National Consortium for Physical Education for Individuals with Disabilities Position Stand on Assessments in Adapted Physical Education

Section 300.304 of the Individuals with Disabilities Education Act (IDEA, 2004) outlines requirements for conducting an evaluation for the provision of special education and related services. Some of these requirements include:

- scores of psychometrically investigated assessments that are non-discriminatory,
- assessments used for the purpose for which scores are valid and reliable,
- use of multiple assessments,
- administered following all standardized procedures in accordance with the instructions provided by the instrument's producer.

Historically, the field of adapted physical education (APE) has not clearly defined the criteria for determining which assessment scores pass psychometric rigor standards potentially leading to the use and misrepresentation of non-standardized tools as a primary data source for evaluation purposes. The National Consortium for Physical Education for Individuals with Disabilities (NCPEID) addresses this concern through this position statement to better promote that the field identifies and only uses stout and psychometrically vetted scores from assessments in compliance with the law.

NCPEID adopts the definitions and meanings of the following terms as outlined in the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) checklist: interpretability, peer reviewed, reliability, re-standardization, sampling factors, and validity as the minimal criteria required to document an assessment as a standardized instrument for assessment in APE (Hopkins et al., 2009; Lohr, 2002; Mokkink et al., 2010a,b; Terwee et al., 2012; see Figure 1).

Using the COSMIN checklist, APE professionals can avoid litigation and better provide quality services to students with disabilities.

Acknowledgements: Melissa Bittner, Amanda Young, Dale Ulrich, Brad Weiner, Kayla Abrahamson, NCPEID Executive Committee, NCPEID Advocacy Committee

References

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

Hopkins, W. G., Marshall, S. W., Batterham, A. M., & Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. *Medicine and Science in Sports and Exercise., 41*(1), 3-12.

Lohr, K. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research, 11,* 193-205.

Maglione, D., Rodin, G., & Kjer, M. (2019). When should I update to the new revision of a test? Retrieved from https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/When-to-Upgrade.pdf

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. (2010a). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63,* 737-745.

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D., Buter, L. M., & de Vet, H. (2010b). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology 2010, 10*(22) http://www.biomedcentral.com/1471-2288/10/22

Terwee, C., Mokkink, L. B., Knol, D. L., Ostelo, R., Bouter, L. M., & de Vet, H. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research, 21*, 651–657.

NCPEID
THE NATIONAL CONSORTIUM FOR PHYSICAL EDUCATION FOR INDIVIDUALS WITH DISABILITIES

COSMIN Checklist Definitions & Meanings

Term	Definition	Meaning
Interpretability	The degree to which one can assign qualitative meaning to an instrument's quantitative scores or change in scores (Mokkin et al., 2010a).	The ability to analyze data results to form an accurate explanation. Only results with statistical significance should be included in a standardized assessment.
Peer Reviewed	Peer review involves subjecting the assessment and standardization procedures to the scrutiny of other experts in the same field to check its validity, reliability, and evaluate its suitability for publication (Hopkins et al., 2009; Lohr, 2002).	A diverse group of APE professionals providing their feedback following an in-depth review of the tool. These results should then published in a peer reviewed journal.
Reliability	The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions (Mokkin et al., 2010a).	Under the same prescribed conditions, individuals receive the same score for multiple trials of the same skill by same and different test administrators. For example, test-retest (over time), interrater (by different persons on the same occasion), or intrarater (by the same persons on different occasions).
Re- standardization	Test specifications should be amended or revised when new research data, significant changes in the represented domain, or newly recommended conditions of test use may reduce the validity of test score interpretations (American Educational Research Association Standards for Educational and Psychological Testing, 2014; Terwee et al., 2012).	Producers of a tool should review and if necessary, revise the standardizations when: · 20 years have passed · A change in how the skill is performed · A change in performance results of the identified normed population · A skill is no longer believed to be important/necessary All assessors should use the most recent tool after one year of being launched (Maglione et al., 2019).
Sampling Factors	Participants must include a sample that is representative of the entire location it is representing (Hopkins et al., 2009).	The research should consist of a large sample of individuals that include diverse attributes (age range, gender, disability, geographic location, ethnicity, etc.) that the tool is intended.
Validity	The degree to which an instrument measures the construct(s) it purports to measure (Mokkin et al., 2010b).	**Content Validity:** The tool consists of skills that will measure the instrument's intended purpose. **Criterion Validity:** The results correspond to a different instrument of the same thing. **Construct Validity:** The tool measures what it was developed to measure.