When diagnosing and treating learning and behavior disabilities in children, intelligence tests are commonly used diagnostic tools. Based on the results of these tests and they are frequently used in educational and vocational contexts for important decisions that impact persons’ academic or professional lives. Research has however shown that important errors may occur despite the application of validation processes and adherence to quality criteria for psychometric tests. At the same time this evidence seems not to be pervasively acknowledged in psychological practice and research. In this article, I will showcase research that places attention on sources of measurement error in pediatric intelligence testing, discuss a process-performance approach to measurement in intelligence testing, and propose the “pretest methods,” methods stemming from the field of survey methodology commonly used in questionnaire construction, as a method to help address the problem of sources of measurement error in pediatric intelligence testing and improve the development of these intelligence tests.

Keywords
Psychometric tests, intelligence, measurement error, child diagnostics, pretesting
understanding of the questions could be very fruitful. In vain, and specifically looking at the response process and that a careful and detailed scrutiny of the test items is not in children” (Burns and O’Leary, 2004: 234). This example shows Comprehension subtest is poorly understood by some children but also performance on intelligence tests (Weiss et al., 2016). In the following paragraphs I will discuss a few specific sources of measurement error in pediatric intelligence testing and where possible specifically the Wechsler scales found in questionnaire construction, understanding of the question or task at hand by the respondent, is seen as a very important aspect when preventing measurement error (see e.g. De Leeuw et al., 2008). Comprehension by itself (understanding of the survey questions), is quite an extensive concern in survey methodology (e.g. Sudman et al., 1996; Tourangeau et al., 2000).

The WISC-V (age range 6–16 years) continues the evolution of the WISC series and represents a substantial update over the previous version. However, independent analysis of the WISC-V has resulted in several substantial critiques of the test and information presented in the manuals (for a review, see Goldstein et al., 2019). Specifically relevant for present discussion is that the WISC-V manual presents traditional metrics of reliability that are biased when multidimensionality is present (Raykov, 1997). Importantly, while an understanding of the abilities being tapped by the index and subtest scores are fundamental to using this test, these scores must also be placed in the larger context of each child and their “world.” There are a huge number of endogenous factors such as genetics and a host of external factors such as culture and education that impact not only the growth and expression of intelligence, but also performance on intelligence tests (Weiss et al., 2016).

**Sources of measurement error in pediatric intelligence testing**

According to Weiler et al. (2019), several sources of error may obfuscate the validity of pediatric neuropsychological evaluations, including issues of incremental validity, demographic characteristics of children, ecological validity, potential for malingering, and errors of clinical decision making. Test of intelligence are generally an essential part of such a neuro-psychological evaluation. Notwithstanding, an up-to-date comprehensive overview of sources of measurement error specifically for pediatric intelligence testing is missing (but see e.g. Hanna et al., 1981; Styck and Walsh, 2016). Hanna et al.’s (1981) estimated the magnitude of four sources of measurement error, namely content sampling, time sampling, scoring and administration in the WISC-R. Some other sources of measurement error in the Wechsler scales may be factors related to the interaction between the examiner and the examinee and situational factors (Styck and Walsh, 2016). Hanna et al.’s (1981) composite was found to produce a far larger standard error of measurement than reported by the manual (Wechsler, 1974). Current estimates of measurement error for most IQ tests rely solely on internal-consistency reliability and therefore the standard error of measurement published in test manuals may not adequately capture error (Styck and Walsh, 2016).

In the following paragraphs I will discuss a few specific sources of measurement error in pediatric intelligence testing and where possible specifically the Wechsler scales found in empirical research.
Emotional and motivational variables and interaction with pediatric test characteristics

Already in the seventies, it was shown that emotional stress can result in decreases in cognitive performance in children (see e.g. Fuerst et al., 1990; Mandler, 1975, 1984; Meyer et al., 1989). Problem-solving and memory processes involved in aptitude testing had as well been shown to be affected by stress levels in children (Kaufman, 1979).

For example, the Arithmetic, Coding and Digit Span sub-tests of the WISC-R had been found in numerous studies to be more susceptible to anxiety, depression, and distractibility than the other WISC-R subtests (e.g. Kaufman, 1975, 1979; Ogdon, 1982; Stewart and Moely, 1983). In two other studies, Plante et al. (1993) and Plante and Sykora (1994) demonstrated that stress and coping variables were associated with aptitude and achievement testing where the WISC-R and WJ-R were used as measures, and with WISC-III performance respectively. However, the studies indicated the necessity of further research to examine the role of specific childhood stressors on testing performance. The instigator of these studies was an experiment on the effects of Thorazine on Wechsler scores of adult catatonic schizophrenics. Gilgash (1957) found that the intellectual functioning of the experimental group was decidedly improved after the introduction of Thorazine and 30days of Thorazine medication. Upon these results, it is hypothesized that, as schizophrenia is usually regarded as an affective rather than an intellectual disorder, the improvement in intellectual functioning was due to an improvement in the patient’s affective responsiveness, which indicates that emotional variables have an impact on the results of an IQ test.

In Steinmayer et al.’s (2014) study, SEM analyses revealed that all in their study included motivational variables played a role in explaining the relation between gender and numerical intelligence in a population of 305 German pupils (average age 17.54 years, SD = 1.08). Specifically, self-estimated numerical intelligence, intrinsic value of math, and worry significantly predicted numerical intelligence. Gender however, still explained a statistically significant amount of variance in numerical intelligence. This could be due on the other hand to other aspects, such as stereotype threat, socialization, and characteristics of the test such as time restriction (Steinmayer et al., 2014). Carr and Davis (2001) proposed that boys compared to girls display better scores if there is a time restriction regarding the tasks on standardized tests, and Wilhelm and Engle (2005) show that imposing a time limit on an intelligence test might threaten its validity. Differences on certain intelligence tests between culturally different groups are explained to a large degree by time limits that are too strict, which is largely explained by emotional and motivational factors. When looking at a specific emotional variable, test anxiety, it has been demonstrated that anxious students in particular suffer from time constraints and that their performance increases when time constraints are relaxed (see Onwuegbuzie and Seaman, 1995).

Fear of failure, among other motivational variables, might function as a significant mediator of the relationship between performance on the administered intelligence test and gender in adolescents (Steinmayer and Spinath, 2008). Motivational variables have been often identified as mediators of the relationship between gender and achievement in standardized math achievement tests (Steinmayer et al., 2014). These motivational variables then interact with traits of the test, such as the mentioned time restrictions. Looking at another aspect of test design, namely test instructions, results of experimental studies (e.g. Johnson et al., 2012) have shown that specific test instructions can be used to increase girl’s performance on different mathematical achievement tests.

In sum, a characteristic of the test, in this case time restriction, or test instruction, interacts with personal characteristics of the subject, favoring certain people with certain characteristics over others, incurring in measurement error and therefore a loss in validity. This would apply to differences between groups, but also to characteristics that may or may not be attributed differently to different groups and thus not easily identifiable and operationalizable. In the case of mentioned gender-research, we have an easily identifiable and measurable variable—gender—that directs the attention to certain differences, where only an in-depth scrutiny points to other variables as the causal attribution. At the same time, other factors that might influence the results of an IQ test might have not been identified and escaped our attention.

Several motivational and emotional variables have been found to be significant predictors of performance in standardized mathematical performance tests in schools in the target population of the WISC, such as expectations for success (e.g. Dickhäuser and Reinhard, 2009); interest and mathematical ability self-concept (e.g. Steinmayer and Spinath, 2009); and test anxiety (Hembree, 1990). It awaits further research to investigate whether they could also be predictors of performance in numerical and general intelligence tests.

Examinee interest and cooperation throughout the session (Glutting et al., 1994) and familiarity with the examiner (Fuchs and Fuchs, 1989; Szarko et al., 2013) have also been shown to be related to variance in FSIQ scores in children.

Segal (2006) asserts that low-stakes test scores, measured in surveys, may be partially determined by test-taking motivation, which is associated with personality traits and not with cognitive ability in a mixed population of mostly minors. Moreover, correlations found in survey data between high test scores during adolescence and economic success later in life seem partially caused by favorable personality traits. The relevance of motivation concerning performance
on pediatric intelligence tests has been investigated also by Duckworth et al. (2011) in a meta-analysis. The authors found an effect of material incentives on performance in the intelligence test in a large study population ($n=2008$) of which 1912 minors ($<18$) and partly using the WISC. Moreover, observer ratings of test motivation were associated with both WISC-R scores and important life outcomes in a population of adolescent boys ($n=251$, average age 12.5 years). Children who tried harder earned higher WISC-R scores. Thus, performance in WISC-R varies depending on the child’s motivation in the testing situation and furthermore test motivation “can act as a third-variable confound that inflates estimates of the predictive validity of intelligence for life outcomes” (Duckworth et al., 2011: 7716). Importantly, Extrinsic motivation thus seems to be important for performance on pediatric intelligence tests.

**Measurement error in pediatric tests related to the scorer**

Examiner error or scorer error is a source of error in the use of intelligence testing that has been amply researched, also specifically for Wechsler pediatric scales (e.g. McDermott et al., 2014; Styck and Walsh, 2016). Examiner or scorer error may have an unsystematic component, but may also be influenced by a more systematic component, such as a halo effect, a generosity error, race and sex stereotypes or other sources of consistent error (Hanna et al., 1981). Just as in a questionnaire, also in a WISC-test, to an unknown extent “responses given are a function of the manner in which the examiner conducts the examination” (Kaspar et al., 1968: 475).

In an early observational study on examiner coding errors on individual test taking with WISC and Binet protocols, Warren and Brown (1972) use a checklist of errors frequently encountered that they developed over years of experience. A total of 1939 errors were found in the 1873 subtests examined, 350 of which were consequential for the IQ reported. Some of these types of errors were: failure to record response, failure to follow procedure given in manual, scoring and tabulating. Some of these different types of examiner errors have been investigated later on in different studies, namely clerical errors, scoring errors, administration errors, failure to record examinee responses as an error, or “correctly” scoring open-ended items (for an overview see Styck and Walsh, 2016), but not altogether.

However, it is unclear how these findings have been translated into practice or improvement of IQ-tests. Practicing psychologists do not always seem to consider the possibility of measurement error in adult or pediatric intelligence testing or specifically measurement error related to themselves—the assessor. In a recent meta-analysis, Styck and Walsh (2016) conclude that assessor errors occur frequently and impact index and FSIQ scores on the WISC. Similarly, McDermott et al. (2014) found that nearly all WISC-IV scores conveyed significant amounts of variation that could not be accounted to children’s actual differences and that the FSIQ and VCI (Verbal Comprehension Index) scores evidenced substantial assessor bias. Consequently, current estimates for the standard error of measurement of the WISC may not adequately capture the variance due to the examiner.

**A process-performance approach to intelligence test results: Stable trait or performance and interaction with the test environment?**

Generally, the results of an intelligence test are considered an indicator of a stable characteristic (Eckert et al., 2006), also by practicing (school) psychologists and in schools. However, some authors indicate that it should be also taken as a kind of performance indicator (Eckert et al., 2006; Steinmayr et al., 2014). In intelligence testing it is commonly ignored that individuals differ among each other concerning motivational factors such as academic self-concept, which endangers the validity of intelligence testing, as research shows that test results under certain conditions are influenced by motivational factors (Eckert et al., 2006).

There have been studies about personality and performance in multiple-choice testing and other achievement tests. Kubinger and Wolfsbauer (2010: 303) contend that from the point of view of personality psychology, examinees in a multiple-choice test are likely to differ in the way they deal with multiple choice items, which might impact their results. Dochy et al. (2001) on their side presume that, if an examinee reaches a particular answer for an item in a multiple-choice test, which, however, is not offered as an option, then, the higher the examinee’s general assertiveness is, the more he/she uses “none of the other options is right.” Ávila and Torrubia (2004) established that extraverted, impulsive, and low-anxiety examinees give more incorrect responses in multiple-choice tasks but make fewer omission errors and Alker et al. (1969) indicate that examinees with a “nonconformist” personality score higher on achievement tests than others because of fewer skipped items and not at all using the “I don’t know the solution” option. Lastly, Stoeber and Kersting (2007) showed that “perfectionist” examinees attain higher scores in achievement tests.

To the best of my knowledge, similar researches have not been carried out in the investigation of personality, emotional and motivational variables, and response process and performance in pediatric intelligence testing. Different personality types might nevertheless also react differently to an intelligence test, because of which the design of the test considering response process and the interaction with the responder deserves attention, such as commonly is done in questionnaire construction procedures (e.g. De Leeuw et al., 2008).

Thus, it is standard to interpret test scores on intelligence tests as an ability and a rather stable trait. However, intelligence test results are preceded by a response process in which persons work on intelligence test tasks and give a
performance that—to a better or worse extent—may reflect their intelligence. From this point of view, a score on an intelligence test can be interpreted as an indicator of performance on this intelligence test (Steinmayr et al., 2014: 141), reflecting both the response process and the potential. This response process, then, is a result of the interaction of aspects such as the test characteristics, personality, and motivational and emotional variables.

Validity: Multidimensionality

From the psychometric point of view, “(…) it has long been known (Stout, 1987) that items, for example in a test, an interview or behavior observation, do not only capture the intended trait but also other aspects, for example, other traits or specific of the test person (…)” (Ziegler and Hagemann, 2015: 232). That is, assuring the unidimensionality of an item is an extremely difficult if not impossible task (Ziegler and Hagemann, 2015: 232). Though when constructing items researchers at least try to ensure the items are not loaded with other traits, unfortunately addressing this aspect is not part of general procedures on test construction (Ziegler, 2014). Following the ABC of test construction according to Ziegler (2014), psychometric criteria such as reliability and validity are of large importance, to which Ziegler and Hagemann (2015) added unidimensionality, which can however be seen as generally treated as a sub aspect. Concerning reliability, reliable findings not necessarily translate into valid findings. Validation syntheses have shown that internal structure and relations to other variables have been the dominant sources of evidences in the literature (e.g. Greiff and Scherer, 2018; Zumbo and Chan, 2014). The first issue here is that even these procedures could not suffice. EFA (Exploratory Factor Analysis) is one of the classical approaches to use in test construction. EFA results obtained during test construction are often used to select items (Ziegler and Hagemann, 2015: 234). With regards to unidimensionality, this item selection is deemed to ensure that no other relevant variance sources are part of the items. Ziegler and Hagemann (2015) show that this can be misleading if all items contain the same (two) sources of variance (see Ziegler and Hagemann, 2015, figures 1a–f: 232–233). Thus, all items loading on one factor does not necessarily imply unidimensionality of the items.

When considering (rather) stable person traits, such as traits that can affect the response process such as perfectionism, fear of failure or insecurity, that is, motivational and emotional variables, or non-stable person traits such as mood, but that may remain stable over the course of the whole response process, these could be such a source of variance contained in all items. This would be overlooked by current common test construction procedures, and would remain unidentified even with repeated measurements over the years in the case the related person traits would not change in this period. There is thus room for a multidimensionality which can be caused by person traits or person related variables.

Nonetheless a step forward compared to EFA, similar considerations hold for CFA (Confirmatory Factor Analysis), and IRT (Item Response Analysis). In response to this, Ziegler and Hagemann (2015) propose certain solutions, using MTMM and other methods. These solutions might however also have their limitations.

This issue can be additionally looked at from the viewpoint of test fairness (Stout, 2002). A procedure to increase test fairness is to do differential item functioning analyses (DIF), and related, differential bundle functioning (DBF) and differential test functioning (DTF) analyses. “(…) by definition, DIF occurs when examinees, matched on the latent variable θ that the test is intended to measure, perform differentially depending on their group membership.” (Stout, 2002: 501). This, however, would become a problem when group membership is not clear. When test fairness is focused at groups formed based on variables such as gender, social and cultural background, or race, then these are straightforward identifiable and measurable variables. The picture changes when it concerns groups that are not directly observable or even measurable, or are measurable but have not yet been (extensively) identified and measured. As would be the case with groups based on motivational or emotional variables, or personality. A test fairness analysis requires a prerequisite validity analysis, that is, any DIF/DBF/DTF analysis is not effective in the case of a serious contamination of the subtest caused by secondary dimensions, moreover, any DIF/DBF/DTF procedure, “is only as effective as the matching subtest is in matching examinees on the construct intended to be measured.” (Stout, 2002: 501).

Multidimensionality and the response process

Unidentified secondary dimensions or confounding variables thus pose a limitation toward differential functioning analyses. “(…) In the prototypical large standardized test setting where one wants to assess pretest items for possible DIF, matching examinees on the score on operational items, which have previously undergone careful test design and psychometric scrutiny, seems reasonable” (Stout, 2002: 501). However, if also in these prior developed tests or test items secondary dimensions remain unidentified, then this is of no use. In survey research, this phenomenon is referred to as response bias and is an extensively researched phenomenon, also with qualitative procedures. Problematic response bias occurs when the variables related to the response bias are not known and thus not measured.

Ziegler (2015: 156) also gave attention to the matter of response bias where it concerns psychological assessment in the form of for example personality tests used for the selection of job applicants, stating that “(…) all research dealing with such phenomena [response bias] should take the answer process into account.” His focus was thus on response bias related to the test takers in what Cattell called Q-data (Cattell, 1958), and not yet on psychometric intelligence testing.
Cattell (1958: 286) defined Q-data as: “Observations of personality which come to the psychologist in terms of retrospective, verbal, self-record and self-evaluation. . .”

Nevertheless, it seems that response processes and test consequences have been largely neglected in validation practice in the area of psychology (Hubley, 2018; Zumbo and Chan, 2014), though there has been some research into sources of model misfit that point to test characteristics and the response process, namely item wording (DiStefano and Motl, 2006; Rauch et al., 2007; Vautier et al., 2003) and position effect (Hartig et al., 2007; Schweizer et al., 2009). These aspects indicate correlated errors (which in turn might indicate undiscovered secondary dimensions) which are however normally not taken into account in test construction (Schweizer, 2010).

On the other hand, item wording and position and their effect on outcomes are extensively researched in survey methodology (see e.g. Schwarz et al., 1991).

Survey research and practices and research in questionnaire construction have particularly addressed the role of psychology in the response process and its relationship with measurement outcomes. Specifically, survey methodologists have placed attention on understanding the cognitive and communicative processes underlying survey responses, increasingly turning the “art of asking questions” (Payne, 1951) into an applied science that is grounded in basic psychological research.

Thus, with the unsolved problem of multidimensionality and empirical research that shows the influence of extraneous variables on measurement in pediatric psychometric tests in mind, it seems fruitful to start looking more closely and in complementary alternative ways at the validity of psychometric tests like the WISC. One way to do this would be to look at the response process such as generally done in the practice of questionnaire construction and survey methodology (see e.g. De Leeuw et al., 2008), by considering results on intelligence tests as indicators of performance in an intelligence test, and considering that variables like motivation are involved in the process that forms that performance. These are then considered confounding variables that impact the target measure. Of course, as Ziegler and Hagemann (2015) propose, these variables can be investigated as independent variables in an experiment, with the target measure as a dependent variable. “An experimental effect of the independent variables on the dependent variables would reveal multidimensionality of the target variable in terms of contamination with undesirable constructs.” (Ziegler and Hagemann, 2015: 236). To be able to do this however, these variables have to first be identified and operationalized. Besides, many of the variables that seem relevant from previous research are not experimentally manipulable.

In conclusion

Research indicates that motivational and emotional variables, but also several other variables, have a contribution to performance in pediatric intelligence testing with the WISC.

These sources of measurement error have moreover not received a lot of attention in psychological practice, and, therefore have not seeped through to this practice, where results of a pediatric IQ test such as the WISC-V are generally considered to be error-free. Neither seem have findings to have received a lot of attention in the development— and research into the development— of psychometric intelligence tests. If they have, it has been from a quantitative psychometric viewpoint, which has not been able to solve the issue satisfactorily.

Psychometric analyses use mathematical models to investigate a complex real-world phenomenon. An additional complementary viewpoint exists that can help address the problem of measurement error in pediatric psychometric testing. Measurement error due to secondary dimensions can be considered and analyzed as a continuous and quantitative variable resulting from the interaction between test taker personality and situational demands, like Ziegler et al. (2015) propose when doing research on faking in personality tests. It can also be quantitatively analyzed based on different response patterns or response styles or characteristics of the test. Yet another complementary way of researching this phenomenon is focusing on and exploring the answer process and cognitive processes taking place when completing an IQ test from a qualitative viewpoint as with questionnaire pretesting methods.

Using procedures from the field of questionnaire construction and survey research as substantive analyses of test
bias, has the potential of achieving one of the imperatives for conducting test fairness research and practice according to Stout (2002: 503), namely integrating substantive analyses with psychometric analyses. It would be interesting to investigate whether standardized test procedures in IQ tests like the WISC-V, including standardized test instructions or time restriction in certain subtests, incur in bias, favoring or disadvantaging children with certain characteristics, including emotional and motivational traits. Complementary to and hopefully in synergy with psychometric procedures, this can be researched focusing on the response process and cognitive processes taking place when completing such an intelligence test.

**Author’s note**
The author is employed at the Section for Research Methods of the Department of Special Education and Rehabilitation of the Faculty of Human Sciences, University of Cologne.

**Declaration of conflicting interests**
The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**
The author received no financial support for the research, authorship, and/or publication of this article.

**ORCID iD**
Vanessa Torres van Grinsven [https://orcid.org/0000-0003-4693-5197](https://orcid.org/0000-0003-4693-5197)

**References**
Alker HA, Carlson JA and Hermann MG (1969) Multiple-choice questions and student characteristics. *Journal of Education & Psychology* 60(3): 231–243.

Ávila C and Torrubia R (2004) Personality, expectations, and response strategies in multiple-choice question examinations in university students: A test of Gray’s hypotheses. *European Journal of Personality* 18(1): 45–59.

Baron IS (2005) Test review: Wechsler intelligence scale for children-fourth edition (WISC-IV). *Child Neuropsychology* 11: 471–475.

Burns TG and O’Leary SD (2004) Wechsler intelligence scale for children–IV: Test review. *Applied Neuropsychology* 11(4): 233–236.

Campanelli P (2008) Testing survey questions. In: De Leeuw E, Hox JJ and Dillman D (eds) *International Handbook of Survey Methodology*. London: Taylor and Francis, pp.176–200.

Carr M and Davis H (2001) Gender differences in arithmetic strategy use: A function of skill and preference. *Contemporary Educational Psychology* 26: 330–347.

Cattell RB (1958) What is “objective” in “objective personality tests”? *Journal of Counseling Psychology* 5(4): 285–289.

De Leeuw ED, Hox JJ and Dillman D (2008) *International Handbook of Survey Methodology*. New York: Taylor and Francis.

Dickhäuser O and Reinhard MA (2009) How need for cognition affects the formation of performance expectancies at school. *Social Psychology of Education* 12: 385–395.

DiStefano C and Motl RW (2006) Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling A Multidisciplinary Journal* 13(3): 440–464.

Dochy F, Moerkerke G, DeCorte E, et al. (2001) The assessment of quantitative problem-solving skills with “none of the above”-items (NOTA items). *European Journal of Psychology of Education* 16: 163–177.

Duckworth AL, Quinn PD, Lynam DR, et al. (2011) Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences of the United States of America* 108(19): 7716–7720.

Eckert C, Schilling D and S dumpfmeier-Pelster J (2006) Einfluss des Fähigkeitselektionskonzep tes auf die Intelligenz- und Konzentrationstestsleistung [the influence of academic self-concept on performance in intelligence and concentration tests]. *Zeitschrift für Pädagogische Psychologie* 20: 41–48.

Eysenck H (1973) *The Measurement of Intelligence*. Baltimore, MD: Williams and Wilkins.

Freeman AJ and Chen YL (2019) Interpreting pediatric intelligence tests: A framework from evidence-based medicine. In: Goldstein G, Allen DN and DeLuca J (eds) *Handbook of Psychological Assessment*. Elsevier Academic Press, pp.65–101.

Fuchs D and Fuchs LS (1989) Effects of examiner familiarity on black, Caucasian, and Hispanic children: A meta-analysis. *Exceptional Children* 55(4): 303–308.

Fuerst DR, Fisk JL and Rourke BP (1990) Psychosocial functioning of learning-disabled children: Relations between WISC verbal IQ-performance, IQ discrepancies, and personality subtypes. *Journal of Consulting and Clinical Psychology* 58: 657–660.

Gilgash CA (1957) Effects of thorazing on wechsler scores of adult catatonic schizophrenics. *Psychological Reports* 3: 561–564.

Glutting JJ, Oakland T and Konold TR (1994) Criterion-related bias with the guide to the assessment of test-session behavior for the WISC-III and WIAT: Possible race/ethnicity, gender, and SES effects. *Journal of School Psychology* 32(4): 355–369.

Goldstein G, Allen DN and DeLuca J (2019) *Handbook of Psychological Assessment*. London: Academic Press.

Greiff S and Scherer R (2018) Still comparing apples with oranges? Some thoughts on the principles and practices of measurement invariance testing [editorial]. *European Journal of Psychological Assessment* 34(3): 141–144.

Hanna GS, Bradley FO and Holen MC (1981) Estimating major sources of measurement error in individual intelligences scales. Taking our heads out of the sand. *Journal of School Psychology* 19(4): 370–376.

Hartig J, Hölzel B and Moosbrugger H (2007) A confirmatory analysis of item reliability trends (CAIRT): Differentiating true score and error variance in the analysis of item context effects. *Multivariate Behavioral Research* 42(1): 157–183.

Hembree R (1990) The nature, effects, and relief of mathematic anxiety. *Journal for Research in Mathematics Education* 21(1): 33–46.

Hulley AM (2018, June) Missed Opportunities in Testing and Assessment: Response Processes and Test Consequences.
Stewart KJ and Moely BE (1983) The WISC-R third factor: What does it mean? *Journal of Consulting and Clinical Psychology* 51(6): 940–941.

Stoeber J and Kersting M (2007) Perfectionism and aptitude test performance: Testees who strive for perfection achieve better test results. *Personality and Individual Differences* 42(6): 1093–1103.

Stout W (1987) A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika* 52(4): 589–617.

Stout W (2002) Psychometrics: From practice to theory and back: 15 years of nonparametric multidimensional IRT, DIF/test equity, and skills diagnostic assessment. *Psychometrika* 67(4): 485–518.

Styck KM and Walsh SM (2016) Evaluating the prevalence and impact of examiner errors on the wechsler scales of intelligence: A meta-analysis. *Psychological Assessment* 28(1): 3–17.

Sudman S, Bradburn NM and Schwarz N (1996) *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass.

Szarko JE, Brown AJ and Watkins MW (2013) Examiner familiarity effects for children with autism spectrum disorders. *Journal of Applied School Psychology* 29: 37–51.

Tourangeau R, Rips LJ and Rasinski K (2000) *The Psychology of Survey Response*. Cambridge, MA: Cambridge University Press.

Vautier S, Raufaste E and Cariou M (2003) Dimensionality of the revised life orientation test and the status of filler items. *International Journal of Psychology* 38(6): 390–400.

Vernon PA (1987) *Speed of Information Processing and Intelligence*. Norwood, NJ: Ablex.

Warren SA and Brown WG (1972) Examiner scoring errors on individual intelligence tests. *Psychology in the Schools* 9: 118–122.

Watkins MW, Wilson SM, Kotz KM, et al. (2006) Factor structure of the Wechsler intelligence scale for children—fourth edition among referred students. *Educational and Psychological Measurement* 66(6): 975–983.

Wechsler D (1974) *Wechsler Intelligence Scale for Children–Revised*. New York: Psychological Corporation.

Wechsler D (1991) *Wechsler Intelligence Scale for Children–Third edn*. San Antonio, TX: Psychological Corporation.

Wechsler D (2003) *Wechsler Intelligence Scale for Children–Fourth Edition*. San Antonio, TX: Psychological Corporation.

Wechsler D (2014) *WISC-V: Technical and Interpretive Manual*. Bloomington, MN: Pearson.

Weiler MD, Willis WG and Kennedy ML (2019) Sources of error and meaning in the pediatric neuropsychological evaluation. In: Goldstein G, Allen DN and DeLuca J (eds) *Handbook of Psychological Assessment*. San Diego, CA: Elsevier Academic Press, pp.193–226.

Weiss LC, Holdnack JA, Saklofske DH, et al. (2016) Theoretical and clinical foundations of the WISC-V index scores. In: Weiss LG, Saklofske DH, Holdnack JA, et al. (eds) *WISC-V Assessment and Interpretation: Scientist-Practitioner Perspectives*. Amsterdam, The Netherlands: Elsevier Academic Press, pp.97–121.

Wilhelm O and Engle RW (2005) *Handbook of Understanding and Measuring Intelligence*. Thousand Oaks, CA: SAGE.

Ziegler M (2014) Stop and state your intentions!: Let’s not forget the ABC of test construction [editorial]. *European Journal of Psychological Assessment* 30(4): 239–242.

Ziegler M (2015) “F*** you, I won’t do what you told me!”—response biases as threats to psychological assessment [editorial]. *European Journal of Psychological Assessment* 31(3): 153–158.

Ziegler M and Hagemann D (2015) Testing the unidimensionality of items: Pitfalls and loopholes [editorial]. *European Journal of Psychological Assessment* 31(4): 231–237.

Ziegler M, Maaß U, Griffith R, et al. (2015) What is the nature of faking? Modeling distinct response patterns and quantitative differences in faking at the same time. *Organizational Research Methods* 18(4): 679–703.

Zumbo BD and Chan EKH (eds) (2014) *Social Indicators Research Series: Vol. 54. Validity and Validation in Social, Behavioral, and Health Sciences*. Switzerland: Springer International Publishing.

**Author biography**

Dr. Vanessa Torres van Grinsven currently is a research fellow at the Section for Research Methods of the Department of Special Education and Rehabilitation of the Faculty of Human Sciences at the University of Cologne. Her current research interests go out mainly to the response process in psychometric measurements and using insights and procedures from qualitative research to improve the operationalization of constructs in quantitative research and quantitative measurement instruments.