NJASP Guidance: In-person Administration, Scoring, and Interpretation of Standardized Assessments During Covid-19

By Terri A. Allen

First Priority: Health and safety of students, families, and school staff

- The health and safety of students, families, and school staff must be the first priority when considering the delivery of services. School districts should adhere to standards released by infection control experts (e.g., NJ Department of Health, CDC guidance). Services (including evaluations and the provision of related services) which require face-to-face interaction should not be conducted during times when there is a substantial risk of contagion, as this practice places the health of the student, family, and practitioner at risk.

- As stated in previous guidance, for many students, sufficient information for eligibility and programming decisions can be gathered using file reviews, interviews, rating scales. However, an IEP team may conclude that the data from standardized testing is necessary so that the evaluation "is sufficiently comprehensive to identify all of the child's special education and related services needs, whether or not commonly linked to the suspected eligibility category."

  The current pandemic underscores the importance of *ongoing* data collection, development and implementation of evidence-based interventions, and progress monitoring, as essential not only when school is in session but also should school be closed suddenly. In doing so, we possess a broader base of data, drawing not only from available "traditional" assessment information, but from multiple sources, providing a framework of evidence that we may use to make decisions with better fidelity and integrity.

- When "stay at home" restrictions are relaxed, school psychologists and other CST members will need **clear and specific** guidelines, consistent with NJDOH and CDC parameters, in order to ensure the safety of school staff and students. School districts are expected to provide clear health and safety standards, including, but not limited to, screening protocols, the wearing of PPE, physical distancing/partitioning procedures and materials, adequate ventilation in testing spaces, and sanitization. Districts should be expected to provide necessary supplies and adjustments to the evaluation setting.

In Person Testing Amidst the COVID-19 Pandemic: Reliability and Validity

- Practitioner decision-making regarding in-person testing with adaptations should be predicated by professional ethical standards. The decision "to test or not to test" should be based on the best interest and individual needs of a student. All decisions must be ***student centered***; not driven by convenience or "routine". The evaluation planning process should focus on defining specific evaluation questions to guide the selection of appropriate evaluation components. Vulnerable students, such as young students, students with emotional conditions, students with language impairments, students with hearing and/or visual impairments, students with developmental disabilities, and others may be greatly impacted by the use of PPE and evaluations results may have limited validity in these cases. For many of these students, alternative "non-testing" evaluative data may be more useful.

- To meet health and safety standards, in person evaluations will require the use of PPE and other adaptations. This is obviously not the way these tests were developed and standardized and potential threats to reliability and validity need to be addressed.

- It is important that the results of an assessment be as precise and consistent as possible. However, no test is 100% perfect and there is always a degree of *error* when we measure something. Although there will be unpredictable fluctuations that are unavoidable, by following standardized procedures we help to eliminate some of that *error.* The less error we have, the more reliable the test is. The more reliable the test, the greater confidence we can have that we are truly measuring what we want to measure (e.g., cognitive abilities, academic achievement). However, deviating from standardized administration procedures introduces a new unknown degree of error and uncertainty into the results.

- Whenever you apply adaptations to a test, you increase the potential of measurement error. School psychologists are obliged to minimize error as much as is possible, in order to obtain reliable and valid results. Further, measurement error related to examiner error and differences may have a greater impact on reliability than what is estimated by test publishers (Lichtenstein, 2020). Therefore, the more conditions depart from standard administration, the greater the chance for measurement error attributable to examiner error or differences. Because a new unknown source of error is introduced to the current testing situation, school psychologists need to be especially diligent in reducing other potential "known" threats to reliability and validity as much as possible. For example, if the test format is changed from traditional paper-pencil to an in-person digital format (e.g., Q-interactive), to minimize the touching of test materials, the school psychologist must be competent and experienced in the administration format in order to reduce *examiner error.* School psychologists should plan ahead if new protocols for testing require additional training and practice. Not only will it be helpful to practice on the new administration platform, but doing a trial run of administering the test under the PPE conditions, could help to minimize error due to examiner administration error or differences (e.g., differences in vocal modulation and clarity when wearing a mask).

- As with all assessments, school psychologists should consider the reliability and validity evidence relevant to their assessment techniques and articulate the limits of their assessment results. Modifications to standardized administration procedures need to be explained to parents and school teams so they are adequately informed of concerns related to reliability and validity and have the opportunity to decline affected portions of the evaluation **before** starting the evaluation procedures.

  {An example of language you can use for informed consent is provided in the appendix of this document.}

  In explaining to parents or school staff, try to communicate in layman's terms: Reliability is "How well are we measuring something?" [consistency and precision]; Validity is "Are we measuring what we want to measure?" [meaning and relevance]. You cannot have validity, i.e., test measures what it is supposed to, if the test does not measure "it"in a consistent, i.e, reliable, manner. Therefore, as our concern about potential error in measuring "X" increases, our confidence in what "X" means will decrease.

Scoring, Interpretation, and Reporting

- A detailed description of all modifications/adaptations to standardized administration procedures should be included in written reports, as well as their potential impact on the reliability and validity of the assessment. Provide specific examples. For example, you observed that the student frequently asked you to repeat questions. Did he have difficulty understanding you because of the mask or did you observe other instances of inattention that might better explain the behavior? What other data can you examine to help determine the impact of the mask on test results? Did last year's teacher report similar behavior when masks were not necessary? Did the testing take longer than usual because of necessary safety measures? Do you think that impacted your test scores? Did there seem to be a concern related to physical distancing? For example, the student is used to "proximity", "pointing/gesturing as cues" but this was not possible due to either the distance or plexiglass divider.

  Address the potential impact of the adaptations but do not assume that the adaptations are the *sole* reason for the difficulties. Examine your results: Is there a pattern observed? The student performs significantly better on nonverbal items that require little, if any, verbal directions. Could there be auditory or language processing concerns that while exacerbated by the current conditions, may not be only because of the need for masks (based on the totality of your information)? These observations could be critical in not only in determining the reliability and validity of your results but also to inform day to day support and intervention that the student may require.

- The potential impact of increased measurement error related to deviations from standard administration should be considered when reporting actual test scores. Because there are

insufficient data to suggest any systematic modifications of norms used to interpret tests administered in this manner, you will need to rely on normative and validity data obtained when the tests were developed. In other words, you will score the test in the same way that you score the test when it is typically administered but with clear documentation in the report in any score tables that you may include. However, because you have introduced unknown *error* that potentially affects the reliability of the scores, the interpretation and reporting of those norm-referenced scores should be done cautiously.

Whereas, reliability refers to measurement consistency, measurement *error* is represented by the confidence interval (as based on the standard error of measurement SEM). This margin of error represents the range of "true" scores around the obtained score. So, for example, in reporting a full scale IQ score of 88, we might say that we can say with 95% confidence that the student's "true" score falls in the range of 83 - 94. These confidence levels ranges were determined when the test was originally developed. In deviating from standardized procedures, we are introducing unknown and unpredictable *measurement error.* We do not have any data to discern how much this error impacted the test's reliability, and, hence the validity of the results. Even without a deviation from standardization, norm-referenced scores from standardized tests are never perfectly accurate and reporting a band of uncertainty around the scores by using confidence intervals helps to account for the potential error. Due to the potential of a greater margin of error, as represented by the confidence interval band, the obtained confidence intervals per the test manual, must also be interpreted with caution. Additionally, the potential effect of adaptations may vary between different subtests depending on task demands. Intuitively, one would think that wearing a mask might be a bigger issue when administering one subtest more than another. But without psychometric evidence - that does not exist - one cannot assume greater error because that "makes sense". Drawing conclusions based on a single subtest should always be avoided and under the current conditions school psychologists need to be even more judicious in interpreting test scores, emphasizing the more reliable global and/or composite scores. Especially during atypical circumstances, school psychologists should remember that "just because the test or its scoring software produces a score, you need not interpret it" (Kranzler & Floyd, 2013, p. 95).

Therefore, although reporting confidence intervals at the widest range (95%) is better than reporting a single obtained score, we can not assume that under these conditions the student's "true" score falls between 83 and 94, for example. Reporting of percentiles ranks is similarly problematic and if included, should be not reported as an *absolute* but rather an approximation. During the current circumstances, school psychologists are advised to refer to the descriptor label in reporting results (best); confidence intervals at 95% with a disclaimer (better than single obtained scores); and avoid reporting single obtained scores. If you *have to* report a "number", clearly and unequivocally state an appropriate disclaimer. "Make generous allowance for measurement error" (Lichtenstein, 2020) is a particularly relevant statement during these times. Your report should not overemphasize numbers (now or ever). This may be a good time to reassess your report writing style - switching to a domain focused and/or referral question guided format from a traditional "test by test" template. **Remember, it is**

never about the "test", but always about the child.

- It cannot be stated too many times: Test score data should be compared to alternative sources (i.e., previous test scores, existing school records/curriculum assessments, teacher interview) in order to confirm or refute obtained test scores. What patterns do you have across all the evaluative data? Where does the data converge? Diverge? A well established and robust MTSS framework can provide a wealth of valuable alternative assessment data.

- In addition to descriptions of modifications/adaptations verbiage, language should be added to your report regarding potential social and emotional factors that could impact the reliability and validity of your assessment results. The current level of distress among students and their families could impact test performance and must be considered when interpreting results. Some students have experienced varying degrees of trauma during this pandemic. Given that research suggests that stress can lead to reduced performance on measures such as working memory and processing speed, knowledge of recent experiences will serve as a critical context for interpretation.

  The extended time out of school may impact results of standardized cognitive and achievement assessments. What was the quantity and quality of the student's remote instruction? How might that factor into interpretation? What was the students "Rate of Improvement" ROI/growth slope prior to the closure? Was he/she struggling even with targeted intervention when school was in session and during remote learning? Was the student making progress prior to the closure but appears to be further behind since the closure (in which case, you probably should be providing interventions rather than jumping to an evaluation)?

  Consider the student's circumstances related to the pandemic when interpreting behavior rating scales. In interpreting the results of the scales, do the results seem to reflect a student's "trait" or his/her current "state"? A response to behavior will differ depending if the presentation seems more related to the current "state" of affairs as opposed to a long standing pattern. Look at behavioral data collected prior to the closure. Assess for potential trauma that the student may have experienced related to the pandemic. Multiple sources of data is especially critical for the evaluation process as we resume in-person services.

- Examine, reflect, evaluate, integrate but still take a stand (albeit, a cautious stand) with regard to the reliability and validity of your assessment. Don't assume that your evaluation cannot be reliable and valid because of all the "cautions" and "disclaimers" - if that were the case, why even bother doing the evaluation in the first place? Rather, use your data-based decision making skills, plus your experience and clinical judgement (yes, I said clinical judgement) in determining what you CAN say about the student with reasonable confidence and certainty. And even more important, indicate how your evaluation can help drive support for this student, whether he or she is "eligible" for special education and related services or not.

Three Take-Aways:

- All decisions must be ***student centered***; not driven by convenience or "tradition". Plan your evaluations around the questions that need to be answered in order to help the child; not around the assessment tools that are available.
- Cast your evaluation net wider and deeper. Draw, not only from "traditional" assessments, but from multiple sources, providing a comprehensive framework, which will provide a comprehensive framework of relevant and meaningful evidence that can be used to make decisions with fidelity and integrity.
- In order for children to thrive, they must be healthy, safe, and have a sense of belonging. Without this foundation, learning will be difficult, if not impossible. Perhaps, these unprecedented circumstances have provided an opportunity to reconceptualize our approach to assessments and evaluation. We have an opportunity to move from the traditional focus on relatively narrow bands of cognition, achievement, and classroom behavior, to a more integrated, thorough, student (not test) focused lens as we seek to observe, understand, and help our children and youth.

Contact Terri @njasp.web@gmail.com with comments/questions

Resources:

APA Guidance on psychological tele-assessment during the COVID-19 crisis

Considerations for Delivery of School Psychological Telehealth Services

Kranzler, John H. & Floyd, Randy G. (2013). *Assessing Intelligence in Children and Adolescents: A Practical Guide.* New York: Guilford Press.

Lichtenstein, Robert (2020). What You Don't Know About Measurement Error - and Why You Should Care. Communique, volume 48, issue 8 https://www.nasponline.org/resources-and-publications/periodicals/communiqu%C3%A9-volume-48-number-8-(june-2020)/what-you-dont-know-about-measurement-error%E2%80%94and-why-you-should-care

NASP Position Paper: School Psychologists' Involvement in Assessment

Virtual Service Delivery in Response to COVID-19 Disruptions

Salkind, Neil J. (2006). *Tests and Measurement for People Who Think They Hate Tests & Measurement* Thousand Oaks, CA: Sage Publications.

Telehealth: Virtual Service Delivery Updated Recommendations

Virtual Service Delivery in Response to COVID-19 Disruptions

Appendix A:

Click here for Word version

Example: Informed Consent/Communication with parents, IEP team, teachers, and administrators - modify as appropriate for your situation

Standard test administration will be modified, and this may affect results in ways that are so far unknown. This has the potential to reduce confidence in the obtained test score(s), conclusions and recommendations. Error may be compounded when adaptations to standardization procedures are used with people who come from culturally and linguistically diverse populations, require an interpreter during the assessment, have attention or auditory processing difficulties or have limited experience/comfort with the PPE adaptations. There may be a loss of some qualitative data usually obtained during the standard in-person test administration related to the use of face coverings and physical distancing and this loss may reduce the richness of the clinical data and further limit conclusions.

No test is 100% perfect and there is always a degree of error when we measure something. The less error we have, the more reliable the test is. The more reliable the test, the greater confidence we can have that we are truly measuring what we want to measure (eg., cognitive abilities, academic achievement). Whenever you apply adaptations to a test, you increase the potential of "error." However, every effort will be made to mitigate the potential impact of concerns related to the necessity of administering the assessment in a manner that deviates from standardization. In addition to standardized test scores, alternative data will be gathered from multiple sources in order to provide a more comprehensive picture of your child's assets and challenges.

If at any point during the evaluation, it appears that the assessment process is not safe for the child or the examiner or circumstances indicate that obtained results are likely to be unreliable, the assessment will be discontinued. A decision will be made with the parent/guardian as to whether rescheduling the assessment is advised. Following the completion of the evaluation, parent/guardian will be provided with a report that will include clear statements about the limitations posed by non-standard administration and the potential impact this might have on test scores, conclusions and recommendations.