SREE 2019

Tensions and Tradeoffs: Responding to Diverse Demands for Evidence

Moderated panel submission

Contact (session moderator) email: agandhi@air.org

Title: Supporting the use of evidence-based screening and progress monitoring tools: Reflections from over a decade of building and maintaining the National Center on Intensive Intervention (NCII) Assessment Tools Charts

First choice conference section: Academic Learning in Education Settings

Second choice conference section: Research Methods

Panel justification: Since 2004, the U.S. Office of Special Education Programs (OSEP) has funded the development and maintenance of a series of "tools charts" which publish results of technical reviews for commonly used academic and behavioral screening and progress monitoring assessments. Now, more than a decade later, these charts, which include information on over 60 products, are available via the National Center on Intensive Intervention (NCII, www.intensiveintervention.org). As demand for information on evidence-based practices has grown, states and districts have increasingly relied on the NCII tools charts as a trusted guide when selecting assessments to use within tiered systems of support such as Response to Intervention (RTI) or Positive Behavioral Interventions and Supports (PBIS). In fact, some states (e.g., Iowa, Washington) require schools to adopt screening or progress monitoring systems that have demonstrated sufficient technical adequacy in accordance with the NCII chart standards. In response to such requirements, commercial publishers of formative assessments are increasing the rigor of their research and development procedures to ensure stronger ratings on the charts. What began as an opportunity for a small group of researchers to engage in informal peer review for academic progress monitoring tools, has evolved into a formal and rigorous review system that has high stakes for tool publishers and tool users.

In this moderated panel, NCII staff and Senior Advisors will reflect on the tensions and tradeoffs involved in the development and ongoing maintenance of the tools charts. While the goal has always been to promote the use of assessments that are technically strong, tensions exist related to: (1) limited technical knowledge among users of screening and progress monitoring tools; (2) new and evolving research about psychometric methods to support the reliability and validity of screening and progress monitoring tools; and (3) motivations for the commercial publishing industry to participate in the reviews. Navigating these tensions has required that NCII make tradeoffs at different times; for example, by allowing time for the publishing field to "catch up" before introducing new evidence standards, by publishing information on tools with limited evidence, or by removing evidence standards to improve the chart's interpretability for users.

The panel will comprise an overview presentation, followed by a moderated discussion among two panelists and the audience. First, Dr. Jill Pentimonti of the American Institutes for Research (AIR) will explain the four assessment tools charts: academic screening, behavior screening, academic progress monitoring, and behavior progress monitoring. She will describe the evidence standards for each chart,

the review process, and the format and features of the charts themselves. Next, Dr. Allison Gandhi (AIR) will moderate a discussion in which she will pose three questions related to tensions and tradeoffs to two panelists who advised NCII on the development of the academic and behavior charts, respectively: Dr. Lynn Fuchs (Vanderbilt University) and Dr. Chris Riley-Tillman (University of Missouri). There will be time for audience participation during discussion of each question.

Abstracts for panel elements:

1. **Overview of NCII Screening and Progress Monitoring Tools Charts**

Presenter: Jill Pentimonti, American Institutes for Research

The National Center on Intensive Intervention (NCII) is the third in a series of technical assistance Centers, funded by OSEP in the U.S. Department of Education and operated by AIR, to support implementation of various aspects of a multi-tiered system of support (MTSS). Because a well-implemented MTSS requires universal screening at least yearly and regular progress monitoring of students who have been identified as at-risk, it is important for practitioners to use screening and progress monitoring tools that have evidence demonstrating reliability and validity. Beginning in 2004 under the National Center on Student Progress Monitoring (NCSPM), AIR developed the academic progress monitoring tools chart, which published results of technical reviews for commonly used progress monitoring tools. In 2007, OSEP discontinued funding for NCSPM and instead funded the National Center on Response to Intervention (NCRTI). Under NCRTI, AIR continued the progress monitoring tools chart and added a chart for screening tools. In 2011, OSEP funded AIR to operate the National Center on Intensive Intervention (NCII), and again ownership of the charts was transferred to the new Center. NCII continues to operate the academic screening and progress monitoring charts and has added charts for behavioral screening and behavioral progress monitoring. As of October 2018, the charts include information on 22 academic screening tools, 33 academic progress monitoring tools, 1 behavior screening tool, and 8 behavior progress monitoring tools.

Review process

Reviews for the chart are conducted on an annual basis, with tool publishers responding to an NCII "call for submissions." Publishers who choose to submit tools for review must fill out a detailed evidence protocol with the required technical information. Each chart's Technical Review Committee (TRC), comprising external experts, establishes evidence standards and conducts reviews based on those standards. During each review cycle, NCII randomly assigns two TRC members to each submission, who review and rate their assigned tools independently and then meet to reconcile ratings as needed. NCII communicates these ratings, termed "interim results," to the publisher, who has the opportunity to submit a response with additional evidence. The TRC members review additional evidence and make their final ratings, which are publicly posted. As a final step, the full TRC conducts a debrief meeting, in which the group reviews submissions and ratings and discusses issues that arose in that review cycle pertaining to inconsistent ratings or unclear evidence and revises evidence requirements as needed for future review cycles. Each review cycle lasts approximately 6 months.

The evidence standards

Current evidence standards reflect the latest knowledge in the field, with an emphasis on relevance for users. To the degree possible, TRCs align language and evidence standards across charts, such that

differences across charts reflect differences in the tools' purposes. For example, screening tools must demonstrate accuracy in classifying students risk status for future academic difficulty or behavior challenges, whereas progress monitoring tools must demonstrate sensitivity to small, frequent changes in student learning or behavior. For all charts, the standards require evidence of technical adequacy for NCII's population of interest: students with intensive needs. However, because schools use screening and progress monitoring tools within their broader MTSS, charts also report information on technical adequacy for the general student population.

Tools are rated for each standard using a "bubble" system: a full bubble indicates that the tool meets the evidence standard, a half bubble that the tool partially meets the standard, and an empty bubble that the tool does not meet the standard. Users may click on any bubble rating and be directed to a pop-up window providing all technical detail that the TRC reviewed and rated. This ensures the review process's transparency and allows users with a more sophisticated technical understanding to dig deeper into the data to assess a tool's fit for their school or district.

For ***screening***, NCII rates the following technical properties: (1) classification accuracy; (2) reliability; (3) validity; (4) sample representativeness; and (5) bias analysis. These categories are the same for academics and behavior, yet the detail for each standard differs somewhat in terms of what is expected.

For ***progress monitoring***, NCII rates technical properties that fall into two categories: *foundational psychometrics*, which apply to the use of the tool with the general population of students; and *growth standards*, which apply to the use of the tool with students in need of intensive intervention. For foundational psychometrics, the TRC rates tools on reliability, validity, and bias analysis. For growth standards, the academic and behavior TRCs use slightly different rubrics. The academic TRC provides ratings on sensitivity to student learning (reliability of the slope); sensitivity to student learning (validity of the slope); alternate forms; decision rules for setting and revising goals; and decision rules for changing instruction. The behavior TRC rates tools on sensitivity to behavior change; reliability (for the intensive population); validity (for the intensive population); data to support intervention change; and data to support intervention choice.

In addition to rating bubbles and pop-up windows with detailed data, screening and progress monitoring charts include a section describing the assessments' usability features, such as administration and scoring time and format.

2. Panel Discussion

Moderator: Allison Gandhi, American Institutes for Research

Panelists: Lynn Fuchs, Vanderbilt University; Chris Riley-Tillman, University of Missouri

The tools chart evidence standards, and the accompanying rating rubrics used by TRC members, have evolved over time in response to sometimes competing demands from chart users, tool publishers, and researchers. For example, as researchers have developed better methods for demonstrating technical adequacy, publishers struggle to keep up with the pace of the field and to find the resources to conduct additional studies using these methods. Meanwhile, users struggle to understand what the standards mean and how they are relevant to them. To meet OSEP's ultimate goal of ensuring that schools are

using evidence-based screening and progress monitoring tools, the TRCs must strike the right balance between technical rigor, relevance for users, and feasibility for publishers.

The second portion of the panel will be a discussion featuring two Senior Advisors to NCII who have been closely involved with the establishment and refinement of evidence standards for the tools chart. Dr. Lynn Fuchs of Vanderbilt University is an original member of the academic progress monitoring TRC and a leading researcher in the field of curriculum-based measurement. Dr. Chris Riley-Tillman of the University of Missouri led the original behavior progress monitoring TRC by adapting the academic review process for behavior; he develops and studies methods for direct measurement of student behavior. Dr. Gandhi, the panel moderator, will ask panelists to provide examples of ways in which the NCII tools charts demonstrate tensions inherent in responding to demands for evidence and how the different TRCs have addressed these tensions. Dr. Gandhi will pose three questions related to three tensions and provide time for panelist response and audience discussion.

Tension #1- Creating a demand for evidence

Question: *To be successful in supporting the use of evidence-based screening and progress monitoring tools, demand must come from the field. In what ways have the tools chart created or increased that demand for technically strong tools?*

When the first TRC was established for academic progress monitoring, the notion of progress monitoring was relatively new. There were few commercially available tools; most available products were being developed and tested by researchers. The push for technically strong tools was coming from the developers themselves. The TRC and tools chart filled a need for these researchers to review each other's work and push each other to develop better tools that could improve practice in the field. There was little demand from practitioners for progress monitoring tools, let alone ones that met strong technical criteria. Over time, the amount of content on the chart increased and was disseminated to users. Commercial publishers also began to develop and market progress monitoring tools to school districts. Users began to recognize the chart as a reliable source of information for selecting high-quality products and began to create a stronger demand for evidence-based tools.

Drs. Fuchs and Riley-Tillman will share their perspectives on this organic process of creating a demand for evidence. They will highlight the fact that the academic field is far more advanced than the behavior field in terms of available tools, and the implications this has for the demand for evidence. For example, on the behavior side there are fewer tools to choose from, so users don't have the luxury of demanding strong evidence. The behavior tools charts were launched only within the past 5 years and are still in the process of building that demand.

Tension #2- Staying aligned with evolving standards of evidence

Question: *Evidence standards for screening and progress monitoring tools have rapidly changed since the charts began. It is difficult for publishers and users to keep up with these changes. How have the TRCs handled this?*

During the past 15 years, standards of evidence for high-quality screening and progress monitoring tools have evolved. For example, traditional methods to assess reliability such as split-half or test-retest are being replaced with model-based indices of item quality such as Item Response Theory estimates (Samejima, 1994). When assessing classification accuracy, the field has moved towards the use of the

Area Under the Curve (AUC) statistic over more traditional statistics such as Kappa (Swets, Dawes, & Monahan, 2000). Over the years, the TRC has taken these advancements into account to adjust its evidence standards. However, these adjustments require publishers to conduct and submit new analyses on a frequent basis to avoid having their ratings downgraded. It also requires ongoing and clear communication to chart users so they understand the need for new standards and the implications for the tools they use. Drs. Fuchs and Riley-Tillman will share their perspectives on the challenges that these tensions have created for the TRCs and describe how they navigated these challenges.

Tension #3- Working with commercial publishers

Question: *Many of the submissions come from large commercial publishing firms, who have a strong financial interest in the outcome of the reviews. They invest considerable resources in conducting the analyses necessary to receive favorable ratings. What kinds of challenges has this introduced into the review process?*

The entities that publish and submit their products for review have ranged from individual researchers to large publishing companies. On the academic side in particular, the development of screening and progress monitoring assessments has become big business, and the tools chart ratings carry high stakes. As evidence requirements have evolved, tool developers have had to invest more resources to maintain positive ratings. For smaller firms or individual researchers who don't have access to the same resources as large publishing companies do, this has been especially challenging. Drs. Fuchs and Riley-Tillman will discuss how this challenge has impacted the types of submissions received over time. They will also discuss issues around which publishing companies have strongly disagreed and pushed back on TRC requirements, and how the TRC has responded.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. Applied Psychological Measurement, 18(3), 229-244.

Swets, J.A., Dawes, R. M., Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1-26

Panelists:

Allison Gruner Gandhi
American Institutes for Research
agandhi@air.org

Jill Pentimonti
American Institutes for Research
jpentimonti@air.org

Lynn Fuchs
Vanderbilt University
lynn.fuchs@vanderbilt.edu

T. Christopher Riley-Tillman
University of Missouri
rileytillmant@missouri.edu