knowledge alliance

May 31, 2022

Mark Schneider
Director of the Institute of Education Sciences (IES)
550 12th Street SW
Washington, DC 20024

Dear Director Schneider,

I am writing on behalf of Knowledge Alliance (KA) with comments and recommendations in response to the National Center for Education Research (NCER) request for information (RFI) on identifying existing large datasets that may be useful for research and regarding the challenges and limitations that may affect access and their value for research. Knowledge Alliance, a non-profit, non-partisan organization, is comprised of leading education organizations committed since 1971 to the greater use of high-quality and relevant data, research, evaluation, and innovation in education policy and practice at all levels. We believe the effective use of rigorous research and evidence is integral to innovating and improving learning and outcomes for every student. Collectively, we promote the use of rigorous research to figure out "what works," and why, to improve student outcomes and then share those findings with policymakers, practitioners, and the general public.

KA members engage with IES in a multitude of ways: as grant recipients, data users, data disseminators, REL operators and What Works Clearinghouse (WWC) reviewers and users, and as contractors. KA members also actively engage with the other primary research work happening at the Department of Education (ED) such as through work undertaken through the Education Innovation and Research (EIR) program and other research and technical assistance work such as the Comprehensive Centers (CCs) program within the Office of Elementary and Secondary Education. Therefore, KA is uniquely situated to take a holistic view of the Institute of Education Sciences (IES) and provide insights across the many activities impacting IES' future goals and priorities. That is why KA's response to this RFI extends beyond the scope of the nine questions in the RFI and incorporates KA's comments on the three-IES commissioned National Academies of Sciences, Engineering and Medicine (NASEM) reports and the Sources Sought Notice (SSN) on the Procurement of Research Evaluation and Statistics Task Orders (PRESTO) released in March 2022, as well as KA member work with EIR and CCs. KA also encourages IES to think holistically about how these different initiatives (this RFI, the NASEM reports, and PRESTO) fit together to impact IES work.

Responses to Questions

Question 1: What public or restricted use education-related datasets are available for training students in data mining/machine learning methods? What training needs are not being met by the datasets that are currently available?

Upgrade the IES Conceptualization of an Education Dataset to Harness the Potential of Data Systems
As Education Analytics, a KA member organization, noted in its response, within a single school system "data are being collected in real time across hundreds of live data systems as a result of thousands of vendors providing tools used by students, teachers, and administrators." Furthermore, "each local

education agency (LEA) is capturing transactional data from all of the LEA's operational systems, including the student information system (SIS) that provides classroom assignment and student demographic data, a Human Resource Information System (HRIS) or Enterprise Resource Planning (ERP) system that provide teacher demographics and salary information, a transportation system, learning management system (LMS), a nutrition system, a Professional Development or Learning system that provides documentation about available and provided professional supports, potentially a learning object repository that identifies what curricular materials are available and in use, an assessment data system, etc."

Therefore, KA encourages IES to upgrade the concept of "*datasets for research*" to reflect the potential of harnessing all of the above data in existing operating systems for research purposes. As Education Analytics noted in its response to this RFI, "**we currently live in a data rich and information poor world**." Knowledge Alliance supports Education Analytics' message that "harnessing the power and promise of technology to unlock operational data for research purposes is the future of educational data." Within existing operating systems exist huge swaths of data that can be useful for a myriad of education research purposes.

As Education Analytics notes in its RFI, KA supports the notion "that the modern version of a dataset is instead a data stream." We therefore encourage IES to ask a different question than the one posed here. Instead of asking, "What public or restricted use education-related datasets are available for training students?" KA encourages IES to ask how IES can support efforts to turn data-rich operating systems into data streams (datasets) for a myriad of education research purposes (including training students in data mining/machine learning methods), as well as training researchers on how to adhere to data privacy standards.

A Holistic View: Large Datasets, the NASEM Report Recommendations, and PRESTO

KA believes that this RFI and the three recently commissioned (IESNASEM reports[1] intersect in important ways, and we encourage IES to consider them in tandem. As noted in KA's response memo on the NASEM reports, KA members engage with IES in a multitude of ways. That is why KA's response memo to the NASEM reports recommended that all the IES centers work together, rather than in silos, to create a strategic plan that cuts across IES.

The NASEM report "*A Vision and Roadmap for Education Statistics*" fundamentally reimagined NCES as a leader in education statistics, evidence building, and data governance. The report proposed expanding NCES' role as a data-access facilitator to be deeply engaged with stakeholders, strengthening data capacity at state and local educational agencies (SEAs and LEAs), and acting as a strong partner to ED's evidence-building and research. KA sees the adoption of data streams as an organizing concept and believes their development is an example of the type of high-value function called for in the report. In that regard, KA encourages IES to support SEA and LEA capacity to link data systems for research purposes. Furthermore, KA strongly supports the report's recommendation that NCES assist SEAs and LEAs through data facilitation.

[1] "*A Pragmatic Future for NAEP: Containing Costs and Updating Technologies*," "*The Future of Education Research at IES*" and "*A Vision and Roadmap for Education Statistics*."

KA sees future IES investments in data streams as having benefits in several key areas for several reasons:

- **Data Governance** - NCES has expertise in data governance that it can and should leverage to help States build capacity to streamline data linkage, prepare and curate data, develop templates, and simplify processes.
- **Reduce Reporting and Response Burdens** - Data streams have the potential to reduce the burden on States for reporting data to IES and also reduce the burden on respondents to IES surveys. For example, for universe data collections such as the Common Core of Data (CCD), if IES developed standardized templates for different levels in a data system, NCES data collection could be as simple as sharing extraction codes that States could use on their systems. For IES sample survey collections, many background and demographic variables could be extracted from data systems, reducing the burden on respondents and (potentially) improving internal data validity.
- **Expedite Federal Reporting –** Robust data streams have the potential to reduce the time from collection to reporting, improving on IES's efforts to release more timely products and information for decision-making purposes.
- **Increase Knowledge Use -** Robust data streams can improve how and when practitioners use research evidence to make decisions, and how existing education research can be made more relevant and useful to practitioners in SEAs, LEAs and individual schools. Projects to build out robust data streams are already being undertaken by KA member organizations.[2]

KA also sees a relationship between this RFI, the NASEM report recommendations, and the Sources Sought Notice (SSN) on the PRESTO. In the SSN, it was noted that IES intends to issue a contract vehicle that supports the entire IES mission and will consist of five distinct scope categories. These categories were:

- Category 1: Education Science Support Activities: *IES requires expert support to conduct surveys and evaluations that collect data and produce products to support the mission of IES.*
- Category 2: Assessment Development Support: *IES requires expert support to develop assessment components for its sample surveys studies and develop and conduct large-scale assessments*

[2] For example, using $2 million made possible through the state's Elementary and Secondary Schools Emergency Relief Fund, South Carolina's Department of Education (SCDE) launched a first-of-its-kind statewide "Teacher as Researcher" initiative. This initiative by Marzano Research and Education Analytics will lead to teachers who are skilled in testing and evaluating instructional strategies as well as the development of an interactive web application that will help teachers select, implement, and examine the effectiveness of evidence-based instructional strategies. The web application will include a database of evidence-based strategies from the WWC practice guides. Marzano Research is in the process of helping SCDE craft an evaluation plan to measure changes in teacher knowledge, skills, and efficacy in generating and using evidence to improve their instructional practice.

- Category 3: Communications and Outreach: *IES requires expert support for communications activities to include strategic communications oversight, external communications support, and internal communications support.*
- Category 4: Recruitment: *IES requires expert support to recruit educational entities and groups traditionally underrepresented in education research on a national and international scale.*
- Category 5: IES Operational Support: *IES requires expert support with project management, logistics, and subject matter expertise to enable ongoing functions across the IES.*

KA would encourage IES to consider how the PRESTO contract vehicle can support the aforementioned suggestions for future IES investments in data streams.

<u>The Training Needs of Future Education Researchers</u>
A number of states are leading the way in harnessing the potential of data streams for research purposes. As states increase their capabilities to harness data streams for education research purposes (among other purposes), education researchers will need to diversify the "tools in their toolbox" by engaging, for example, in:

- Training in Application Program Interface (API) technologies to tap into the "data stream."
- Training in traditional SQL servers, open-source modularizable software, etc. to extract data from various operating systems.
- Training in open-source statistical software like R and Python that interact natively with the aforementioned technologies through API calls.
- Training in data privacy.

The above methodologies, used on data streams rather than datasets, open up a universe of research questions that are not currently answerable in the static large-scale datasets typically available in education.

Question 2: What research needs are not being met by the datasets that are currently available?
Question 6: How likely is it that existing datasets, especially those that come out of education technology, contain data that are valuable for researchers and of sufficient quality that research could be conducted with a high amount of rigor?
Question 8: What are the best practices for creating new datasets or linking existing datasets and sharing them with researchers (open or restricted use) while prioritizing the privacy of individuals and adhering to local, State, and Federal laws? What barriers and limitations exist?

KA sees questions 2, 6, and 8 as interrelated is responding to them collectively rather than individually. Among the "datasets" that are currently "available" but not meeting researchers needs are the Statewide Longitudinal Data Systems (SLDS). Thus, we take question 2 and modify it to ask, "why and how do the SLDS fail to meet researchers needs?" We also ask how SLDS differ from the concept of a data stream. KA sees Question 6 as pertaining to question 2 and therefore discusses challenges linking SLDS data that affect the ability of researchers to conduct impact analyses. Finally, for question 8, KA shares feedback on some of the barriers to accessing and using SLDS data.

The Potential and Limitations of SLDS

The SLDS grants program has propelled the design, development, implementation, and expansion of K12 and P20W longitudinal data systems. These systems are intended to enhance the ability of states to efficiently and accurately manage, analyze, and use education data, including individual student records. The SLDS are intended to help states, districts, schools, educators, and other stakeholders make data-informed decisions to improve student learning and outcomes, as well as facilitate research to increase student achievement and close achievement gaps.

Thus, it seems that SLDS are innately related to this RFI on large datasets, as SLDS have the potential to be the largest, most robust source of data for each state. The question then becomes, are SLDS serving education research purposes? As users of SLDS, KA members posit the following responses as to the potential and limits of SDLS "datasets" that are currently available but are not meeting research needs.

The short answer is that SLDS are a very valuable resource, but difficult and costly to access given States' varying rules. The barriers to access likely mean that only academics and research firms have the resources to use them. Further, in-state academics working closely with SEAs have an easier time using their own State SLDS than do non-state academics.

The longer answer is that there are several challenges to the existing SLDS. These challenges include:

- **Data Access and Merging** Some states (Texas and Tennessee) will not allow student data to leave SLDS servers. While the data can now be accessed remotely, researchers are not permitted to combine TX or TN data with data from other states in a single impact analysis. For evaluations of multi-site interventions, this means that researchers can only combine results across sites via meta-analysis. This will work for particular studies but not other kinds of multi-state designs. Other states (such as Virginia and Louisiana) will only allow their data to be used in secure data rooms on dedicated computers not connected to the internet – this has been a challenge during COVID, when most researchers are working from home.

- **Missing Information** SLDS is often missing critical data and context necessary to conduct research. For example, to run impact analyses using SLDS, researchers need data about the intervention merged with outcome data. At a minimum, researchers need a treatment flag added to the data received from the state, but often researchers want to include teacher-level variables related to their participation in the intervention (or business-as-usual conditions). In this example, because the intervention data does not have teacher-level data, SLDS staff have to do a data merge using a state-assigned teacher ID, which researchers can sometimes collect from districts but usually not. Without the state teacher ID, SLDS staff use teacher email, which is a time-consuming and error-prone matching process. If randomization happens at the student level, researchers need a treatment flag linked to the same student ID that the state uses in its SLDS. This ID has to be collected from districts so that researchers can link it to a treatment flag; accessing the ID can raise all kinds of IRB and privacy concerns.

- **Expense** In our experience, states charge quite a bit for access to SLDS data – this covers their costs of extracting the variables researchers have requested and merging in the researchers' intervention variables. KA member organizations report that a one-year license to access TX data costs $18K and that KY is charging $5K per data pull. For a multi-state study, those costs add up.

- **Access** The applications to access SLDS data are time-consuming to complete and take months to approve. Approval by the SEA is usually required and is not guaranteed, which makes planning difficult. Sometimes studies can get caught up in political crosscurrents. It is often quite difficult to tell from State web sites whether they make data available to external researchers and what the process is. The bottom line is that data are not easy to access, and access is not guaranteed. This means, probably, that data access, rather than more important concerns (the context or the research questions) drives design decisions.

- **Data Context** The SDLS generate the topmost aggregated data. This also means that the data in SLDS can lack context. Context gives more value to data. Context also is important for ensuring equity in that the more context data have, the less likely the users of those data will be prone to biased assumptions and inequitable decisions. Data streams, however, enable context to travel with data (rather than be lost in transportation).

- **Data Uniformity** The SLDS do not require uniform formatting and structure across all states. Instead, each state has adopted its own SLDS strategy and format. This makes it very hard (and expensive) to use SLDS data from a national perspective.

To be fair, when SLDS was conceived a lot of current technologies (such as those that are part of the modern data stack) did not exist, so it was not possible to envision the current potential to harness the systems under the concept of data streams. To its credit, the SLDS grant program has laid the groundwork for the vision of data streams by incentivizing communities of SEA analytics teams to collaborate on strategies, incentivizing the development of Federal standards like CEDS, and, in a roundabout way, incentivizing the need for Ed-Fi. To its detriment, the investment in SLDS has created a myriad of 50 different SLDS that operate under different rules, costs, and structures. If IES invests in SLDS without considering the full, underlying ecosystems of data-rich operating systems that trap data, we are going to continue living in a data-rich and information poor society. KA therefore strongly encourages IES to think about how it can support the underlying ecosystems for research systems or modify the SLDS program to address the program’s current shortcomings.

7. To what extent do existing datasets capture enough information to address research questions related to diversity, equity, inclusion, and accessibility? What additional data should be collected to address these questions?

KA believes that IES should not be asking what information exists in datasets to answer questions around diversity, equity, inclusion, and accessibility (DEIA) but, rather, should ask how education researchers be enabled to tap into the vast amounts of data in operating systems that can answer DEIA questions. We think that a lot of the data needed to address research questions related to DEIA presently exist in data streams but we do not invest in the infrastructure to tap into those data. Once we can tap into data streams, we will be better equipped to uncover what additional data are needed to answer questions around DEIA.

With regard to the question around what additional data should be collected to address DEIA research questions, KA encourages IES to leverage the efforts of the Comprehensive Center (CCs) program, which supports the establishment of 20 CCs that provide capacity-building services to (SEAs, LEAs, and schools to improve educational outcomes for all students, close achievement gaps, and improve the quality of instruction. The National Comprehensive Center (NCC), one of the centers. is already making headway on the data SEAs and LEAs can turn to in order to answer questions relating to DEIA. Starting in February of this year, the NCC began a community of practice with teams from seven states using the 2019 NASEM report, "*Monitoring Educational Equity,"* to identify key indicators for measuring and monitoring the extent of equitable opportunities and outcomes in low-performing schools.

The aforementioned NASEM report notes that "among indicators related to K-12 education, there is a range of measures that are readily available (performance on standardized assessments, on-time graduation, access and enrollment in rigorous coursework); measures that research indicates are important, but where there is not enough political or academic consensus on how to measure these variables (curricular breadth, access to high-quality supports, nonacademic supports); and indicators that are "in-between," that is, for which there is some consensus on how to measure these variables but more research is needed. This core group of indicators was selected by the Committee with the expectation that different educational agencies could select constructs or add indicators tailored to each local system's needs."

9. What role can IES play in developing infrastructure that supports the use of large-scale datasets for education research?

Use the Regional Educational Laboratories

KA encourages IES to think about how the RELs, which already work in partnership with SEAs, LEAs, and other education stakeholders, can leverage and expand on their work and relationships to support the use of data streams for education research and to inform policymaking. RELs already partner with districts, states, and other education stakeholders to identify high-priority needs and conduct applied research to address such needs, helping stakeholders understand problems and learn what is working in their schools. RELs produce clear, objective, and peer-reviewed research products designed to be actionable for partners and national audiences alike. RELs also develop toolkits that support the scaling up of best practices, such as those identified through the What Works Clearinghouse's Practice Guides.

RELs also support Training, Coaching, and Technical Support (TCTS) for Use of Research. TCTS projects leverage RELs' unique expertise in designing and interpreting rigorous, relevant research, as well as the identification and application of evidence-based practices. TCTS includes intensive training involving hands-on, direct instruction from experts in research or practice; coaching, or "thought partnering," that supports decisionmakers in applying research evidence to inform high-leverage decisions and actions; and technical support to build partners' capacity to identify, collect, analyze, and visualize data.

KA also encourages IES to consider how RELs can disseminate information about data stream best practices. REL dissemination activities are already designed to communicate research and evidence in a timely, accessible, and actionable manner. RELs are honest brokers and effective synthesizers of scientifically valid information in an age where information of varying quality is ubiquitous and readily transmitted. REL dissemination activities, products, and strategies are co-developed in partnership with policymakers and educators to help ensure that they can leverage and apply research evidence in their local context.

KA also encourages IES to think about the role of RELs with regard to the NASEM report recommendations, specifically the recommendations regarding knowledge mobilization and the recommendation to involve data users in the development of IES products that better meet their needs. RELs can serve as an important feedback loop to IES so to enable IES to engage in continuous improvement processes.

State Data Stream Coordinator

As IES considers what is needed to support a robust data steam for research purposes, we encourage IES to look to the success of the National Assessment of Educational Progress (NAEP) State Coordinator as a model that helps supports IES goals in states. The NAEP State Coordinators play a crucial role in acting as a liaison between the state, districts, and IES. KA can envision a similar type of position for a data stream coordinator within each REL who supports LEA and SEA adoption of best practices to help unlock the potential of data-rich operating systems to improve education research and find out what works, for whom, and under what conditions.

Data Privacy

KA believes IES can play a central role in helping SEAs and LEAs think about data privacy as it relates to tapping streams. To start, KA encourages IES to leverage the work NCES already undertakes to de-identify student record data and make restricted use-data available for research purposes. NCES is already required, "by law to develop and enforce standards designed to protect individually identifiable information of its respondents. This requirement includes protecting the information during the collection analysis, reporting, and publication of the data. The NCES Statistical Standards Program (SSP) has two major functions in the area of information protection. First, SSP houses the Disclosure Review Board (DRB), comprised of members from each NCES division, representatives from the SSP, and a member from each of the IES Centers. The DRB reviews disclosure risk analyses conducted by IES and NCES staff and contractors to ensure that data released do not disclose the identity of any individual respondent. The DRB approves the procedures implemented to de-identify the data that are released in public-use files. The DRB also approves the procedures used to remove direct identifiers from restricted-use data files. In the case of nationally representative sample surveys, the DRB also approves the procedures used to add perturbations to the data as a method of data protection.

The second major activity in this area is the restricted-use data license program. This licensing program provides external researchers access to individually identifiable IES and NCES data that are covered under Federal statutes and regulations by subjecting authorized researchers to the laws, regulations, and penalties that apply to use of confidential data held by IES. Under the license agreement, authorized researchers are subject to unannounced site inspections.”

KA strongly believes that as the capacity to tap into rich data streams increases, IES could and should play a role in issues around data privacy.

Please reach out to Rachel Dinkes at rdinkes@knowledgeall.net with any questions.

Best,

Rachel Dinkes

Rachel Dinkes, President