IDC IDEA DATA CENTER
Building Capacity for High-Quality *IDEA* Data

IDEA Data Quality: Outlier Analyses Brief

An introduction to the principles of outlier analysis

Authors
Chris Lysy
Danielle Crain

Westat®

May 2016
Version 1.0

The *IDEA* Data Center (IDC) created this document under U.S. Department of Education, Office of Special Education Programs (OSEP) grant number H373Y130002 (IDC). Richelle Davis serves as the project officer.

For more information about IDC's work and its partners, see https://ideadata.org/.

Suggested citation:
Crain, D., and Lysy, C. (2016, May). *IDEA Data Quality: Outlier Analyses Brief* (Version 1.0). *IDEA* Data Center. Rockville, MD: Westat.

Version Date: May 2016

IDEA Data Center (IDC)

IDEA Data Quality: Outlier Analyses Brief

IDC's principles for high-quality data include addressing data quality at all stages—collecting, submitting, analyzing, reporting, and using.
At the most fundamental level, high-quality data are timely, accurate, and complete.

- ***Timely data are current per a specific period of time.***
- ***Accurate data are:***
  - **Reliable, that is, consistent across time, methods, and locations; and**
  - **Valid, that is, representative of what they are designed to measure.**
- ***Complete data represent the intended population (e.g., national, state, or local level) and relevant subgroups (e.g., race/ethnicity, grade level, socioeconomic level, gender).***

Beyond these fundamental components, high-quality data are also *accessible* and *usable*.

- ***Accessible* data are readily available in formats that are understandable, user friendly, and practical.**
- ***Usable* data promote sound management, strong governance, and dedication to improving results for children and youth with disabilities and their families.**

Data security is essential. *Secure* data are collected and stored with due consideration to maintaining confidentiality and with electronic and physical protections commensurate with the sensitivity of the data.

Purpose and Intended Audience

Outlier analysis provides an important tool for examining data to identify observations (local education agencies (LEAs)/local lead agencies (LLAs), schools, students) with data that deviate from an established norm so that they can be investigated as possible data errors. This brief introduces the principles of outlier analysis. It is part of a suite of three technical assistance products IDC designed to be used by the state personnel responsible for the IDEA 618 and/or 616 data. In addition to this brief, these products include a tutorial on completing an outlier analysis and a tool state staff can use to conduct outlier analyses with their local data. All of these products may also be used by IDEA Parts B and C state staff working with LEAs and LLAs to analyze their local data. Any state staff with the ability to examine and analyze IDEA 618 and/or 616 data would also benefit from these technical assistance products. Such staff might include data managers, IT personnel, coordinators, and/or directors.

This brief provides an overview of outlier analyses. It is organized around four questions:

1. What is an outlier?
2. Why are outlier analyses important for data validity and reliability?
3. What steps should states take after an outlier analysis?
4. How can states conduct and display an outlier analysis?

What Is an Outlier?

"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins, 1980). Outlier analyses also include investigating whether the data are valid or invalid. In the fields of statistics and data mining, outliers may be referred to as anomalies, abnormalities, deviants, and discordant data. When conducting outlier analyses, states define what value or combination of values are outside the expected norm. These parameters can help specify the LEAs/LLAs that have data outside of the "normal" parameters set by the state. There may be times when states see

differences in LEAs/LLAs that are considered normal. It will be up to the state to determine what constitutes a "sufficient" anomaly.

Outlier analysis may identify valid as well as invalid data. Invalid outliers are the target of outlier analysis, as they represent errors in the data. On the other hand, valid outliers may appear to be outside the norm, but investigation demonstrates that the data are not in error. Valid outliers may occur due to random variation, which occurs due to chance and is inherent in a system.

Why Are Outlier Analyses Important for Data Validity and Reliability?

Outlier analysis is primarily important because it helps to identify errors in the data and, when investigated, may reveal systematic errors in data collection, coding, or entry. **Invalid outliers should be corrected, and the processes that resulted in such errors should be fixed**.

Outlier analysis can also be important because it may identify LEAs/LLAs that are performing better or worse than the norm. Identifying these high or low performers provides opportunities for understanding the factors behind high performance or providing targeted technical assistance where it is needed.

What Steps Should States Take After an Outlier Analysis?

After conducting an outlier analysis, states should investigate any identified outliers to understand why the data are so different from the norm. If the data are outside the parameters set for valid outliers, then states should follow up with the LEAs/LLAs to determine the root cause of the outlying data. Lists of investigative questions are provided below in the section "Questions to focus outlier investigations."

For more information on examining root cause, states can review *Equity, Inclusion, and Opportunity: How to Address Success Gaps, White Paper*.

How Can States Conduct and Display an Outlier Analysis?

States can use several possible approaches to conduct outlier analyses. The IDC outlier analysis step-by-step guide includes six different tutorials covering different methods states can use to identify and visualize outliers. Five of the six tutorials explain approaches using Microsoft Excel. The sixth approach uses Tableau to visualize outliers. See the *IDC Outlier Analysis Step-by-Step Guide.*

IDC also created an Excel-based tool states can use to identify outliers using the interquartile range approach described in the step-by-step tutorials.

Questions to focus outlier investigations[1]:

1. **Are the outliers found in just one LEA or LLA?**
   Knowing that all outliers are in just one LEA or LLA will help state personnel focus their investigation into the cause of outliers.
2. **Are the same outliers identified in more than one dataset?**

[1] The Office of Elementary and Secondary Education (OESE) compiled a list of possible causes of data quality problems related to the *Elementary and Secondary Education Act* (ESEA) and other data reporting. States may want to review the list to help determine areas of data quality that are affected by the outlier analysis. U.S. Department of Education. (April 2006). *Improving Data Quality for Title I Standards, Assessments, and Accountability Reporting: Guidelines for States, LEAs, and Schools.* http://www2.ed.gov/policy/elsec/guid/standardsassessment/nclbdataguidance.pdf.

State personnel may want to review LEAs/LLAs that have outlier data in more than one data submission. The outliers may indicate a need for the LEA/LLA to review the data entry or coding policies. It may also indicate that the LEA/LLA lacks understanding in the data required for the data submissions.

3. **Are multiple outliers commonly identified in the same LEAs/LLAs?**
   State personnel may want to review the similarities in demographics and/or data collection practices in these LEAs/LLAs.

4. **Are the LEAs/LLAs with outliers using non-standard data collection definitions?**
   State personnel may want to review the definitions used by LEAs/LLAs to ensure that, within the state context, they understand and use the definitions provided by the Office of Special Education Programs (OSEP) for the IDEA 618 data collections. For example, outliers in the discipline data could be due to an LEA's interpretation of the terms "suspension" and "expulsion." Outliers in the Part C exiting data may be due to how a local Part C program interprets the definition for "attempts to contact unsuccessful."

5. **Are the LEAs/LLAs with outliers using non-standard methods for aggregating the data?**
   States that collect aggregated data from LEAs/LLAs may want to review the methods locals use to aggregate child-level data to create totals. Inconsistencies in how LEAs/LLAs aggregate data could lead to outliers. For example, the state education agency (SEA) may want to review the methods LEAs/LLAs use to aggregate the number of children with disabilities by race/ethnicity to ensure that they appropriately count each child in the seven categories OSEP requires.

6. **Are the LEAs/LLAs with outliers using non-standard methods to collect the data?**
   State personnel may want to review whether LEAs/LLAs are using similar policies and procedures for collecting the data. This can include ensuring that all LEAs/LLAs use the U.S. Department of Education's race/ethnicity guidelines. It could also include ensuring all locals consistently define suspension and expulsion.

7. **Did the small *n* size affect the analysis?**
   States can analyze *n* size in two different ways. The *n* size can skew the results of the analysis. If an LEA/LLA has a small population, then it may lead to outliers because of the proportion to the rest of the state's LEAs/LLAs. To investigate if this scenario has occurred, users would disaggregate the data and determine if they are sensible based on the local small populations. The second scenario can lead to outliers when reporting IDEA data. States can review the *n* size used to calculate the Annual Performance Reports (APR) for Part B Indicator B4 around discipline and significant discrepancy or Part C Indicator C4 around child outcomes to determine if outliers have skewed the results of the analyses.

Resources

Ben-Gal, I. (2005). Outlier Detection. In: O. Maimon and L. Rockach L. (eds.), *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers.* Norwell, MA: Kluwer Academic Publishers. ISBN 0-387-24435-2.

National High School Center. (2007, May). *Dropout Prevention for Students With Disabilities: A Critical Issue for State Education Agencies.* Washington, DC: American Institutes for Research.

Early Childhood Technical Assistance Center. (ECTA) (2013). *Looking at Data*. Retrieved from http://ectacenter.org/eco/assets/ppt/LookingAtData_revised.ppt.

Gupta, P. (n.d.). *The ABCs of Statistics*. Retrieved from www.accelper.com/pdfs/ The_ABCs_of_Statistics.pdf.

Hawkins, D.M. (1980). *Identification of Outliers*. Netherlands: Springer.

Kriegel, H.P., Kroger, P., and Zimek, A. (2010). *Outlier Detection Techniques*. The 2010 SIAM International Conference on Data Mining, Columbus, Ohio. Retrieved from http://www.siam.org/meetings/sdm10/tutorial3.pdf.

Munk, T., Reedy, K., D'Agord, C., Inglish, J., and DuRant, S. (2014, October). *Equity, Inclusion, and Opportunity: Addressing Success Gaps White Paper* (Version 2.0). *IDEA* Data Center. Rockville, MD: Westat.

U.S. Department of Education. (2006, April). *Improving Data Quality for Title I Standards, Assessments, and Accountability Reporting: Guidelines for States, LEAs, and Schools.* Retrieved from http://www2.ed.gov/policy/elsec/guid/standardsassessment/nclbdataguidance.pdf.