



King's Research Portal

DOI:

[10.1007/s00702-016-1543-4](https://doi.org/10.1007/s00702-016-1543-4)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

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>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

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 reevaluation.

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

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

5 Thomas Nagel, *Subjective and Objective* in *Mortal Questions* (1979)
6

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

11 In order to get a comprehensive picture of John’s concerns across many different
12 situations she chooses to ask John and different persons who know him – typically relatives,
13 peers or teachers – to report on his symptoms. The clinician obtains self-reports from John
14 and an informant-report from his mother (method 1 and 2). Moreover, she may use her
15 observations of his behaviour during the mildly stressful clinical assessment (method 3) and
16 interview his teacher about John’s behaviour at school (method 4). This method is commonly
17 referred to as a *multi-informant approach* (De Los Reyes, 2013). Likely all perspectives may
18 contribute valid observations about John’s concerns. Yet, would they tell a coherent story
19 altogether? Interviewing multiple sources informs the assessment process on a variety of
20 different symptom levels. However, a satisfactory convergence, is rarely attained because the
21 relationship of informants’ reports is predominantly characterised by random noise (Burns &
22 Haynes, 2006). Even if identical or parallel – i.e. psychometrically identical -- measures were
23 applied (De Los Reyes, 2011), informants’ reports share little variance (see Achenbach,
24 McConaughy, and Howell (1987) for a comprehensive meta-analysis of correspondence
25 between informants in 119 studies): parents’ and teachers’ reports overlap by approximately
26 15% for internalizing symptoms (with informants underestimating the presence of respective
27 symptoms) and 30% for externalizing behaviour problems. The convergence of children’s and
28 adults’ reports, however, circles around 20% for either condition (McConaughy, Stanger, &
29 Achenbach, 1992)

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

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

52 Third, low cross-informant agreement raises questions about how to classify mental
53 disorders. For instance, John's recent irritability may have gotten him into trouble with his
54 peers due to his temper outbursts. To his teacher such behaviour may present as a symptom of
55 a conduct disorder. John, however, may report that his excessive worry made him be more

66 easily annoyed by others. How – on a general level – should a condition be classified that one
67 informant reports as externalising and the subject itself as internalising disorder? What
68 becomes evident is that in order to estimate true nomological relations of the constructs
69 assessed, source effects need to be partitioned out from the measures, because associated
70 biases will likely distort their covariance (see Greenbaum, Decrick, Prange, and Friedman
71 (1994) for a comprehensive examination of source effects on the relation of internalising,
72 thought, attention and externalising problems).

73 This so-called *grand discrepancy* (De Los Reyes, Thomas, Goodman, & Kundey,
74 2013) presents the clinician with a dilemma: Empirical science assumes that there is such a
75 thing as *truth*. To the clinician in our example John's recent condition has a *true* underlying
76 cause. She applies multiple instruments that are specifically designed to identify this cause (e.
77 g. anxiety disorder). Each of these measures underwent the process of validation – a test of
78 whether the empirical relations between test scores match the relations in the nomological
79 network (Borsboom, 2005). Theory holds that each of the measures properly represents the
80 construct of interest. However, if they differ so radically – which is the correct one? And, if
81 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

82 estimates of child characteristics on the basis of collateral information has been left
83 unresolved. As a first step towards a solution of this challenging status quo we give an
84 overview of a) conceptual and theoretical foundations of valid multi-informant assessment
85 and b) discuss why our common concepts of validity need reevaluation. Here, we focus on
86 child and adolescent clinical assessments in particular, because multi-informant approaches
87 are of fundamental importance in this population.

88 ***The problem of truth.***

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

101 Measurement instruments can be broadly defined as vehicles “(...) *that uncover*
102 *psychological attributes and procedures of objects and transform these attributes into*
103 *symbols that can be processed (...)*” (Schmitt, 2006). Yet, by definition, these symbols are
104 imperfect. Psychological measurement theories put forward that each person has a *true score*
105 on the attribute assessed. Evidently, the average observed score of a person is only an
106 approximation of the latent, hypothesised construct. Beyond variance that is entirely
107 attributable to the trait of interest (Judd, Smith & Kidder, 1991; Schmidt & Hunter, 1999),

108 this reflection, however, is assumed to contain another component: In Classical Test Theory
109 (CTT; Lord, Novick, & Birnbaum, 1968) any discrepancy between the hypothetical *true* score
110 and an observed estimate is explained by measurement error, a random source of variance
111 (see Sutcliffe (1965) for the *platonic true score interpretation*). Other than the estimate of the
112 true score, the error term varies unsystematically and becomes virtually zero when the number
113 of measurements tends to infinity. In accordance with this equation from CTT, maximizing
114 the number of measurements implies *approximating the truth*. More informants, in this case,
115 increase the a) reliability and b) validity of our measurement (Roberts & Caspi, 2001).

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

134 perspective, a true score cannot ever be attained at the individual level unless the subject
135 “were repeatedly tested in a long run of testing occasions with intermediate brainwashing
136 and time travel” (Borsboom, 2005; p. 45).

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

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

150 In spite of lacking convergence, individual measures uniquely contribute to the
151 prediction of trait-specific behaviours (Asendorpf, Banse, & Mücke, 2002; Egloff &
152 Schmukle, 2002; Hirschmüller, Egloff, Nestler, & Back, 2013). Interestingly, not only
153 information provided by different informants is characterised by little amounts of shared
154 variance. Also, specific trait estimates based on different methods filled in by one and the
155 same person show very little to no convergence (e.g. implicit and explicit measures of
156 shyness; Asendorpf et al., 2002). This may allow disentangling the 70-80% into meaningful
157 components of inter-informant variation (De Los Reyes, Alfano, & Beidel, 2010; Kraemer et
158 al., 2003). Such perspective puts emphasis on epistemological issues – i.e. their ability to
159 represent reality – of the construct under investigation because the divergence of different

160 accounts may be meaningful because they compensate each other's shortcomings by
161 complementary information. This information – in turn – leads to increased levels of
162 explained trait variance. From this standpoint, traditional definitions of *traits* (Campbell &
163 Fiske, 1959) may not apply, because variance attributable to the construct of interest is
164 uniquely linked to specific determinants of the individual of each informant (e.g. situations in
165 which behaviours are observed).

166 ***Truth matters.***

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

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

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

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

208 Clearly, each measurement depends on its respective source. Generalisability Theory
209 (GT; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) was established as a theoretical
210 framework to investigate the effects of multidetermination on convergence among
211 information sources (e.g. informants, methods). According to GT, each sample of

212 measurements represents a *universe* of all possible measurements (Cardinet, Tourneur, &
213 Allal, 1976). With the assumption of the universe being infinite, two measurements cannot
214 ever be identical. However, central to GT is the issue to what degree observed scores match
215 average scores obtained under all possible circumstances. Here, variance of a test score is
216 distinguishable into several factors, that were carefully derived from theoretical and practical
217 considerations.

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

231 Jöreskog (1969) suggests to partition distinct facets of variance by a *covariance*
232 *structure modeling approach*, i.e. confirmatory factor analyses (CFA). This analysis allows
233 disentangling *trait*, *source* and *error variance* simultaneously in each individual symptom
234 rating. An assessment is considered to be valid, if trait variance outweighs source variance.
235 Only in this case the measurement is not inflated by variance attributable to the informants
236 and the assessment allows to generalise across informants' individual reports (Eid,
237 Lischetzke, Nussbeck, & Trierweiler, 2003). However, studies that systematically review the

238 ratio of trait and source variance are few and specific patterns of results indicate the
239 inappropriateness of MTMM or GT conceptualisations of *trait variance* for multi-informant
240 assessments. Burns & Haynes (2006) demonstrate that in specific cases, generalisation is
241 possible only across one set of informants: For instance, parent-ratings may consist of 10%
242 trait and 83% source variance, whereas teacher-ratings indicate 56% trait and 28% source
243 variance (Burns, Walsh, & Gomez, 2003; Gomez, Burns, Walsh, & De Moura, 2003).
244 Whether strong source effects reflect situation specificity of child behaviour or measurements
245 that are predominantly influenced by biases may – according to the authors – only be clarified
246 with two separate CFAs: one specifying situations at school (e.g. reports provided by teachers
247 and peers) and another specifying situations at home (e.g. reports provided by mothers and
248 fathers; see Figure 1 B). If the strong source effects in the first analysis result from behaviour
249 that is situation specific, then each CFA should lead to an increase of trait over source
250 variance.

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

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

262 The authors hold that fusing diverging, independent perspectives on one individual
263 helps to capture the whole diversity of possible indicators of the construct, thereby offsetting

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

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

288

289

<< insert Figure 1 here >>

290

291 ***So, truth lies in the eye of the beholder?***

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

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

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

316 *evaluativeness* – socially (un)desirable traits whose judgment poses a threat to the self-esteem
317 of the assessee (e.g. intelligence).

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

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

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

339 ***A framework towards the integration of meaningful divergence.***

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

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

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

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

362 Our example again serves to illustrate how reflective and impulsive processes
363 distinctively manifest within one person. The clinician asks John to fill in a questionnaire
364 about his experienced levels of anxiety. Also, she indirectly assesses his anxiety with an
365 implicit test where he is asked to sort words of anxious and non-anxious content to categories
366 of the *self* or *other* respectively. Because John wants to remain his image as someone who is
367 confident or because he may trace back his symptoms to a physiological cause or simply

368 because he feels uncomfortable talking about his concerns he may (deliberately) underscore
369 his recent levels of anxiety in his self-report. The implicit test, however, allows to control for
370 faking tendencies or response biases due to low levels of face validity. Also, this approach
371 uncovers automatic and non-conscious aspects of John's implicit self-concept that he cannot
372 be aware of. These non-conscious aspects may include processes of *evaluative conditioning*.
373 Here emotional contents of words or objects are semantically associated with another
374 stimulus. In our example words like afraid, nervous, anxious, uncertain or fearful may be tied
375 to John's implicit self-representations thus leading to quicker reaction times in the sorting
376 task, when anxious words need to be paired with the self vs. other. As a consequence, he may
377 provide the clinician with two estimates of his anxiety that do not overlap at all.

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

393

394

Discussion

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

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

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

420 Second, with regards to the BPMP it is possible that not all indicative behaviours are
421 captured by established measures of clinical and research practice. This question of content
422 validity, however, is difficult to answer, because research in this domain exhibits a strong
423 single-method approach. When it comes to the validation of new instruments researchers
424 repeatedly chose to establish how much variance is shared with a gold-standard measure of
425 the same construct. The *tautology* of this approach becomes highly evident, when the items of
426 both methods are semantically similar (or even the same). Such an approach sheds light on
427 very specific aspects of the trait assessed. As a result, little evidence is unveiled that may
428 inform construct validity and conclusions are restricted to this operationalization, because
429 very specific aspects of the construct assessed are illuminated (Burns & Hayes, 2005). From
430 this perspective, high levels of clinical, pathophysiological and behavioural *heterogeneity*
431 may be a result of little construct validity (see Corvin et al. (2013) for a discussion of
432 heterogeneity in schizophrenia). This aspect emphasises the importance of *divergence* on a
433 more general level: Evidently, the agreement between John’s mother and his teacher about his
434 anxiety alone is not sufficient for a valid assessment. Importantly, their reports need to
435 discriminate between the trait assessed and other factors. Yet, this step in the process of
436 validation is much more difficult to achieve. The divergence of two methods indicates their
437 discriminant validity only to the extent that the attributes under investigation are *truly*
438 unrelated. In the absence of valid measures, a solid theory that specifies nomological relations
439 among different constructs is therefore indispensable (Schmitt, 2006). With regards to the
440 descriptive approach applied in clinical research, this line, however, is blurred. The clinician
441 from our example relies on a lot of questions about phenomena that are related to an anxiety
442 disorder. But these phenomena may also have a range of other causes (Pickles & Angold,
443 2003; Block, 1995). For instance, irritability is represented in six different psychiatric
444 childhood disorders – both, internalising and externalising (Stringaris, 2015). The overlap of
445 symptoms across different conditions may present as *diagnostic overshadowing bias* to

446 clinical reality. Also, anxiety disorders are likely to be missed by clinicians in children with
447 Autism Spectrum Disorders, because both conditions are characterised by irritability, fear and
448 avoidance (Mason & Scior, 2004). Similarly, in research designs that explore the incremental
449 value of an additional measurement, the problem of *criterion contamination* arises (Garb,
450 2005). A criterion is labeled as contaminated if predictors and criteria are not independent of
451 each other. For instance, if we aim at predicting the clinical diagnosis from clinical files and
452 parent reports, contamination occurs if the clinician based her judgment on this information.

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

457

458

Conclusion

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

467 **Figure Caption**

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

471

472 *Note. X and Y: informants/methods; X₁/Y₁ and X₂/Y₂ multiple assessments across same*
473 *sources; Z = construct assessed; Z_A and Z_B = functional subfacets of the constructs assessed;*
474 *dashed lines denote trait-variance exclusively explained by one informant/method.*

475 **References**

476

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

480 Angold, A., & Costello, E. J. (2000). The child and adolescent psychiatric assessment
481 (CAPA). *Journal of the American Academy of Child & Adolescent Psychiatry*, 39(1),
482 39-48.

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

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

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

491 Bögels, S. M., Alden, L., Beidel, D. C., Clark, L. A., Pine, D. S., Stein, M. B., & Voncken, M.
492 (2010). Social anxiety disorder: questions and answers for the DSM- V. *Depression*
493 *and anxiety*, 27(2), 168-189.

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

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

498 Burns, G. L., & Haynes, S. N. (2006). Clinical psychology: Construct validation with multiple

499 sources of information and multiple settings. In M. Eid & E. Diener (Eds.), *Handbook*
500 *of multimethod measurement in psychology* (pp. 401–418). Washington, DC:
501 American Psychological Association.

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

505 Cairns, R., & Green, J. (1979). How to assess personality and social patterns: Observations or
506 ratings. *The analysis of social interactions: Methods, issues, and illustrations*, 209-
507 226.

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

510 Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory:
511 Applications to educational measurement. *Journal of Educational Measurement*, 119-
512 135.

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

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

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

523 De Los Reyes, A. (2011). Introduction to the special section: More than measurement error:
524 Discovering meaning behind informant discrepancies in clinical assessments of

525 children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, 40(1),
526 1-9.

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

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

533

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

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

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

543 Dirks, M. A., De Los Reyes, A., Briggs- Gowan, M., Cella, D., & Wakschlag, L. S. (2012).
544 Annual research review: Embracing not erasing contextual variability in children's
545 behaviour—theory and utility in the selection and use of methods and informants in
546 developmental psychopathology. *Journal of Child Psychology and Psychiatry*, 53(5),
547 558-574.

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

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

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

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

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

559 Garb, H. N. (2005). Clinical Judgment and Decision Making*. *Annu. Rev. Clin. Psychol.*
560 *2005*, 1, 67-89.

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

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

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

570 Hirschmüller, S., Egloff, B., Nestler, S., & Back, M. D. (2013). The dual lens model: A
571 comprehensive framework for understanding self–other agreement of personality
572 judgments at zero acquaintance. *Journal of personality and social psychology*, 104(2),
573 335.

574 Jöreskog, K. G. (1969). Efficient estimation in image factor analysis. *Psychometrika*, 34(1),
575 51-75.

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

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

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

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

586 Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J.
587 (2003). A new approach to integrating data from multiple informants in psychiatric
588 assessment and research: Mixing and matching contexts and perspectives. *American*
589 *Journal of Psychiatry*, 160(9), 1566-1577.

590 Kwon, K., Kim, E. M., & Sheridan, S. M. (2012). A contextual approach to social skills
591 assessment in the peer group: Who is the best judge? *School Psychology Quarterly*,
592 27(3), 121.

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

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

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

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

- 602 McConaughy, S. H., Stanger, C., & Achenbach, T. M. (1992). Three-Year Course of
603 Behavioural/Emotional Problems in a National Sample of 4- to 16-Year-Olds: I.
604 Agreement among Informants. *Journal of the American Academy of Child &*
605 *Adolescent Psychiatry, 31(5)*, 932-940.
- 606 Müller, J. M., Achtergarde, S., & Furniss, T. (2011). The influence of maternal
607 psychopathology on ratings of child psychiatric symptoms: an SEM analysis on cross-
608 informant agreement. *European child & adolescent psychiatry, 20(5)*, 241-252.
- 609 Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., & Lipman,
610 E. L. (1996). Integrating assessment data from multiple informants. *Journal of the*
611 *American Academy of Child & Adolescent Psychiatry, 35(8)*, 1078-1085.
- 612 Owens, J. S., & Hoza, B. (2003). The role of inattention and hyperactivity/impulsivity in the
613 positive illusory bias. *Journal of Consulting and Clinical Psychology, 71(4)*, 680.
- 614 Pickles, A., & Angold, A. (2003). Natural categories or fundamental dimensions: On carving
615 nature at the joints and the rearticulation of psychopathology. *Development and*
616 *psychopathology, 15(03)*, 529-551.
- 617 Pine, D. S., Helfinstein, S. M., Bar-Haim, Y., Nelson, E., & Fox, N. A. (2009). Challenges in
618 developing novel treatments for childhood disorders: lessons from research on anxiety.
619 *Neuropsychopharmacology, 34(1)*, 213-228.
- 620 Polanczyk, G. V., Willcutt, E. G., Salum, G. A., Kieling, C., & Rohde, L. A. (2014). ADHD
621 prevalence estimates across three decades: an updated systematic review and meta-
622 regression analysis. *International journal of epidemiology, 43(2)*, 434-442.
- 623 Roberts, B. W., & Caspi, A. (2001). Personality Development and the Person-Situation
624 Debate: It's Déjà Vu All Over Again. *Psychological Inquiry, 12(2)*, 104-109.
- 625 Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence,*
626 *27(3)*, 183-198.

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

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

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

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

639 Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology,*
640 *1904-1920, 3*(3), 271-295.

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

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

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

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

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

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

658

659

660 **Acknowledgements**

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