Rejoinder

School Psychology International

On the validation of ability measures in school psychology: Do established psychometric standards matter?

School Psychology International
2021, Vol. 42(2) 210–216

DOI: 10.1177/0143034320985209
journals.sagepub.com/home/spi
SAGE

Ryan J. McGill and Thomas J. Ward
William & Mary School of Education, USA

Gary L. Canivez
Eastern Illinois University, USA

Abstract

In this rejoinder, we address Kettler's comments regarding our article for this special section regarding the validation practices employed with recently translated and adapted versions of the Wechsler Intelligence Scale for Children (WISC), which are used by school and educational psychologists in clinical practice around the world. Whereas we seek to briefly clarify points of minor contention, there is much that we agree on from the commentary. We reiterate the need to take seriously established psychometric standards in our discipline when validating commercial ability measures for the benefit of our ethical charges.

Keywords

evidence-based assessment, psychometric standards, WISC-V

We thank the editor for the opportunity to respond to Kettler's (2020) commentary on our article, "Use of Translated and Adapted Versions of the WISC-V: Caveat Emptor" (McGill et al., 2020), which was published recently in *School Psychology International*. In following Rapoport's (1961) rule of argumentation,[1]

Corresponding author:
Ryan J. McGill, William & Mary School of Education, P. O. Box 8795, Williamsburg, VA 23187, USA.
Email: rmcgill@wm.edu

we begin by acknowledging several points of agreement with the author before briefly addressing a few issues that deserve additional clarification. Nevertheless, we stipulate that it is clear that Kettler (2020) is committed to advancing evidence-based assessment in our discipline; thus, we read his commentary with earnest, and it left us with much food for thought to guide our future work on these matters. We consider any matters of disagreement expressed below should be regarded as minor.[2]

Matters of agreement

Much of Kettler's (2020) commentary was devoted to critiquing what was perceived as an over-emphasis on the structural validity evidence presented for the WISC-V or lack thereof. We agree that our article does devote disproportionate attention to discussion of these matters; however, a major reason for this was that additional explication on these issues was specifically requested by the manuscript's reviewers, given that many translated and adapted versions of the WISC point users to the *Technical and Interpretive Manual* for the United States version (Wechsler, 2014) to understand the structural validity of their local version. Specifically, reviewers noted that the potential shortcomings associated with the methods used to validate that particular measure were important to explore in depth so that readers could fully contextualize our methodological concerns.

Construct validity is multidimensional

Nevertheless, we agree with Kettler (2020) and Messick (1989) that construct validity in applied psychological assessment is inherently a multidimensional concept and does not rest solely on the evidence of structural validity. As per Messick (1980), "...construct validity is indeed the unifying concept of validity that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships" (p. 1015). This is the very reason why we also devoted space in our review to discussing issues pertaining to reliability, treatment validity, and diagnostic utility. Consonant with our commitment to advancing the evidence-based assessment movement in school psychology, we contend that the latter two elements represent what Youngstrom et al. (2015) regard as the clinical bottom line. Nevertheless, structural validity is a critical element of construct validity, and though a researcher may prefer additional forms of evidence (i.e., relationships with external variables) over others, this preference does not obviate the need to explore each aspect of construct validity when attempting to establish the validity of a new test (Cronbach & Meehl, 1955). We agree with Kettler (2020) that some essential evidence for validating a new test can be obtained from existing WISC-V normative samples; in fact, the omission of this information from some versions, despite having adequate target samples from which to obtain it, motivated our review.

Reliability and validity evidence inform clinical test interpretation

Why do we call attention to established psychometric standards (i.e., International Test Commission, 2001, 2017) in psychological science as a guide for our work? Put simply, whereas Lilienfeld et al. (2006) acknowledge that it is *possible* for a test to lack psychometric integrity in the most fundamental aspects of reliability and validity and yet somehow confer useful information to a skilled clinician (e.g., the Rorschach test) in some circumstances, these unicorn examples represent the exception not the norm. Unfortunately, many clinicians overestimate their detective skills in being able to overcome these limitations and are unable to resist the invitation to engage in risky decision making. As noted by Borsboom and Markus (2013),

> In an ideal Cartesian tree of knowledge, only certainly true beliefs would be admitted as premises, and thus conclusions would follow with certainty. Thus, truth and justification hold together. However, even formally valid deductive arguments allow for some slippage: A valid argument can lead to a false conclusion if it begins with false premises. Thus, what is required is not just validity but soundness of the argument (validity plus true premises [pp. 110-111]).

Furthermore, such examples do not obviate the need to adhere to the basic rules of measurement in clinical science (Lilienfeld & Strother, 2020). As a consequence, we implore stakeholders to make decisions on how to use and interpret ability measures, such as the WISC-V, based on the *best* available research evidence and not solely on the basis of preferred methods of test interpretation or what is presented in the latest incentivized workshop within our contemporary zeitgeist (e.g., profiles of strengths and weaknesses, cross-battery assessment; McGill et al., 2018). Validity and reliability evidence are important because they inform what we likely can and cannot infer from an ability measure.

Points of clarification

Kettler (2020) took issue with our discussion of the results furnished from both exploratory and confirmatory factor analytic studies with the WISC-V. To wit, he stated "I may be at odds with the field, I do not agree that EFA is a gold standard. EFA is conducted with no a priori theory taken into consideration" (p. 478). Admittedly, the EFA versus CFA debate is a minor methodological matter, and we respect Dr. Kettler's right to take a different position on these matters. We agree with Gorsuch (2003) that there is no gold standard method of factor analysis. Both approaches are capable of recovering correct models when those models best explain underlying data. However, there is nothing inherently *exploratory* or *confirmatory* about the programs that are commonly used for either method; instead, it is how those programs are used that should define how a factor analytic study is classified (Loehlin & Beaujean, 2017). We disagree with the prevailing myth that EFA is inferior to CFA as they complement each other so that we can have greater

confidence in the veracity of a measurement model when results from both approaches are in agreement. We view EFA as an important check on the use of inappropriate CFA procedures (e.g., specification searches, application of undisclosed constraints) to recover preferred models, even when those models are untenable for the data (Canivez & Youngstrom, 2019; Dombrowski et al., 2020). Further, EFA can, and should, be used to let "the data to speak for themselves" (Carroll, 1995, p. 436) and is instructive for determining plausible (and implausible) models to test in CFA, especially in a new or revised test version.

Other forms of validity evidence

It is clear that Kettler (2020) would have preferred more in-depth discussion about other aspects of reliability and validity (i.e., relationships with external measures, content validity, treatment utility, etc.), seeming to suggest that omission of this information was somehow misleading. First, we note that Kettler (2020) acknowledges that he did not have access to the Technical Manuals in question. However, much of the additional information that is sought appears to be presently unavailable in many of those sources. Thus, we are unable to comment further on many of those important validity elements; in particular, treatment utility, which is not an evidentiary lacuna that is limited to the WISC-V. Our substantive discussion about structural validity issues with the WISC-V does not mean that we do not value other forms of validity evidence for various versions of that instrument. Our focus on the former stems from our agreement with Benson's (1998) approach to construct validation, which positions structural validity as the foundation for progressively exploring other aspects of construct validity. As noted by Keith and Kranzler (1999), external validity evidence is veritably meaningless if the statistical rational for those scores is not firmly established. Again, we return to our central contention. When structural validity evidence is questioned in the literature or, in some cases, not reported at all, school psychologists' confidence that they are interpreting scores that reflect legitimate psychological dimensions is necessarily compromised (Kranzler & Floyd, 2020).

Conclusion

In closing, we reiterate that issues pertaining to clinical test interpretation are important for our field. In particular, during the COVID-19 pandemic, previously unused assessment technologies were introduced, at scale, in our business (Farmer et al., 2020). Given the departures from psychometric best practice that we highlighted in our review, it is fair to ask whether established psychometric standards (i.e., American Educational Research Association et al., 2014; International Test Commission, 2001, 2017) matter in our profession. Nevertheless, we encourage practitioners and trainers to familiarize themselves with relevant test standards and ethical codes that govern test use so that they can become more informed and critical consumers of psychological instrumentation. Familiarity with these

standards and codes may also better enable practitioners and trainers to spot potential gaps in the literature prior to adopting an instrument for use in clinical practice and to identify which scores or score comparisons are psychometrically worthy of clinical inferences.

In closing, we look forward to partnering with those who agree as well as disagree on these matters to improve the educational outcomes of all children in schools. Thus, we were pleased to read and will respectfully end this discourse in agreement with Kettler's (2020) overarching conclusion, which further raises the ethical bar on test use and interpretation famously articulated by Weiner (1989). The onus rests on both the test publisher *and* user; act accordingly.

Acknowledgements

The authors wish to thank Dr. Ryan J. Kettler, Associate Professor and Dean of Academic Affairs at the Graduate School of Applied and Professional Psychology at Rutgers University for his thoughtful commentary on our article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Ryan J. McGill https://orcid.org/0000-0002-5138-0694
Gary L. Canivez https://orcid.org/0000-0002-5347-6534

Notes

1. We invoke the term loosely so as not to suggest that we have any particular point of contention that bears argumentation in a classical sense with the principle involved.
2. We draw attention to this point, as academic discourse is frequently marked by agonism (Tannen, 2002), a ritualized adversativeness elevating the perceived stakes associated with minor disagreements.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement on Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practices*, *17*(1), 10–17. doi:10.1111/j.1745-3992.1998.tb00616.x

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement*, *50*(1), 110–114. doi:10.1111/jedm.12006

Canivez, G. L., & Youngstrom, E. A. (2019). Challenges to the Cattell-Horn-Carroll theory: Empirical, clinical, and policy implications. *Applied Measurement in Education*, *32*(3), 232–248. doi:10.1080/08957347.2019.1619562

Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research*, *30*(3), 429–452. doi:10.1207/s15327906mbr3003_6

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. doi:10.1037/h0040957

Dombrowski, S. C., McGill, R. J., Canivez, G. L., Watkins, M. W., & Beaujean, A. A. (2020). Factor analysis and variance partitioning in intelligence research: Clarifying misconceptions. *Journal of Psychoeducational Assessment*. Advance online publication. doi:10.1177/0734282920961952

Farmer, R. L., McGill, R. J., Dombrowski, S. C., Benson, N. F., Smith-Kellen, S., Lockwood, A. B., Powell, S., Pynn, C., & Stinnett, T. A. (2020). Conducting psychoeducational assessments during the COVID-19 crisis: The danger of good intentions. *Contemporary School Psychology*. Advance online publication. doi:10.1007/s40688-020-00293-x

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. *2*, pp. 143–164). John Wiley.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, *1*(2), 93–114. http://www.intestcom.org

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). www.InTestCom.org

Keith, T. Z., & Kranzler, J. H. (1999). The absence of structural fidelity precludes construct validity: Rejoinder to Naglieri on what the cognitive assessment system does and does not measure. *School Psychology Review*, *28*(2), 303–321. doi:10.1080/02796015.1999.12085967

Kettler, R. J. (2020). Interpretation of the translated WISC-V: Caveat venditor and caveat emptor. *School Psychology International*, *41*(5), 475–482. doi:10.1177/0143034320942281

Kranzler, J. H., & Floyd, R. G. (2020). *Assessing intelligence in children and adolescents: A practical guide for evidence-based assessment* (2nd ed.). Rowman and Littlefield.

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology*, *61*(4), 281–288. doi:10.1037/cap0000236

Lilienfeld, S. O., Wood, J. M., & Garb H. N. (2006). Why questionable psychological tests remain popular. *Scientific Review of Alternative Medicine*, *10*, 6–15.

Loehlin, J. C., & Beaujean, A. A. (2017). *Latent variable models: An introduction to factor, path, and structural equation analysis* (5th ed.). Routledge.

McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of School Psychology*, *71*, 108–121. doi:10.1016/j.jsp.2018.10.007

McGill, R. J., Ward, T. J., & Canivez, G. L. (2020). Use of translated and adapted versions of the WISC-V: Caveat emptor. *School Psychology International*, *41*(3), 276–294. doi:10.1177/0143034320903790

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012–1027. doi:10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Macmillan.

Rapoport, A. (1961). Three modes of conflict. *Management Science*, *7*(3), 195–322. doi:10.1287/mnsc.7.3.210

Tannen, D. (2002). Agonism in academic discourse. *Journal of Pragmatics*, *34*(10–11), 1651–1669. doi:10.1016/S0378-2166(02)00079-6

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children-Fifth Edition: Technical and interpretive manual*. NCS Pearson.

Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. *Journal of Personality Assessment*, *53*(4), 827–831. doi:10.1207/s15327752jpa5304_18

Youngstrom, E. A., Choukas-Bradley, S., Calhoun, C. D., & Jensen-Doss, A. (2015). Clinical guide to the evidence-based assessment approach to diagnosis and treatment. *Cognitive and Behavioral Practice*, *22*(1), 20–35. doi:10.1016/j.cbpra.2013.12.005

Author biographies

Ryan J. McGill, PhD, BCBA-D, NCSP is Associate Professor of school psychology chair of the department of school psychology and counselor education at the William & Mary School of Education. His research focuses on applied psychological measurement and SLD identification in school psychology.

Thomas J. Ward, PhD is Professor of education and chair of the department of educational policy, planning, and leadership at the William & Mary School of Education. His research focuses on research methodology and applied measurement in education.

Gary L. Canivez, PhD is Professor of psychology and principally involved in the Specialist in School Psychology training program. Dr Canivez is a Fellow of the American Psychological Association Division of Quantitative and Qualitative Methods and Division of School Psychology, a Charter Fellow of the Midwestern Psychological Association, and a member of the Society for the Study of School Psychology. He is currently a Senior Editor for School Psychology Review and is an editorial board member for several school psychology and assessment journals. His research interests are in applied psychometrics in evaluating the psychometric fitness of psychological and educational tests (including international applications), and empirically supported test interpretation.