# Quality Indicators of Secondary Data Analyses in Special Education Research: A Preregistration Guide

**Allison R. Lombardi** (ID) [1], **Graham G. Rifenbark**[1], **and Ashley Taconet**[1]

## Abstract

Secondary data analyses occur when new analyses are proposed for existing data. Although they are prevalent in special education research, there is little guidance on how to prepare secondary data analyses studies. Preregistration of secondary data analyses studies provides a nice opportunity and structure for fellow researchers to share innovative questions and analytic approaches to existing data sets as well as increase transparency. In this manuscript, we (a) describe quality indicators of secondary data analyses consistent with open science practices and (b) provide applied examples of these indicators from a sampling of published studies based on two iterations of data from the National Longitudinal Transition Study (NLTS2 and NLTS2012) with the overall goals to provide guidance to authors and peer reviewers and promote collaboration among fellow researchers engaged in secondary analyses for a range of purposes.

Secondary data analyses occur when new analyses are proposed for existing data. Researchers may opt for secondary data analyses for a variety of reasons, such as examining a different unit of analysis, posing different research questions that are more in-depth than in the original study, or combining multiple extant data sets, to name a few. Although secondary data analyses are widely practiced among education researchers, very little guidance exists. In the context of open science, researchers are encouraged to preregister studies in order to remain transparent and reduce researcher biases (Nosek et al., 2018). As a result, research consumers will be provided with adequate information in order to make sound judgments about the study findings and implications. Although preregistration guidance specific to special education research has been established in certain methodologies (Cook et al., 2022; Johnson & Cook, 2019), no guidance is available on preregistration of studies using secondary data sources. Such guidance will be useful

considering the broad applications of secondary data analyses in education research.

Preregistration of secondary data analyses studies provides a nice opportunity for fellow researchers to share innovative questions and analytic approaches to existing data sets. Even though the data are already collected, preregistration is especially important in this context. Prior knowledge of the data is more likely, and thus preregistration provides a means to avoid potential researcher bias. Specifically, a practice known as "HARKing," or hypothesizing after results are known (Kerr, 1998; van den Aker et al., 2021), is a concern with secondary data studies. Preregistration also has the potential to reduce publication bias specifically with regard to publishing null findings. A refreshing attribute is

[1]University of Connecticut

**Corresponding Author:**
Allison R. Lombardi, University of Connecticut, 249 Glenbrook Rd. Unit 3064 Storrs, CT 06269-3064, USA.
Email: allison.lombardi@uconn.edu

that null findings are not viewed as a roadblock to publishing; rather, the results are fixed and therefore should be represented in the literature. We posit the goal should be reproducibility; that is, enough information must be reported in the manuscript so that another team or individual who has no prior knowledge of or connection with the study could independently reproduce the results.

> *As such, the proposed quality indicators are aligned with best practices in open science and will allow for seamless preregistration of secondary data analyses studies.*

In turn, this approach will provide guidance to and promote collaboration among fellow researchers engaged in secondary analyses for a range of purposes.

The quality indicators to designing secondary data analyses were informed by education and social science research (Baldwin et al., 2022; Logan, 2020; van den Akker et al., 2021). Specifically, van den Akker and colleagues (2021) described 21 detailed elements, embedding examples throughout as a tutorial. Baldwin and colleagues (2022) provide guidance on determining exploratory and testable hypotheses and implications for study design. Finally, Logan (2020) describes a decision-making framework easily adopted by education researchers that provides five concrete steps: selection, preprocessing, transformation, data analysis, and interpretation. Drawing from these resources, we emphasize quality indicators that are especially salient in special education research for secondary data analyses studies. Adherence to these quality indicators will enable researchers to follow the tenets of open science and preregister their studies accordingly. Table 1 lists the proposed indicators and criteria.

To further illustrate, we examined a specific area of the special education literature, secondary transition, for evidence of the quality indicators. We selected secondary transition because of the prevalence of secondary data analyses studies using multiple iterations of the National Longitudinal Transition Study (NLTS). From 1987 to 2015, the U.S. Department of Education funded three NLTS studies (Wave 1, NLTS in 1987; Wave 2, NLTS2 in 1990−1991; Wave 3, NLTS2012 in 2012−2013) to examine the characteristics, secondary school experiences, and postschool outcomes of youth with disabilities who receive services under the Individuals With Disabilities Education Act (IDEA). As such, the NLTS data collection efforts are the most comprehensive to date with regard to nationally representative sample of youth from all 13 disability categories recognized by IDEA. The NLTS uses a complex sampling scheme that ultimately allows for appropriate inferences regarding the population of interest, youth with disabilities. Specifically, we focused on the two most recent iterations, NLTS2 and NLTS2012. As such, by focusing on examples from the NLTS-based literature, our proposed quality indicators are grounded in large-scale data sets that focus almost entirely on youth with exceptionalities. At the same time, the quality indicators are applicable to large-scale data sets that may or may not focus on youth with exceptionalities.

Previously, Mazzotti and colleagues (2016) comprehensively reviewed NLTS-based studies specifically with the goal to identify articles that met the standards for quality indicators for correlational research (Test et al., 2009) and identified 14 peer-reviewed publications that confirmed predictors of postschool success (Mazzotti et al., 2016). Of these 14 articles, three were considered a priori, meaning the study was theoretically driven and hypotheses were planned prior to conducting analyses, and 11 were considered exploratory, meaning hypotheses were not stated prior to conducting analyses. However, this systematic review was limited to NLTS2-based studies and did not address open science, which at the time was in its infancy. As such, open science practices have not been addressed in a published review of NLTS-based studies to date. Even so, we found the Mazzotti et al. study especially informative in understanding the importance of establishing decision rules about group sizes prior to conducting analyses, a practice we posit is especially important in special education research given the importance on categories of exceptionality.

## Proposed Quality Indicators

We propose the following quality indicators that build on published recommendations for

**Table 1.** Quality Indicators and Detailed Descriptions.

| Quality indicator | Details |
| --- | --- |
| Research questions and hypotheses | Research questions are clearly stated and linked to testable hypotheses |
| | Planned analysis and potential interpretation of findings, including null results, are stated a priori |
| | An advanced organizer that maps these components is provided |
| Variable reporting | Actual variables names as listed in codebook are reported |
| | Variable types (e.g., parent reported, teacher reported, youth reported) are specified |
| | Newly computed variables are clearly described well enough so they could be reproduced |
| | Software code is available as open materials to show how to compute the variables |
| Statistical power | Power analyses are conducted a priori to determine decision rules regarding minimal group sizes |
| | The type of power analyses is described (e.g., simulation-based or prepackaged software program) |
| | Power analysis software code is available as open materials in accessible file formats |
| Complex sampling designs | Survey weights are reported specifying the variable name of the weight as it appears in the codebook and whether the weights are altered in any way (e.g., rescaling the weights) |
| | Cluster and stratum variables are reported as they appear in the codebook and specified in annotated software code files |
| Analytic approach | The analytic approach is described and could include either or both conventional methods (e.g., *t* test, ANOVA or MANOVA, chi-square) and modern methods (e.g., latent variable modeling, mixture models, factor analysis, item response theory, structural equation modeling, multilevel modeling) |
| Demographic variables: Race, ethnicity, and disability | Race, ethnicity, and disability of the analytic sample are reported, specifying actual variable names as they appear in the codebook |
| | Source of data is clear (e.g., parent report, district report, self-report, teacher report) |
| | New variables created based on information from both race and ethnicity variables are described (e.g., if race and ethnicity were combined into a single race-ethnicity variable) |
| | New variables created based on disability category are described (e.g., if certain categories are collapsed, this is clearly described and justified) |
| | Software code is available as open materials to show how to compute the variables. If multiple sources of data are consulted, robustness checks are conducted and reported |
| Missing data | Method of missing data treatment is reported and could include conventional methods (e.g., listwise deletion, hot-deck imputation) or modern methods (e.g., full information maximum likelihood, multiple imputation) |
| | Use of auxiliary variables is reported and specified |
| | For multiple imputation, the number of imputations (generated data sets), software, and seed should be specified as well as how variables were treated (e.g., Bayesian regression, logistic regression, predictive mean matching) |
| Software packages | The software package and version are reported |
| | Changes to default settings are reported |
| | The annotated code file is available as open materials in accessible file formats |

preregistered secondary data analyses studies within and outside of educational research (Baldwin et al., 2022; Logan, 2020; van den Akker et al., 2021) and address the following areas: research questions and hypotheses, statistical power, variable reporting, complex sampling designs, treatment of missing data, software packages, and knowledge of data. Table 1 shows definitions for each quality indicator. Throughout our description of the quality indicators, we provide examples from a sampling of published studies based on NLTS2 and NLTS2012 data to determine if the proposed fundamental elements were discernible. A full list of article citations by type of NLTS data (e.g., NLTS, NLTS2, NLTS2012), a list of frequency of publication by journal, and more details about the search process and inclusion and exclusion criteria are available in Appendix A in the online supplemental materials.

## Research Questions and Hypotheses

As in any study, the research questions should address an important and timely topic in special education. We recommend authors clearly state research questions and hypotheses together. Research questions can be thought of as more of a bridge or stepping stone to the testable hypotheses (van den Akker et al., 2021) and thus may be phrased more conceptually. Study hypotheses should be written in a testable way; in other words, testable hypotheses are specific, are quantifiable, and may broadly reference the methodology that will be used for testing or state directionality. It is important to include not only research questions but also testable hypotheses that are more specific and lead to planned analyses. In this vein, the research questions and testable hypotheses should be linked to

the planned analyses, sampling plan, and potential interpretation of results with the ultimate goal that readers could compare the initial study plan with the final results (Baldwin et al., 2022). Ideally, these steps are clearly mapped in a table that can serve as an advanced organizer of the planned study, as shown in Figure 1.

If researchers are unable to determine testable hypotheses, the study may be considered more exploratory in nature. If the study is exploratory, this will be important to clearly state. Even with exploratory research, authors should provide some plan for how results will be interpreted, especially in the case of null results. Further, authors might consider using a holdout sample (i.e., a subset of the full sample) for exploratory work that could then inform study hypotheses to be confirmed in a preregistered study (Baldwin et al., 2022). Similarly, Logan (2020) describes the initial step, selection, where the researcher identifies a target data set and sample. Feasibility of the holdout- or target-sample approach depends in large part on sample size. Many secondary data analyses studies incorporate large-scale data and thus allow for the potential to split the sample to use in two-part exploratory and confirmatory phases. If sample size is a restriction, researchers should make it clear the hypotheses are exploratory. The concept of a holdout sample is not new; in fact, in the context of measurement studies, it may be referred to as split sample or cross-validation (Kline, 2015), a well-accepted practice in assessment research.

Importantly, researchers who take this approach will be well equipped to submit their study as a Stage 1 registered report for peer review, a step that offers a high probability of publication to a specific journal regardless of null findings (Nosek & Lakens, 2014). If pursuing a Stage 1 registered report with a specific

| Research Question | Hypothesis | Sampling plan (e.g., power analysis) | Analysis plan | Interpretation |
|---|---|---|---|---|
|  |  |  |  |  |
|  |  |  |  |  |

**Figure 1.** Suggested format of advanced organizer of secondary data analyses studies.
*Note.* From Open Science Framework guidance on registered reports: https://osf.io/8mpji/.

journal (Nosek et al., 2018), authors should contact editors prior to submission to clarify if exploratory analyses are considered or if preregistered secondary-data-analyses studies must include testable hypotheses. Although not all journals offer the option to submit Stage 1 registered reports, researchers who follow these steps will be able to successfully preregister their study to a third-party registry, such as Open Science Framework (https://osf.io). An example of a table template to use as an aid in mapping a secondary data study is shown in Figure 1 (also available at https://osf.io/gcyzq). For an example of this approach in the context of an accepted Stage 1 registered report, see Lombardi and colleagues (2021; available at https://osf.io/f26pa). Finally, we elaborate on the sampling plan component in the Statistical Power section, as secondary data studies will inevitably have special considerations given the data are already collected and post hoc power analyses must be conducted.

Although we provide examples in the context of publicly available large-scale data sets that are focused on youth with exceptionalities, it is important to acknowledge the wide variety of secondary data sources. In some studies, secondary data are coupled with new data collection as a Stage 1 preregistration (see McCoach et al., 2022, for an example of how this might be presented as a Stage 1 registered report). In this example, researchers proposed to collect new data on teacher ratings as well as gather extant data on relevant academic and behavioral variables. On the surface, this may seem a study with primary data collection, yet gathering the extant school data is a type of secondary data analyses. Another approach is to combine two or more extant data sets, such as school-, district-, and state-level data (Cruz et al., 2022; Hall-Mills, 2019). Cruz et al. (2022) combined variables across multiple data sources to examine the impact of state-level policy changes on student outcomes. These researchers extracted data from multiple secondary sources in order to examine school-level (finance data, teacher demographics) and student-level data (state test scores), whereas Hall-Mills (2019) combined two large extant data sets to examine changes in language impairment prevalence after a statewide

mandate of response to intervention was put in place. Across all of these examples, multiple sources of data are combined whether they are primary and secondary data or multiple secondary data sources. Preregistration of these types of studies not only will increase transparency but also will potentially increase communication and collaboration among researchers with like interests who wish to access and analyze the data.

## Variable Reporting

Although it may seem straightforward, transparency in variable selection, computation, and reporting is a quality indicator of preregistered secondary data analyses. In order to ensure transparency, researchers should report the actual variables names as listed in the codebook. Describing or paraphrasing the variables and possible response options is not enough, especially because many large-scale data sets have several variables that are similar in wording. In addition to reporting actual variable names, researchers must thoroughly describe any newly computed variables. This includes simple decisions, such as collapsing categorical variables or reverse scoring items, as well as more complex computation of new scales, index scores, or latent constructs based on a combination of variables. Including annotated code files from statistical software programs of these data analyses is crucial. We recommend providing annotated code files as supplemental materials if page limit requirements are a concern. Fortunately, many journals allow for the addition of supplemental files that do not count toward page limits.

As mentioned, describing the variables using the codebook naming conventions is likely to increase the possibility for reproducibility. This is particularly important in the context of NLTS-based studies because these data include youth and parent report surveys as well as school- or district-reported variables, which is not unlike other large-scale data sets that include teacher-, parent-, and student-reported variables. Within the NLTS data, multiple variables can have the same names, and the codebook is the only reliable source to discern which variable was used.

An example item from the NLTS2012 data is the variable "Youth attended IEP/transition planning meeting," an item with the same text represented in the youth report (L1) and parent report (E1a). If authors do not provide the codebook variable name, it is difficult to know which item was used in the study.

In our review of variable reporting within the NLTS-based studies, we found 19% of the studies listed the variables as they appear in the codebook ($n = 24$). Although 70% of the studies reported newly computed variables ($n = 89$); none of the studies, and none overall, provided annotated software syntax or code files to show how these variables were computed so that independent researchers could reproduce the results. Despite the absence of annotated code files, of the 70%, some study authors described the computation of the new variables well enough in text that a reader could likely reproduce it (71%, $n = 63$), whereas other study authors did not describe their process sufficiently enough to reproduce it (29%, $n = 26$). It is promising that most study authors provided sufficient detail to reproduce newly computed variables, yet the absence of annotated code files will inevitably create roadblocks to reproducibility attempts.

We were particularly interested in how NLTS-based study authors treated race and ethnicity variables. We found race (87%) and ethnicity (84%) were reported frequently across studies and often together. Fewer studies (30%) reported newly computed variables that either collapsed race categories (e.g., White and non-White) or combined information from both race and ethnicity variables. Of the studies that reported newly computed variables ($n = 39$), some study authors described the computation of the new variables well enough in text that a reader could likely reproduce it (77%, $n = 30$), whereas other study authors did not sufficiently describe the new variable computation so that it could be reproduced (23%, $n = 9$), and none provided software syntax or code files to show how these variables were computed.

Overall, this examination of NLTS-based studies implies inadequate attention has been given to disentangling race and ethnicity in published studies. Specifically, it was typical to see race and ethnicity combined as one race-ethnicity variable, where it would be listed with the categories White, Black, and Hispanic. Given the nature of NLTS data collection procedures, it was entirely possible to check "Yes" for the ethnicity variable (e.g., "Are you Hispanic?") and check "White" and/or "Black" for a race category. Thus, in reporting race-ethnicity as White, Black, or Hispanic, it is possible for cases to count in multiple categories; therefore, without explicit decision rules, reported replication attempts would be futile. In a recent study using NLTS2012 data, race and ethnicity were combined by taking both into account rather than collapsing (Lombardi et al., 2021); however, these types of efforts remain the exception rather than the rule.

Many secondary analyses of large-scale data are too simplistic with regard to addressing the complexities of race and ethnicity. Moreover, these complexities become exponentially more challenging when disability categories are also considered in analytic study-sample groupings. Transparency in the decision-making of these groupings is an important yet often overlooked aspect to secondary data studies. In some cases, researchers found the analytic approach could even produce unexpected results that are largely contrary to the field's understandings on crucial issues in special education (e.g., Morgan et al., 2017).

Clearly, large-scale data sets are a wealth of information and provide either a snapshot of phenomena or, in the case of longitudinal studies, trends of the phenomena over time. In order to gain a complete understanding, researchers should undertake robustness checks on multiple sources of data to examine to what extent their inferences are invariant across data sources, subgroups, and methodological approaches. Commonly, different approaches to data analysis are available; therefore, it is important to include such robustness checks (i.e., sensitivity analyses) in the preregistration or amend the preregistration as the study progresses (van den Akker et al., 2021). Ultimately, if robustness checks are conducted, this benefits research and, perhaps, more so for research on diversity, equity, and inclusion.

A recent example of a robustness check was applied to disability data (Shogren et al., 2022). Self-report and administrative disability data were gathered and compared in order to test if both had the same impact on the study outcomes. After conducting these robustness checks, Shogren et al. (2022) determined that disability data source had meaningful differences related to their study outcomes. It is common for multiple sources of information regarding disability status and disability type to be available to researchers analyzing secondary data sets. Specifically, disability status or type could be self-reported or parent, teacher, school, or district reported. If two sources are available (e.g., self- and district reported), researchers could employ McNemar's test, which is strictly a $2 \times 2$ contingency table method. If the McNemar's test is significant, difference exists between the data sources and researchers must carefully decide which data source to use as this may have an impact on study findings. If more than two sources are available or agreement over disability types is of interest, more general chi-square tests could be used to determine if there is an association between disability status or type and source. Ultimately, it is important for researchers to consider multiple sources of disability data if possible (Shogren et al., 2022); thus, if multi-informant data are available, robustness checks are an important step to take in the data analysis plan.

## Statistical Power

Post hoc statistical power analyses on secondary data are crucial with regard to preventing "data snooping" as well as establishing a priori decision rules on adequate group sizes. We cannot emphasize enough the importance of empirically establishing a priori rules pertaining to group sizes, especially in the context of research on children with exceptionalities, as well as factoring in race and ethnicity in these study designs.

Given reasonable population parameters and expected effects, Monte Carlo power simulations can be executed to determine the sample size required to detect an effect while balancing type I and type II errors. However, with secondary data analysis, as the name implies, the data have already been collected; therefore, researchers are limited to the data available. It stands, then, that any power analysis conducted using information from the available data (e.g., group sample sizes available) would be considered post hoc and could give rise to research questions that stem from data snooping. Of course, this should be avoided at all costs.

*Thus, power analyses allow for researchers to determine a* priori *decision rules about group sizes without compromising best practices.*

For example, after conducting a power simulation, researchers might determine a group size must meet a minimum threshold of 125 participants. Any group that does not meet this threshold would then not be included in the data analyses. Stating these decision rules in a Stage 1 report is essential for transparency. Examples of power simulation code in R can be found online (https://osf.io/g5wkb). Researchers might also consider prepackaged programs, such as Optimal Design (Raudenbush et al., 2011) or PowerUp! (Dong & Maynard, 2013).

With regard to statistical power in the NLTS-based studies we examined, only one study reported conducting a power analysis to ensure the sample was sufficiently powered (Leppo et al., 2013). However, five studies (4%) described a priori decision rules in order to determine adequate group sizes (Grigal et al., 2011; Schuck et al., 2018; Shogren et al., 2014; Zablocki & Krezmien, 2013; Zhou et al., 2012). Although these a priori rules varied quite a bit across these studies, all were tied in some way to requirements of the planned analyses, with two studies providing citations to justify these decisions (Schuck et al., 2018; Shogren et al., 2014). Thus, although the NLTS-based literature has some examples, generally the published literature does not address statistical power adequately.

*Small samples.* Routinely, special education researchers face the challenge of making inferences with small sample sizes, especially when

examining subgroups. For example, it may be of interest to determine whether a given effect holds across subgroups, such as disability types. Conducting a power analysis assists in determining, based on a chosen analytic approach, the minimum group size needed prior to conducting the analyses. In this way, researchers ensure enough information is available per group for acceptable parameter recovery by examining root mean square error, standard deviation, and bias over Monte Carlo simulation replications or by evaluating parameter coverage to determine what proportion of replications have confidence intervals that include the population value (i.e., parameter coverage). Additionally, when researchers are faced with small sample sizes, careful consideration should be given to their chosen methodology. For instance, if the chosen analytic approach involves latent variable regression, a feasible approach could be to conduct factor score regression (FSR) with Croon's correction, which accounts for the uncertainty in the factor scores (Croon, 2002; Devlieger et al., 2019). Alternatively, Bayesian approaches, such as Bayesian hierarchical modeling, could be applied, wherein hyperpriors (or adaptive priors) are estimated from the data and information is pooled across subgroups, affording robust predictions even when subgroup sample sizes are small (Gelman & Hill, 2006), assuming effects will not differ wildly across subgroups. It is also worth mentioning that Bayesian hierarchical modeling could be used in lieu of FSR with Croon's correction. Additionally, researchers should consider examining Bayes factors to aid in interpreting null findings.

Moreover, small samples may be addressed by combining extant data sets. Open data repositories are becoming more available in light of the open science movement; see Talbott and colleagues (2023) for an examination of open data sets with multiple data sources. In addition to large-scale national data sets, there are school, district, and statewide data, to name a few. It is important to consider a wide variety of sources as secondary data and key considerations when analyzing secondary data within special education research.

*Dependency in data.* Dependency is often present in education data where students are nested in classrooms, which are in turn nested in schools, or alternatively, repeated measures are nested within students and so on. In such cases, the proportion of the total variance attributed to school, classrooms, and students can be determined by estimating the intraclass correlation coefficient (ICC). To determine whether it is necessary to use multilevel modeling to account for the dependency in the data, researchers should evaluate the estimated ICCs but also determine whether other approaches for accounting for clustering, such as using generalized estimating equations or cluster robust standard errors, are feasible (McNeish et al., 2017). Effectively, statistical power in the context of multilevel modeling depends on the number of clusters at the top of the hierarchy (Maas & Hox, 2005); therefore, alternative methods may provide more statistical power than multilevel modeling when few clusters are available.

## Complex Sampling Designs

Another aspect of secondary data analyses that should be considered is the type of sampling that was employed to collect the data. For instance, nationally representative data sets do not utilize simple random sampling (SRS); rather, these types of data sources utilize a more complex sampling scheme that ultimately allow for appropriate inferences regarding the population of interest. Observations collected from an SRS scheme are weighted equivalently and therefore have the same influence on the likelihood function (e.g., each observation has a weight of 1); however, an exception to this would be in the context of propensity score weighting (PSW). In the case of both SRS and PSW, when the weights are added, the sum arrives at the total number of observations. On the other hand, given a complex sampling design, observations are either up-weighted or down-weighted, thus exacting a purposeful influence on the likelihood function. When the weights are added, the sum arrives at the target population (e.g., the target population for NLTS2012 was 22.5 million students).

When analyzing data that stem from a complex sampling design, researchers must account for these complexities by introducing

stratum, cluster, and weights into their statistical models; otherwise, inferences are sure to be incorrect. Researchers should report all sampling weights and cluster and stratum variables that were used in data analyses as they appear in the codebook. This is important because it is common for nationally representative data sets to include multiple sampling weights that serve different purposes; if studies are opaque regarding which complex sampling variables were used, then replication is less likely. With regard to complex sampling designs, most studies (84%) described the complex sampling design and weighted procedures associated with the NLTS data. However, fewer studies provided a level of detail that would be required to reproduce data analyses; specifically, altering or rescaling the weights (10%) and reporting the cluster (5%) and stratum (5%) variables that were used in the statistical models are needed in order to reproduce the results. Because no studies provided annotated code files (see upcoming Software Packages section), opportunities were missed to openly share how stratum, cluster, and weights are factored into the model-building process. Moving forward, it is crucial for authors to include annotated code files that show this information as open materials.

## Missing Data

Properly reporting details about treatment of missing data is essential to reproducibility. It is likely that all data, including secondary data, will have missing values on key variables. When data are missing on analytic variables, this has implications on statistical power, parameter bias, and recovery (Little et al., 2014). Rubin (1976) delineates the possible mechanisms that result in nonresponse: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data are MCAR when the reason for the missingness is unrelated to all variables in a data set (e.g., an individual is unreachable at the time of the survey). Data are MAR when missingness is related to other variables collected in the study in a predictable way (e.g., student refuses to respond). Finally, data are MNAR when missingness is due to

low levels on the attribute (or phenomena) being measured (e.g., nonresponse on self-determination items are due to the individual having low self-determination).

The strategies for handling missing data are categorized as either conventional (traditional) or modern. Examples of the former include mean substitution, hot-deck imputation, and list-wise or pairwise deletion (also referred to as complete case analysis). These traditional approaches to handling missing data assume that the missingness is due to an MCAR process, with an exception being hot-deck imputation, as this strategy involves imputing conditional means instead of an overall mean (i.e., mean values are determined based on values on some key variable).

Modern strategies for handling missing data include full-information maximum likelihood (FIML) and multiple imputation (MI). Specifically with large-scale data sets, secondary data analyses studies often showcase these approaches. Although FIML and MI perform similarly, these strategies differ from one another considerably. Specifically, FIML is a model-based approach to missing data in which missing values are estimated simultaneously along with all model parameters; therefore, to aid in the estimation of the missing data, the modeled relations among observed variables assists with the recovery of missing data. Additionally, if the data are missing due to an MAR process, the identified variable can be inserted as an auxiliary variable, which informs unbiased estimation. On the other hand, MI is an asymptotic procedure and can be described as a data-based approach to missing data. As such, MI is conducted as a first step (e.g., data preprocessing) from which $M$ imputed data sets result, where $M$ corresponds to the number of imputations. Due to MI being an asymptotic approach, it is necessary for $M$ to be sufficiently large (e.g., $M = 100$). Once the imputations are executed, the statistical model is fitted $M$ times and the results (estimates, standard errors) are pooled using Rubin's rules (Rubin, 1987). Importantly, when data are missing due to MCAR or MAR processes, FIML and MI will maximize statistical power, provide unbiased estimates, and will sufficiently recover the missing values (Enders, 2010). However, when data are

missing due to an MNAR process, both MI and FIML will not produce unbiased estimation of the missing data, and special modeling is required to account for this process (e.g., use of proxy measures that are related to the missing values).

Depending on the context and the type of analyses, researchers may or may not make conscious decisions on how to treat missing data. For example, a novice researcher may not understand that, by default, Mplus will employ FIML when a factor-analytic model is fitted or that listwise deletion is employed by default when estimating logistic regression. Therefore, it is imperative that specific information be shared to increase the chances for reproducibility. Depending on what strategy was employed to recover the missing data, different information is needed—here we focus on FIML and MI. Namely, if FIML is used, it is necessary for the hypothesized model to be fully specified (ideally in an annotated syntax file) along with any variables in the data set that were employed as auxiliary variables. For example, if the data analysis is a factor-analytic model (e.g., structural equation modeling [SEM] of confirmatory factor analysis [CFA]), the factor structure is highly important as it has ramifications on the model-implied relationship among observed variables, which aids in the estimation of the missing data. Seeing as MI is a simulation-based approach, when this strategy is utilized, more details are needed in order for analyses to replicate. Specifically, the software program, the number of imputations requested (specify $M$), the manner in which each observed variable is treated (e.g., Bayesian regression for continuous variables, logistic regression for binary variables), the seed for number generator, and in the context of complex sampling, the variables that capture the nature in which data were collected (e.g., sampling weights, stratum, cluster) must be specified—with all MI code annotated and made available. Importantly, imputation software programs differ from one another, although with a sufficiently large $M$ and proper MI diagnostics, results should converge on the same solution. Regardless of what strategy is used, if its respective information is not supplied, then replication is not guaranteed.

Treatment of missing data was reported by a majority of study authors (67%) in our review of published NLTS-based studies. Of those studies that had addressed missing data ($n = 86$), nearly half (47%) reported using conventional methods (e.g., listwise deletion, mean replacement), followed by MI (37%) and FIML (7%). For those studies reporting MI ($n = 32$), most reported the number of imputations and specified the software (87%).

*Missing disability data.* When missing values occur on variables such as disability type, the task of recovering this sort of missing data is challenging seeing as this variable and others like it (e.g., race, ethnicity) are nominal in nature. To gain insight into the mechanism (e.g., MAR) responsible for the missingness, we recommend researchers to estimate a series of logistic regressions to determine whether any measured variables are related to the missing values. The proportion of missing data should be reported along with any insights gained from estimating the logistic regressions, which will help to describe the missingness. If researchers are interested in recovering such data, Lang and Wu (2017) recommend using multiple imputation with chained equation via polytomous regression as implemented in the R package mice (Van Buuren & Groothuis-Oudshoorn, 2011). Their recommendation stems from a Monte Carlo simulation they conducted that compared the five most common MI applications in the context of nominal data.

## Software Packages

Regarding reproducibility, it is important for researchers to not only disclose what software program and version they used for the analyses but also provide annotated code for the analyses they report. This is especially important for reproducibility. Namely, if researchers attempt to replicate findings and do so with a different program, then such an attempt may be futile. Given the number of software programs available, it is important to understand how they might differ from one another in order to troubleshoot should issues arise.

Depending on the planned analysis, different software programs may be available. When researchers plan to use simple statistical

approaches (e.g., linear or logistic regression), more software packages will be available, but it is necessary to disclose the package used as their defaults are likely to be different across packages. For example, when estimating a logistic regression in SAS using PROC SURVEYLOGISTIC, the default is to model the probability of failure, whereas when estimating the same model in R using the glm function, the default is to model the probability of success. If defaults are changed, this should be made clear. These types of decisions are easily chronicled in a supplemental annotated software syntax or code file.

In situations where more advanced statistical methods are used (e.g., multilevel modeling [MLM], CFA, SEM), not as many software programs may be available. Moreover, not all programs perform the same or share the same defaults. For instance, when testing for the presence of a random slope in MLM (e.g., attempting to estimate small variance components), software programs will differ with respect to convergence rates and estimation results (McCoach et al., 2018), and these differences are mostly due to actions taken "behind the scenes" to ensure model convergence. In terms of latent variable methods (e.g., CFA and SEM), software programs may differ based on the default method of model identification (e.g., fixed factor vs. marker variable), the strategy for recovering missing data, or the default parameterization used when estimating categorical CFAs (e.g., delta vs. theta).

Finally, software programs will vary with respect to the way sampling weights are treated. Specifically, paid-for programs (e.g., Mplus and SAS) differ from open-source programs (e.g., the stat package in R), with the latter putting the onus on the researcher. In other words, when sampling weights are included in regression models estimated in R using the stat package, for example, lm() for linear regression and glm() for generalized linear models, the model will not fit due to the sampling weights not summing to the number of observation put forth for the analysis, whereas programs like Mplus and SAS will rescale the sampling weights such that the weighted sum comes to the total number of observations. Obviously, failure to disclose what program is used and what options are called for will impact the potential for reproducibility and replication.

Within our review of NLTS-based studies, software packages used on all study analyses were reported across a majority of the studies (67%). The most frequently used software products were SPSS (36%), SAS (29%), Stata (22%), Mplus (17%) and R (2%). As previously mentioned, annotated code files for these software programs were not made available in any published studies as open materials. It seems common practice to report the software product name and version, yet researchers should go beyond that to include the finer details of their analyses. Providing an annotated code file in supplemental materials would resolve this issue.

*Availability versus accessibility.* While we call for authors of future secondary data analyses studies to be transparent and openly share materials, it is important to make a distinction between making materials available and making them accessible. Simply, when materials are made available, it is not enough to provide them in a format that is foreign to the program (e.g., R), such as pdf; rather, all annotated code for data preprocessing, power analyses, and model syntax should be available in its native format. In this way, it is ensured that those who are interested in replicating analyses are able to download the original file(s) and conduct the analysis on their own without risking any loss of information that may result from copying and pasting text from a pdf. In the same vein, when data are made available, it is important to be considerate of potential consumers of the data, and therefore the data should be made available in a universal and *accessible* format, such as comma-separated value or tab delimited—this should also be the minimum when making materials such as code or syntax available. For instance, if data are available in an Excel format, there may be loss of information due to the use of functions or links to materials not available; on the other hand, data made available as a pdf is usually more of a hindrance than a help.

## Knowledge of Data

In this section, authors should report their prior knowledge of the data, including previous

studies, variables, analyses, and results. The purpose is to be transparent about what the authors already know about the data and to reference previous related studies in order to determine if the proposed hypotheses present some amount of bias. Of course, with secondary data analyses, researchers will inevitably become more acquainted with the data set as each study progresses. There must be some point of saturation where the researchers cannot preregister studies anymore without exhausting a certain amount of the variables. However, this threshold has yet to be determined. It could vary by journal or research question and may boil down to a value judgment. Authors should contact journal editors prior to submission with any questions about prior knowledge of data. In some ways, this section could be similar to the Dissemination History section in a grant proposal, where the authors describe previous related work along with relevant citations. If authors have used the data set in previous studies, it is important to describe which variables have already been used and for what group of observations (e.g., overall vs. a subsample). To the greatest extent possible, newly proposed studies should focus on other variables of interest that are less familiar to the research team or on groups of observations that have yet to be examined. In our review, no published NLTS-based studies included a dedicated Knowledge of Data section.

## Implications for Future Research

Overall, our review of published NLTS-based studies through an open science lens revealed promising and concerning trends related to the proposed quality indicators. Promising trends include the consistent reporting of variables as they appear in the codebook as well as sufficient-enough descriptions that could potentially be reproduced. These trends suggest study authors may aim to be transparent, yet the absence of established reporting guidelines makes it difficult to determine precisely what information is essential to report. On the other hand, post hoc power analyses were rarely reported. In anticipation of this result and based on the Mazzotti et al. (2016) study, we

specifically searched for reporting of a priori decision rules about group sizes and found very few study authors had these in place. Also, the finer details about complex sampling designs (e.g., cluster, stratum variables included in models) were lacking. Furthermore, the absence of annotated software code files as supplemental materials and a dedicated Knowledge of Data section reveals a major gap in author publishing practices that are far from open science approaches. Similarly, most authors reported on treatment of missing data. However, fewer were able to elaborate on the implications for selecting conventional (listwise deletion) or modern methods (MI or FIML), and in most cases, not enough information was provided so that procedures could be followed and successfully reproduced by independent researchers.

Although we propose quality indicators to conducting secondary data analyses, we acknowledge these types of studies certainly are not new or novel. In fact, secondary data analyses are already largely prevalent, and special education researchers have been conducting secondary data analyses over the past several decades without any agreed-upon standards or guidance related to reproducibility consistent with open science practices. It is unlikely that reproducibility was ever a goal in published studies. Moving forward, we recommend reproducibility should underpin all efforts in secondary data analyses. It will be important to determine how the indicators could be used with smaller secondary data sets. Most of the quality indicators we propose in Table 1 could be applied to smaller secondary data sets from schools and districts. One exception may be complex sampling designs. All other indicators are applicable. Although we touched on the potential of combining extant data sets, there is much more to explore that goes beyond the scope of this manuscript.

In order to reach the goal of reproducibility of secondary data analyses studies, transparent reporting of the quality indicators must become more widespread.

> *Authors must become more accustomed to sharing not only data but also materials to analyze data, including syntax and code files.*

Moreover, code files must be annotated so that independent researchers can logically follow along. This means authors must devote substantial preparation time to write annotated code files. They should not be an afterthought or hastily thrown together at the time of submission. Moreover, journal editors might consider implementing requirements to provide these files for any secondary-data-analyses study. Recently, Fleming and Cook (2022) conducted a review of special education journal and publisher guidelines examining 51 journals. They focused on open-access features, such as preprints, postprints, open-access publishing, article processing charges, and embargo periods. Although this review was especially helpful with regard to understanding the prevalence of these open-access practices within special education research, there was not a focus on open materials within these journal and publishing guidelines. In order to promote the proposed quality indicators for secondary data analyses, it will be crucial for special education journal editors to require that authors provide open materials, such as annotated code files, and specific details about software program and specifications. Finally, given the recent prioritization on open-access data and materials by the White House (2022), infrastructure and supports for university researchers need greater attention. Funding for infrastructure to house data and materials that reside outside of journal and university paywalls must be in place.

Given the prevalence of secondary-data-analyses studies in special education and research broadly, the preregistration process allows for the reporting of null results with the goal to reduce publication bias. If we are to progress in the era of open science, it will be especially important to provide guidance and support to study authors and peer reviewers. Adhering to the proposed quality indicators will increase transparency, reproducibility, and collaboration and ultimately improve the overall quality of education research.

## References

Baldwin, J. R., Pingault, J. B., Schoeler, T., Sallis, H. M., & Munafò, M. R. (2022). Protecting against researcher bias in secondary data analysis: Challenges and potential solutions. *European Journal of Epidemiology*, *37*, 1–10. https://doi.org/10.1007/s10654-021-00839-0

Cook, B. G., Fleming, J. I., Hart, S. A., Lane, K. L., Therrien, W. J., van Dijk, W., & Wilson, S. E. (2022). A how-to guide for open-science practices in special education research. *Remedial and Special Education*, *43*(4), 270–280.

Croon, M. (2002). Using predicted latent scores in general latent structure models. In Marcoulides, G., & Moustaki, I. (Eds.), *Latent variable and latent structure models* (pp. 195–223). Lawrence Erlbaum.

Cruz, R. A., Lee, J. H., Aylward, A. G., & Voulgarides, C. K. (2022). The effect of school funding on opportunity gaps for students with disabilities: Policy and context in a diverse urban district. *Journal of Disability Policy Studies*, *33*(1), 3–24. https://doi.org/10.1177/1044207320970545

Dong, N., & Maynard, R. (2013). PowerUp! A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24–67. https://doi.org/10.1080/19345747.2012.673143

Devlieger, I., Talloen, W., & Rosseel, Y. (2019). New developments in factor score regression: Fit indices and a model comparison test. *Educational and Psychological Measurement*, *79*(6), 1017–1037. https://doi.org/10.1177/0013164419844552

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Fleming, J. I., & Cook, B. G. (2022). Open access in special education: A review of journal and publisher policies. *Remedial and Special Education*, *43*(1), 3–14. https://doi.org/10.1177/0741932521996461

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Grigal M., Hart, & & Migliore, D., A. (2011) Comparing the transition planning, postsecondary education, and employment outcomes of students with intellectual and other disabilities. *Career Development for Exceptional Individuals*, *34*(1), 4–17. https://doi.org/10.1177/0885728811399091

Hall-Mills, S. (2019). A comparison of the prevalence rates of language impairment before and after response-to-intervention implementation. *Language, Speech, and Hearing Services in Schools*, *50*(4), 703–709. https://doi.org/10.1044/2019_LSHSS-18-0144

Johnson, A. H., & Cook, B. G. (2019). Preregistration in single-case design research. *Exceptional Children*, *86*(1), 95–112. https://doi.org/10.1177/0014402919868529

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford Press.

Lang, K. M., & Wu, W. (2017). A comparison of methods for creating multiple imputations of nominal variables. *Multivariate Behavioral Research*, *52*(3), 290–304. https://doi.org/10.1080/00273171.2017.1289360

Leppo, R., Cawthon, S., & Bond, M. (2013). Including deaf and hard-of-hearing students with co-occurring disabilities in the accommodations discussion. *Journal of Deaf Studies and Deaf Education*, *19*(2), 189–202. https://doi.org/10.1093/deafed/ent029

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, *39*(2), 151–162. https://doi.org/10.1093/jpepsy/jst048

Logan, T. (2020) A practical, iterative framework for secondary data analysis in educational research. *Australian Educational Researcher*, *47*, 129–148. https://doi.org/10.1007/s13384-019-00329-z

Lombardi, A., Rifenbark, G., Hicks, T., Taconet, A. V., & Challenger, C. (2021). *College and career readiness support for youth with and without disabilities based on the National Longitudinal Transition Study 2012*. [Stage-1 Registered Report] https://doi.org/10.17605/OSF.IO/C9D4P

Lombardi, A. R., Rifenbark, G. G., Hicks, T. A., Taconet, A., & Challenger, C. (2022). College and career readiness support for youth with and without disabilities based on the National Longitudinal Transition Study 2012. *Exceptional Children*, *89*(1), 5–21. https://doi.org/10.1177/00144029221088940

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Mazzotti, V. L., Rowe, D. A., Sinclair, J., Poppen, M., Woods, W. E., & Shearer, M. L. (2016). Predictors of post-school success: A systematic review of NLTS2 secondary analyses. *Career Development and Transition for Exceptional Individuals*, *39*(4), 196–215. https://doi.org/10.1177/2165143415588047

McCoach, D. B., Peters, S., Long, D., & Siegle, D. (2022). *Examining teacher variability in teacher rating scales for gifted identification* [Stage 1 registered report]. https://osf.io/ykv5h/

McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D, Gambino, A. J., & Bellara, A. (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics*, *43*(5), 594–627. https://doi.org/10.3102/1076998618776348

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114. https://doi.org/10.1037/met0000078

Morgan, P., Farkas, G., Hillemeier, M. M., & Maczuga, S. (2017). Replicated evidence of racial and ethnic disparities in disability identification in U.S. schools. *Educational Researcher*, *46*(6), 305–322. https://doi.org/10.3102/0013189X17726282

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., & Hill, C. (2011). Optimal design software for multilevel and longitudinal research (Version 3.01) [Computer software].

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Schuck, L., Emerson, R. W., Kim, D. S., & Nelson, N. W. (2018). An approach to using orientation and mobility (O&M) variables from the second National Longitudinal Transition Study. *Journal of Visual Impairment & Blindness*, *112*(2), 203–208 https://doi.org/10.1177/0145482X1811200208

Shogren, K. A., Kennedy, W., Dowsett, C., Garnier Villarreal, M., & Little, T.D. (2014). Exploring essential characteristics of self-determination for diverse students using data from NLTS2. *Career Development and Transition for*

*Exceptional Individuals*, *37*(3), 168–176. https://doi.org/10.1177/2165143413486927

Shogren, K. A., Pace, J. R., Wittenburg, D. C., Raley, S. K., Hicks, T. A., Rifenbark, G. G., Lane, K. L., & Anderson, M. H. (2022). Self-report and administrative data on disability and IEP status: Examining differences and impacts on intervention outcomes. *Journal of Disability Policy Studies*. https://doi.org/10.1177/10442073221094811

Talbott, E., De Los Reyes, A., Kearns, D. M., Mancilla-Martinez, J., Cook, C. R., & Wang, M. (2023). Evidence based assessment in special education research: Advancing the use of evidence in methods, tools, and empirical processes. *Exceptional Children*, *89*(4).

Test, D. W., Mazzotti, V. L., Mustian, A. L., Fowler, C. H., Kortering, L., & Kohler, P. (2009). Evidence-based secondary transition predictors for improving postschool outcomes for students with disabilities. *Career Development for Exceptional Individuals*, *32*(3), 160–181. https://doi.org/10.1177/0885728809346960

van den Akker, O. R., Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P, Hall, A., Kosie, J., Kruse, E., Olsen, J., Ritchie, S., Valentine, K., van't Veer, A., & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2020.2625

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

White House. (2022). *Breakthroughs for all: Delivering equitable access to America's research*. https://www.whitehouse.gov/ostp/news-updates/2022/08/25/breakthroughs-for-alldelivering-equitable-access-to-americas-research/

Zablocki, M., & Krezmien, M. P. (2013). Drop-out predictors among students with high-incidence disabilities: A National Longitudinal and Transitional Study 2 analysis. *Journal of Disability Policy Studies*, *24*(1), 53–64. https://doi.org/10.1177/1044207311427726

Zhou, L., Griffin-Shirley, N., Kelley, P., Banda, D. R., Lan, W. Y., Parker, A. T., & Smith, D. W. (2012). The relationship between computer and internet use and performance on standardized tests by secondary school students with visual impairments. *Journal of Visual Impairment & Blindness*, *106*(10), 609–621. https://doi.org/10.1177/0145482X1210601005

## Authors' Note

## ORCID iD

Allison R. Lombardi 🆔 https://orcid.org/0000-0002-7254-8820

## Supplemental Material

The supplemental material is available in the online version of the article.