively explained by one informant/method.*

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioural and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological bulletin, 101*(2), 213.

Angold, A., & Costello, E. J. (2000). The child and adolescent psychiatric assessment (CAPA). *Journal of the American Academy of Child & Adolescent Psychiatry, 39*(1), 39-48.

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: the case of shy behaviour. *Journal of personality and social psychology, 83*(2), 380.

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting actual behaviour from the explicit and implicit self-concept of personality. *Journal of personality and social psychology, 97*(3), 533.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, 117(2), 187.

Bögels, S. M., Alden, L., Beidel, D. C., Clark, L. A., Pine, D. S., Stein, M. B., & Voncken, M. (2010). Social anxiety disorder: questions and answers for the DSM- V. *Depression and anxiety, 27*(2), 168-189.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Brown, W. (1910). Some Experimental Results in the Correlation of Mental Abilities1. *British Journal of Psychology, 1904-1920, 3*(3), 296-322.

Burns, G. L., & Haynes, S. N. (2006). Clinical psychology: Construct validation with multiple sources of information and multiple settings. In M. Eid & E. Diener (Eds.), Handbook of multimethod measurement in psychology (pp. 401–418). Washington, DC: American Psychological Association.

Burns, G. L., Walsh, J. A., & Gomez, R. (2003). Convergent and discriminant validity of trait and source effects in ADHD-inattention and hyperactivity/impulsivity measures across a 3-month interval. *Journal of Abnormal Child Psychology*, *31*(5), 529-541.

Cairns, R., & Green, J. (1979). How to assess personality and social patterns: Observations or ratings. *The analysis of social interactions: Methods, issues, and illustrations*, 209-226.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, 119-135.

Corvin, A., Buchanan, R. W., Carpenter, W. T., Kennedy, J. L., Keshavan, M. S., MacDonald, A. W., Sass, L. & Wessa, M. (2013). Which aspects of heterogeneity are useful to translational success? In S. M. Silverstein, B. Moghaddam & T. Wykes (Ed.), Schizophrenia – Evolution and Synthesis (pp. 77-92). Cambridge, Massachusetts: The MIT Press.

Cronbach, Gleser, G., Nanda, H., & Rajaratnam, N. (1972). Theory of generalizability for scores and profiles. The dependability of behavioural measurements: New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

De Los Reyes, A. (2011). Introduction to the special section: More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child & Adolescent Psychology, 40*(1), 1-9.

De Los Reyes, A. (2013). Strategic objectives for improving understanding of informant discrepancies in developmental psychopathology research. *Development and psychopathology, 25*(03), 669-682.

De Los Reyes, A., Alfano, C. A., & Beidel, D. C. (2010). The relations among measurements of informant discrepancies within a multisite trial of treatments for childhood social phobia. *Journal of Abnormal Child Psychology, 38*(3), 395-404.

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology, 9*, 123-149.

DiBartolo, P. M., Albano, A. M., Barlow, D. H., & Heimberg, R. G. (1998). Cross-informant agreement in the assessment of social phobia in youth. *Journal of Abnormal Child Psychology, 26*(3), 213-220.

Dirks, M. A., Boyle, M. H., & Georgiades, K. (2011). Psychological symptoms in youth and later socioeconomic functioning: do associations vary by informant? *Journal of Clinical Child & Adolescent Psychology, 40*(1), 10-22.

Dirks, M. A., De Los Reyes, A., Briggs- Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Annual research review: Embracing not erasing contextual variability in children's behaviour–theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry, 53*(5), 558-574.

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society*, 51(3), 599-635.

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an Implicit Association Test for assessing anxiety. *Journal of personality and social psychology, 83*(6), 1441.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: a multiple-indicator CT-C (M-1) model. *Psychological methods, 8*(1), 38.

Funder, D. C. (1995). On the accuracy of personality judgment: a realistic approach. *Psychological review, 102*(4), 652.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21*(3), 177-182.

Garb, H. N. (2005). Clinical Judgment and Decision Making*. *Annu. Rev. Clin. Psychol. 2005, 1*, 67-89.

Gomez, R., Burns, G. L., Walsh, J. A., & De Moura, M. A. (2003). Multitrait-multisource confirmatory factor analytic approach to the construct validity of ADHD rating scales. *Psychological Assessment, 15*(1), 3.

Greenbaum, P. E., Decrick, R. F., Prange, M. E., & Friedman, R. M. (1994). Parent, teacher, and child ratings of problem behaviours of youngsters with serious emotional disturbances. *Psychological Assessment, 6*(2), 141.

Grills, A. E., & Ollendick, T. H. (2003). Multiple informant agreement and the anxiety disorders interview schedule for parents and children. *Journal of the American Academy of Child & Adolescent Psychiatry, 42*(1), 30-40.

Hirschmüller, S., Egloff, B., Nestler, S., & Back, M. D. (2013). The dual lens model: A comprehensive framework for understanding self–other agreement of personality judgments at zero acquaintance. *Journal of personality and social psychology, 104*(2), 335.

Jöreskog, K. G. (1969). Efficient estimation in image factor analysis. *Psychometrika, 34*(1), 51-75.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). Research methods in social relations. Sixth Edition. New York, NY: Holt, Rinehart, and Winston.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American psychologist, 58*(9), 697.

Klonsky, E. D., & Oltmanns, T. F. (2002). Informant- reports of personality disorder: Relation to self- reports and future research directions. *Clinical Psychology: Science and Practice*, *9*(3), 300-311.

Kolko, D. J., & Kazdin, A. E. (1993). Emotional/behavioural problems in clinic and nonclinic children: correspondence among child, parent and teacher reports. *Journal of Child Psychology and Psychiatry, 34*(6), 991-1006.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry, 160*(9), 1566-1577.

Kwon, K., Kim, E. M., & Sheridan, S. M. (2012). A contextual approach to social skills assessment in the peer group: Who is the best judge? *School Psychology Quarterly, 27*(3), 121.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores.

Luby, J. L. (2012). Dispelling the “they’ll grow out of it” myth: implications for intervention. *American Journal of Psychiatry, 169*(11), 1127-1129.

Luby, J. L., Belden, A., Sullivan, J., & Spitznagel, E. (2007). Preschoolers’ contribution to their diagnosis of depression and anxiety: Uses and limitations of young child self-report of symptoms. *Child psychiatry and human development, 38*(4), 321-338.

Mason, J., & Scior, K. (2004). ‘Diagnostic overshadowing’amongst clinicians working with people with intellectual disabilities in the UK. *Journal of Applied Research in Intellectual Disabilities*, *17*(2), 85-90.

McConaughy, S. H., Stanger, C., & Achenbach, T. M. (1992). Three-Year Course of Behavioural/Emotional Problems in a National Sample of 4− to 16-Year-Olds: I. Agreement among Informants. *Journal of the American Academy of Child & Adolescent Psychiatry, 31*(5), 932-940.

Müller, J. M., Achtergarde, S., & Furniss, T. (2011). The influence of maternal psychopathology on ratings of child psychiatric symptoms: an SEM analysis on cross-informant agreement. *European child & adolescent psychiatry, 20*(5), 241-252.

Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., & Lipman, E. L. (1996). Integrating assessment data from multiple informants. *Journal of the American Academy of Child & Adolescent Psychiatry, 35*(8), 1078-1085.

Owens, J. S., & Hoza, B. (2003). The role of inattention and hyperactivity/impulsivity in the positive illusory bias. *Journal of Consulting and Clinical Psychology, 71*(4), 680.

Pickles, A., & Angold, A. (2003). Natural categories or fundamental dimensions: On carving nature at the joints and the rearticulation of psychopathology. *Development and psychopathology, 15*(03), 529-551.

Pine, D. S., Helfinstein, S. M., Bar-Haim, Y., Nelson, E., & Fox, N. A. (2009). Challenges in developing novel treatments for childhood disorders: lessons from research on anxiety. *Neuropsychopharmacology, 34*(1), 213-228.

Polanczyk, G. V., Willcutt, E. G., Salum, G. A., Kieling, C., & Rohde, L. A. (2014). ADHD prevalence estimates across three decades: an updated systematic review and meta-regression analysis. *International journal of epidemiology, 43*(2), 434-442.

Roberts, B. W., & Caspi, A. (2001). Personality Development and the Person-Situation Debate: It's Déjà Vu All Over Again. *Psychological Inquiry, 12*(2), 104-109.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence,* 27(3), 183-198.

Schmitt, M. (2006). Conceptual, theoretical, and historical foundations of multimethod assessment. In M. Eid & E. Diener (Ed.), Handbook of multimethod measurement in psychology (pp. 9-25). Washington, DC: American Psychological Association.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*: Wadsworth Cengage learning.

Silverman, W. K., & Ollendick, T. H. (2005). Evidence-based assessment of anxiety and its disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 380-411.

Smith, B. H., Pelham Jr, W. E., Gnagy, E., Molina, B., & Evans, S. (2000). The reliability, validity, and unique contributions of self-report by adolescents receiving treatment for attention-deficit/hyperactivity disorder. *Journal of Consulting and Clinical Psychology*, *68*(3), 489.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, *3*(3), 271-295.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behaviour. *Personality and social psychology review*, *8*(3), 220-247.

Stringaris, A. (2015). Emotion regulation and emotional disorders: conceptual issues for clinicians and neuroscientists in Rutter's Child and Adolescent Psychiatry, Sixth Edition, Eds. Thapar A, Pine DS, Leckman JF, Scott S, Snowling MJ, Taylor EA. Wiley-Blackwell.

Sutcliffe, J. P. (1965). A probability model for errors of classification. I. General considerations. *Psychometrika*, 30(1), 73-96.

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of personality and social psychology, 98*(2), 281.

Weisbrot, D. M., Gadow, K. D., DeVincent, C. J., & Pomeroy, J. (2005). The presentation of anxiety in children with pervasive developmental disorders. *Journal of Child & Adolescent Psychopharmacology, 15*(3), 477-496.

Youngstrom, E. A., Findling, R. L., Calabrese, J. R., Gracious, B. L., Demeter, C., Bedoya, D. D., & Price, M. (2004). Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. *Journal of the American Academy of Child & Adolescent Psychiatry, 43*(7), 847-858.

Acknowledgements

AK wants to thank Stefan Berti, Henning Müller and Jan Matti Dollbaum for their helpful comments on a version of this paper.