Published in final edited form as:

Qual Life Res. 2020 May ; 29(5): 1147–1158. doi:10.1007/s11136-019-02387-3.

Investigating child self-report capacity: a systematic review and utility analysis

Katherine B. Bevans[1], **Isaac L. Ahuvia**[2], **Taye M. Hallock**[1], **Rochelle Mendonca**[3], **Stephanie Roth**[1], **Christopher B. Forrest**[4,5], **Courtney Blackwell**[2], **Jessica Kramer**[6], **Lauren Wakschlag**[2]

[1]College of Public Health, Temple University, 1913 N Broad Street, Philadelphia, PA 19122-6092, USA

[2]Northwestern University, Chicago, USA

[3]Columbia University, New York, USA

[4]Children's Hospital of Philadelphia, Philadelphia, USA

[5]University of Pennsylvania Perelman School of Medicine, Philadelphia, USA

[6]University of Florida, Gainesville, USA

Abstract

Purpose—To identify and evaluate methods for assessing pediatric patient-reported outcome (PRO) data quality at the individual level.

Methods—We conducted a systematic literature review to identify methods for detecting invalid responses to PRO measures. Eight data quality indicators were applied to child-report data collected from 1780 children ages 8–11 years. We grouped children with similar data quality patterns and tested for between-group differences in factors hypothesized to influence self-report capacity.

Results—We identified 126 articles that described 494 instances in which special measures or statistical techniques were applied to evaluate data quality at the individual level. We identified 22 data quality indicator subtypes: 9 direct methods (require administration of special items) and 13 archival techniques (statistical procedures applied to PRO data post hoc). Application of archival techniques to child-report PRO data revealed 3 distinct patterns (or classes) of the data quality indicators. Compared to class 1 (56%), classes 2 (36%) and 3 (8%) had greater variation in their PRO item responses. Three archival indicators were especially useful for differentiating plausible item response variation (class 2) from statistically unlikely response patterns (class 3). Neurodevelopmental conditions, which are associated with a range of cognitive processing challenges, were more common among children in class 3.

✉Katherine B. Bevans katherine.bevans@temple.edu.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-019-02387-3) contains supplementary material, which is available to authorized users.

Conclusion—A multi-indicator approach is needed to identify invalid PRO responses. Once identified, assessment environments and measurement tools should be adapted to best support these individuals' self-report capacity. Individual-level data quality indicators can be used to gauge the effectiveness of these accommodations.

Keywords

Patient-reported outcome measures; Data quality; Self-report capacity; Pediatric

Patient-reported outcome (PRO) measures are a ubiquitous source of information used in pediatric health research and clinical care. An assumption of PRO assessment is that individuals are the most reliable and accurate reporters of their own health experiences. Many child-reported outcome measures are intended for children as young as 8 years of age [1]. This assumes that most 8-year-old children are able to execute the mental functions needed to comprehend items and response options, evaluate and summarize their experiences relative to item meaning, and select response options that best represent their self-evaluation. The numerous cognitive capacities underlying these functions include reading comprehension and/or auditory processing, attention, working memory, long-term memory, temporal sequencing, and judgment. Variation in cognitive capacities, which is both developmentally normative for school-aged children and heavily influenced by exposure to home and formal learning environments, poses significant challenges to the reliability and accuracy of child PRO measurement.

Developmentally sensitive PRO measures (PROMs) attempt to accommodate for variation in children's cognitive capacities by minimizing assessment demands [2]. For example, item wording is refined to maximize item relevance and understandability [3, 4]. Other accommodative techniques intended to bolster children's self-report capacities include auditory or multimodal presentation of items, illustrated content, and pictorially represented response categories. These accommodations are presumed to enhance the reliability and validity of children's self-report, but there have been few systematic attempts to evaluate their impact on data quality among children with a broad range of cognitive ability levels and education and socio-cultural experiences. Such research is limited by the lack of validated techniques for detecting potentially invalid PROM responses at the individual level. This can include a lack of response variability, excessive response variation, and extreme, inconsistent, or improbable response patterns.

A variety of methods for detecting invalid PRO data have been proposed [5–8]. These methods fall into two categories: archival and direct. Archival methods are applied to data that have already been collected and thus, are useful for flagging cases with problematic response data post hoc. Direct measures require administration of special items and, thus, may be used to screen for self-report capacity before administering full PRO batteries. For example, direct measures may prompt for a clear correct answer as a way of measuring attentiveness [9, 10]. We conducted a systematic literature review to establish an exhaustive list of methods previously used to evaluate PRO data quality at the individual level and to assess their frequency of use. We reviewed studies involving children and adults because we anticipated that few pediatric studies directly assess child-report data quality. Moreover,

techniques applied to adult samples may be applicable to children. We also evaluated the feasibility and usefulness of archival data quality indices by applying them to survey data collected from children ages 8–11 years. Both the literature review and data analyses were conducted to inform recommendations for gauging PRO data quality and children's self-report capacity. Use of the methods will encourage uptake of child PROs in pediatric clinical care and research by increasing end-user confidence in the accuracy of these important outcomes. Additionally, individual-level data quality indicators are needed to support the development of assessment techniques that maximize opportunities for child self-report while also ensuring measurement precision.

Systematic literature review

We conducted a systematic literature review to identify articles that describe methods for evaluating the reliability and/or validity of individuals' PRO responses. Results are presented in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [11]. The PRISMA flowchart and checklist are included in Appendix.

Data sources

The search was conducted in collaboration with a medical librarian with systematic review expertise. We developed a detailed search strategy and conducted the search using the Web of Science Core Collection (Clarivate Analytics), Embase (Elsevier), ERIC (Ebscohost), and PsycInfo (Ebsco-host). The search strategy is shown in Appendix. The search strategy used a combination of subject headings and free text terms. The search was limited to peer review publications published from 2007 to present. The final search was completed on April 10, 2018. The search identified 4937 sources. Duplicate articles (n = 857) were omitted using Endnote X.7 for the deduplication of records.

Study selection

Article inclusion criteria were (1) written in English; (2) included a self-report measure; and (3) included a data quality indicator. Data quality indicators were defined as methods for determining whether individuals' responses to self-report measures are reliable, valid, trustworthy, or credible. Two blinded and independent reviewers screened the unique references (n = 4080) by title and abstract using Rayyan, a systematic review web application. A third independent reviewer resolved disagreements. Title and abstract screening resulted in the elimination of 3752 articles (33 were not written in English, 3230 did not include a self-report measure, 489 did not include a data quality indicator).

Review process

We reviewed full text for the remaining 328 articles. An additional 202 articles were eliminated upon full-text review for violating inclusion criteria: 14 articles were not written in English, 10 did not include a self-report measure, and 93 did not include a data quality indicator. An additional 85 articles were eliminated for their focus on malingering, the intentional manipulation of data for specific gains. Measures of malingering are designed to detect purposeful attempts to alter one's response for gain or to comply with social

desirability [8]. Such measures assume that respondents have a clear understanding of the self-report measure and possess the skills needed to intentionally alter their responses. Our goal was to identify methods for detecting biased responses due to poor item understanding, carelessness, or insufficient effort, as opposed to intentional faking. Therefore, we eliminated articles that focused exclusively on malingering.

Abstraction of data quality indicators

Two independent reviewers abstracted information about data quality indicators from the remaining 126 articles. A third independent reviewer resolved disagreements. We recorded 494 instances in which data quality assessment measures or techniques were applied. This included 133 unique direct quality measures representing 9 subtypes and 13 archival indicator subtypes. Table 1 describes the types of data quality indicators and the frequency of their use. A complete list of articles and data quality indicators is included in Appendix.

Data quality indicators used in pediatric research

Of the 126 identified articles, 12 (10%) described 23 instances in which data quality indicators were applied in pediatric populations. These articles are identified in Online Appendix Table 1A. Table 1 shows the frequency with which data quality indicators were applied in research involving children. Four of the 12 studies (33%) used a single type of data quality indicator; 7 (58%) used direct indictors only; 3 (25%) used archival indicators only; and 1 (8%) used a combination of direct and archival methods. Self-report measures of honest responding (e.g., “I am telling the truth on this survey”) were used in 5 studies [12–16] and 2 studies asked children to report how well they paid attention during the assessment [12, 13]. Unlikely symptoms and virtues were assessed in 3 studies, which were conducted in residential treatment facilities [17], juvenile detention centers [18], and psychiatric inpatient settings [19]. A single study used bogus/infrequency items to detect questionable data quality. Furlong et al. (2017) screened for endorsement of unusual response options (< 5% incidence) across multiple items (e.g., weighing > 225 lb).

In research involving children, 2 studies used psychometric synonyms/antonyms, which measures within-person correlations across item pairs with strong positive associations or strong inverse associations [19, 20]. Two studies assessed response invariability (“longstring”), the maximum number of consecutive items with the same response [21, 22]. Univariate outlier statistics were applied in 2 studies [21, 23] and 1 study assessed consistency in responding to reverse-coded items to identify invalid responders [21].

Application of archival data quality measures

We evaluated 8 archival data quality indicators in secondary analysis of self-report data that were previously collected from a general population sample of children ages 8–11 years. The dataset did not support the use of 4 of the identified archival indicators and we were unable to evaluate direct indicators because they were not administered at the time of data collection.

Participants

Participants included 1780 children ages 8–11 years. Children were 48% male, 81% White, 17% African-American, 3% of another race, and 3% Hispanic. Approximately 21% of children were living in poverty as indicated by U.S. Census Bureau poverty thresholds, and 39% were living in single parent households. Parents or primary caregivers of 1221 participating children (69%) completed a parent-report questionnaire.

Measures

Child-reported health

Children completed the Healthy Pathways Child-Report Scales as part of a study on associations between child health and school performance [24]. Analyses were conducted using children's initial (baseline) responses to six scales selected to represent diverse aspects of children's health and functioning: physical comfort (8 items, physically experienced distress such as pain, fatigue, and somatic complaints); emotional comfort (10 items, emotions and mood with emphasis on anxiety, anger, and depression); self-worth (7 items, one's satisfaction with self); family connectedness (8 items, feelings of belonging in one's family); peer connectedness (9 items, making friends, quality of friendships); and student engagement (6 items, the degree to which children are interested and invested in learning). All items assess the frequency of children's experiences using a 5-point response scale (never, almost never, sometimes, almost always, always). Three of the 6 scales (emotional comfort, self-worth, and student engagement) include a total of 5 items that are reverse coded so that for all items, higher values indicate better health or functioning (e.g., fewer physical symptoms, greater student engagement). Reverse-scored items afforded the opportunity to assess consistency in children's responses to positively and negatively oriented items from the same scale.

Chronic health/neurodevelopmental conditions and special healthcare needs

Parents completed the Children With Special Health Care Needs (CSHCN) Screener, a non-categorical measure of long-term health problems that require health services or cause functional limitations [25]. The screener identifies children who have a condition lasting at least 12 months that results in (1) needing or using medicine prescribed by a doctor, other than vitamins; (2) needing or using more medical care, mental health, or educational services than is usual for most children of the same age; (3) limitations in their ability to do the things most children of the same age can do; (4) receipt of special therapy, such as physical, occupational, or speech therapy; or (5) emotional, developmental, or behavioral problems for which they need treatment or counseling. Parents also indicated whether children have been diagnosed with asthma, epilepsy, attention deficit hyperactivity disorder (ADHD), a learning disability, or a speech impairment or delay.

Academic performance

Concurrent with administration of the child- and parent-report measures, reading and math standardized test scores were collected for 1762 (99%) of the participating children. The test score metrics differed across states and grade levels. Therefore, scores were transformed to

state-grade-specific mean of 100 and standard deviation of 15. Parents (n = 1240) completed the Healthy Pathways Parent-Report Academic Performance scale, a 6-item measure of children's performance on schoolwork, homework, reading, math, and remembering what they learned [26].

Procedures

Children were recruited from regular education 4th, 5th, and 6th grade classrooms. Children ranged in age from 8–13 years (grade 4: M=9.6, SD = 0.6; grade 5: M = 10.6, SD = 0.6; grade 6: M = 11.6, SD = 0.6) in 34 schools in Maryland (2 school districts) and West Virginia (1 school district). Students completed the Healthy Pathways Scales at school in the presence of research and school staff. Parent-report measures were sent home with children and returned directly to the researchers via U.S. mail. Research staff collected participants' standardized test scores from school records. Study procedures were approved by Institutional Review Boards at the Children's Hospital of Philadelphia, Johns Hopkins Bloomberg School of Public Health, and Marshall University.

Analyses

Data analyses were conducted using R version 3.5.3. Annotated R scripts are presented in Appendix. We calculated 8 data quality indicators for each respondent: 5 indicators of intra-individual response variation (invariability/longstring, individual reliability, psychometric synonyms, inter-item standard deviation, reverse-coded item inconsistency) and 3 indicators that an individual's item responses differed from those of other children in the sample (person-total correlation, Mahalanobis distance, polytomous Guttman errors).

Invariability (longstring) (LS)

The maximum number of consecutive items with the same response prior to item re-coding. This metric assumes that repeated use of the same response category indicates lack of understanding or careless responding [5].

Individual reliability (IR)

Average of within-person correlations between scores on randomly selected halves of each subscale, sampled 100 times [5, 27]. Higher correlations indicate greater response consistency within scales.

Psychometric synonyms (PS)

Within-person correlation across 20 item pairs that have sample-level item-to-total correlations > 0.50. Correlations were computed using the psychsyn function from the *careless* package in R [28]. Higher correlations indicate greater response consistency.

Inter-item standard deviation (SD)

Average of within-person standard deviations of items on each scale. Higher values indicate less consistency in responses to conceptually and empirically related items.

Reverse-coded item inconsistency (RI)

Average of differences between each of 5 reverse-coded items (after re-coding) and scores on the item's parent scale with the reverse-coded item removed. Higher scores indicate greater inconsistency in responding to reverse-coded and non-reverse-coded items that measure the same construct.

Person-total correlation (PT)

Correlation between the person's item responses and the mean of all others' item responses. Higher values indicate greater compliance with others' response patterns.

Mahalanobis distance (MD)

The distance between an individual's response pattern and the multivariate center of responses [5, 8]. We calculated MD for each scale using the *PerFit* package in R and averaged the scores to generate a single MD value [29]. Higher MD values indicate that a respondent is an outlier relative to the multivariate distribution formed by responses to all items.

Polytomous Guttman errors (GE)

Within each scale, Guttman errors were registered for each item pair when children endorsed a higher response category for the more difficult item than for the less difficult item. In its polytomous form, Guttman errors also account for the magnitude of this discrepancy (the distance between expected and actual response categories). We used the GPoly function in the *PerFit* package in R to assess the frequency of polytomous Guttman errors for each scale and summed the errors across scales [29]. Higher values indicate greater deviation from expected response patterns based on item difficulty.

Because all items on the Healthy Pathways Scales are unique, we were unable to assess consistency in responses to repeated items ('direct item repetition'). We were also unable to count omitted items because the computer-based survey administration platform required a response before advancing to the next item. Additionally, we chose not to calculate basic outlier statistics because they may reflect children's actual levels of the measured constructs rather than invalid responding [5].

We calculated descriptive statistics (mean, standard deviation, and range) for each data quality indicator and assessed their associations using Pearson correlation coefficients. We conducted latent profile analysis using the *tidyLPA* package in R to explore if and how scores on the archival data quality indicators contributed to the characterization of unique PRO response patterns that may be useful for identifying children with underdeveloped self-report capacity [30]. Using an iterative model testing process, we compared latent profile analysis (LPA) models with 2 to 5 classes according to Bayesian Information Criterion (BIC), Sample-Adjusted BIC (SABIC), and Akaike's Information Criterion (AIC) [31–33]. We characterized and compared scores on the data quality indicators and Healthy Pathways scales across classes. Lastly, we tested for between-class differences in child age, gender, health and neurodevelopmental conditions, special healthcare needs (SHCNs), and academic performance. We expected that younger children and those with health or

neurodevelopmental challenges that may interfere with the cognitive functions used in self-reporting (e.g., attention, reading comprehension, working memory) would be overrepresented in classes with problematic response patterns. We also expected that children in these classes would have relatively poorer standardized test scores and parent-reported academic performance. We tested these hypotheses using analysis of variance for continuous outcomes (child age, school performance) and Chi square for categorical outcomes (SHCNs, chronic health or neurodevelopmental conditions).

Results

Descriptive statistics (mean, standard deviation, and range) and associations among archival indices (Pearson correlation coefficients) are shown in Table 2. LPA model fit criteria and classification probabilities are shown in Appendix (Table A2). The 3- and 4-class LPA models fit the data better than the 2-class model. Classification probabilities were better for 3 classes than for 4 and 5 classes. Figure 1 is the 3-class profile plot in which the bars reflect 95% confidence intervals for class centroids of standardized data quality indicators (*z* scores). Descriptive statistics and class comparisons for unstandardized data quality indicators and Healthy Pathways scales are shown in Table 3 and Table A3 (Appendix), respectively.

Over half of the sample was assigned to class 1 ($n = 998$, 56.1%). Class 1 children were consistent in their responses to items within scales as evidenced by their relatively small inter-item standard deviation and high within-person correlation on psychometric synonyms. The internal consistency of scales (α) ranged from 0.74 to 0.87 in class 1. For these children, response consistency may be partially attributable to their negatively skewed responses or tendency to use the “healthiest” response category. Children in class 1 had the highest Healthy Pathways scale scores (*M* range: 3.61–4.18) and the least response variation within scales (*SD* range: 0.57–0.69). Negative skew may also help explain the relative infrequency of Guttman errors, since no errors are registered when responses to item pairs are the same (e.g., endorsed “Always” for both items). Still, the relatively small average difference between negatively and positively worded items on the same scale (reverse-coded item inconsistency) suggests that children in class 1 understood the items and responded attentively. Children in class 1 had the highest person-total correlations and smallest Mahalanobis distance values, indicating that their response patterns were consistent with those of other children in sample.

Class 2 included 637 (35.8%) children. Compared to class 1, children in class 2 had more variation in their item responses. On average, they had relatively greater inter-item standard deviations (scale-level *SD* range: 0.64–0.85) and smaller within-person correlations on psychometric synonyms than class 1. Scale internal consistency (α) ranged from 0.65 to 0.85 in this class. Class 2 had more reverse-coded item inconsistency than class 1, but less than class 3. Similarly, their scores fell between those of children in classes 1 and 2 on person-total correlation, Mahalanobis distance, and polytomous Guttman errors.

Class 3 included 145 (8.1%) children. These children had the lowest Healthy Pathways scale scores (*M* range: 2.97–3.23) and the highest within-scale variation (*SD* range: 0.70–0.98).

As evidence by their high Mahalanobis distance and Guttman error values and low person-total correlation, the response patterns of children in class 3 deviated significantly from those of children in classes 1 and 2. With an average difference of nearly 1.5 response categories, children in class 3 had the largest discrepancy between positively and negatively worded items on the same scale. For children in class 3, internal consistency was below conventionally acceptable levels ($\alpha < 0.70$) for 3 of the 6 Healthy Pathways scales (range α: 0.57–0.82).

Compared to class 1, children in both classes 2 and 3 were significantly more likely to have ADHD, learning disability, speech impairment or delay, and all types of SHCNs. As shown in Table 4, children in class 3 were more likely to have neurodevelopmental conditions and SHCNs needs than children in class 2. Compared to class 1, average standardized tests scores were about 0.5 SDs lower in class 2 and 1.0 SDs lower in class 3. Similar trends were observed for parent-reported academic performance. On average, children in class 3 were about 4 months younger than those in classes 1 and 2. Class 3 also included a greater proportion of boys than the other classes.

Discussion

Despite recent advances in child PRO assessment, barriers to the widespread use of child-report measurement tools remain. Developmentally normative variation in the cognitive capacities needed to accurately self-report, which is especially pronounced in middle childhood, makes it difficult for end-users to determine whether they should administer PROMs to a child and once data are collected, whether they accurately reflect a child's true health state. Accommodative techniques, such as illustrated or multimodal administration of items, are intended to support children's self-report capacity. However, research on these techniques is limited by the lack of validated methods for detecting invalid PRO responses at the individual level. We conducted a systematic literature review to identify previously used methods for detecting invalid PRO response data. Thereafter, we explored associations among some of the identified procedures by applying them to existing child-report data.

There is limited consensus on the best approaches for evaluating the quality of PRO data quality at the individual level. In research involving both children and adults, nearly 60% of the identified data quality indicators were measures of unlikely virtues/social desirability and unlikely symptoms. In pediatric research, these measures were used exclusively with special populations (e.g., youth in residential treatment, juvenile detention, psychiatric inpatient settings) [17–19]. These measures assess *content responsive faking*, which assumes that respondents understand the items and alter their responses for gain (e.g., to avoid punishment). Such measures may be problematic for children, because they are often administered by people in positions of power. Even when PROMs have no direct bearing on children, they may misunderstand the potential consequences of the measures and intentionally alter their responses accordingly.

Self-reported measures of honesty, attention, effort, and/or interest were the second most frequently used data quality indicator overall, and the most commonly used strategy for children. Rather than administer unrealistic bogus items, Furlong et al. (2017) developed a

youth-report scale comprised of legitimate questions with both plausible and implausible response options. Adolescents who endorsed several low-incidence response options (e.g., having more than 10 siblings) were flagged as unreliable. Lengthening assessments with self-report indicators and bogus items may be impractical in some settings, especially in clinical care contexts. Mirroring Furlong et al.'s approach, a simple count of the number of times an individual endorses low-incidence response options on PRO items already being administered may provide important information about data quality without increasing assessment burden.

We identified 13 types of archival data quality indicators, 5 of which were previously applied in research involving children. Whereas prior studies used between 1 and 3 archival indicators to assess child-report data quality, we applied 8 archival techniques to child-report survey data. This enabled us to assess interrelationships among indicators and their relative utility. In prior studies involving children, the most commonly applied archival strategies were long-string, psychometric synonyms/antonyms, and univariate outlier statistics [19–21, 23]. We found these indicators to have limited utility compared to multivariate and IRT-based statistics. Longstring was uncorrelated with other indicators, and it failed to differentiate among data quality classes. In our general population sample, repeated endorsement of the same (usually most positive) response option may reflect children's true perceptions of their health, rather than lack of understanding or careless responding. Curran (2016) suggests that identification of extreme longstring outliers may be an efficient way to eliminate some of the "worst of the worst" responders (p. 8). However, in our sample, children in classes 1 and 3 were equally likely to have extreme long-string values (> 15 items). Resampled individual reliability was comparable across all three classes. Children in class 1 had higher psychometric synonym values than remaining children and there were no differences in psychometric synonyms between classes 2 and 3.

Classes were differentiated by inter-item SD, reverse-coded item inconsistency, Pearson-total correlation, Mahalanobis distance, and polytomous Guttman errors. As evidenced by inter-item SD, children in classes 2 and 3 had greater response variation than those in class 1. Poorer self-reported health among children in classes 2 and 3 may reflect their higher rates of special healthcare needs and neurodevelopmental conditions. This highlights a critical challenge in assessing the veracity of self-report data—distinguishing variable response patterns that reflect plausible symptom variation from those that indicate careless responding or possible limitations in self-report capacity. In our analyses, we found Mahalanobis distance and polytomous Guttman errors to be especially useful for differentiating plausible item response variation (class 2) from statistically unlikely response patterns (class 3). Given the extremely high correlation between these methods ($r = 0.95$), we recommend these approaches to identifying potentially invalid PROM data. The computational intensiveness of these methods may reduce their utility for some users. They can only be applied to large datasets a priori and, thus, may be most useful for health outcomes researchers. In contrast, they lack utility in clinical settings where the goal is to determine whether individual children can self-report prior to administering a full PROM battery. Reverse-coded item inconsistency is a simpler approach that also distinguished classes 2 and 3. Pediatric clinicians could use this method as a "quick check" of self-report data quality. However, because negatively worded items are inherently more difficult than

positively worded items, reverse-coded item inconsistency should be interpreted with caution, especially among youth with developmental or learning challenges.

We cannot definitively conclude that children in class 3 responded carelessly or lack self-report capacity. Notably however, conditions associated with a range of cognitive processing challenges that may compromise self-report validity were most common among children in class 3. These children were younger and significantly more likely to have SHCNs, ADHD, learning disability, and/or speech impairment or delay. Their poorer standardized test scores and parent-reported academic performance may reflect difficulty with attention, reading comprehension, sequencing, and other cognitive processes needed to reliably and accurately self-report. Rather than exclude these children from PRO assessments, assessment environments and measurement tools should be adapted to best support their self-report capacity [2]. Individual-level data quality indicators are essential for gauging the effectiveness of these accommodations.

In sum, the systematic literature review and secondary data analyses suggest that no single metric would definitively identify all problematic response patterns. Robust data screening practices should include multiple approaches to detect different types and patterns of potentially invalid data [5, 27]. This study has several notable limitations that highlight areas for future research. Our literature search was limited to 10 years of publication. Prior studies may describe additional approaches; however, redundancy in the identified methods suggests that we were able to establish a comprehensive list of data quality indicators. Direct data quality indicators, which were unavailable for these analyses, may have strengthened our capacity to identify problematic response patterns. In future research, archival and direct data quality indicators should be compared to more objective measures of inattentiveness (e.g., eye-tracking during PROM administration) and cognitive capacity (e.g., individually administered standardized tests, cognitive interviews that reveal children's understanding of PROM items) to establish cut points on quality indicators that signal the need for caution in interpreting PROMs. Simulation studies may also be helpful for identifying thresholds that indicate meaningful shifts in data quality. Finally, research on predictors of robust data quality indicators will inform the development of screeners that could be administered prior to PROM administration to identify children who may be unable to accurately self-report and those who would benefit from specific assessment accommodations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding Research reported in this publication was funded by the Office Of The Director, National Institutes Of Health (OD) under Award Number 4U24OD023319-02, with co-funding from the Office of Behavioral and Social Sciences Research (OBSSR) and by a grant from the National Institute of Child Health and Human Development (R01HD048850, PI Forrest). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Health.

References

1. Forrest CB, Bevans KB, Tucker C, Riley AW, Ravens-Sieberer U, Gardner W, et al. (2012). Commentary: The Patient-Reported Outcome Measurement Information System (PROMIS(R)) for children and youth: Application to pediatric psychology. Journal of Pediatric Psychology, 37, 614–621. 10.1093/jpepsy/jss038. [PubMed: 22362923]
2. Kramer JM, & Schwartz A (2017). Reducing barriers to patient-reported outcome measures for people with cognitive impairments. Archives of Physical Medicine and Rehabilitation, 98, 1705–1715. 10.1016/j.apmr.2017.03.011. [PubMed: 28400180]
3. Brod M, Tesler LE, & Christensen TL (2009). Qualitative research and content validity: Developing best practices based on science and experience. Quality of Life Research, 18, 1263–1278. 10.1007/s11136-009-9540-9. [PubMed: 19784865]
4. Lasch KE, Marquis P, Vigneux M, Abetz L, Arnould B, Bayliss M, et al. (2010). PRO development: Rigorous qualitative research as the crucial foundation. Quality of Life Research, 19, 1087–1096. 10.1007/s11136-010-9677-6. [PubMed: 20512662]
5. Curran PG (2016). Methods for the detection of carelessly invalid responses in survey data. Journal of Experimental Social Psychology, 66, 4–19. 10.1016/j.jesp.2015.07.006.
6. DeSimone JA, Harms PD, & DeSimone AJ (2015). Best practice recommendations for data screening: Data screening. Journal of Organizational Behavior, 36, 171–181. 10.1002/job.1962.
7. Johnson JA (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. Journal of Research in Personality, 39, 103–129. 10.1016/j.jrp.2004.09.009.
8. Meade AW, & Craig SB (2012). Identifying careless responses in survey data. Psychological Methods, 17, 437–455. 10.1037/a0028085. [PubMed: 22506584]
9. Huang JL, Bowling NA, Liu M, & Li Y (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. Journal of Business and Psychology, 30, 299–311. 10.1007/s10869-014-9357-6.
10. McKibben WB, & Silvia PJ (2017). Evaluating the distorting effects of inattentive responding and social desirability on self-report scales in creativity and the arts. The Journal of Creative Behavior, 51, 57–69. 10.1002/jocb.86.
11. Moher D, Liberati A, Tetzlaff J, Altman D, & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med, 6, e1000097. [PubMed: 19621072]
12. Cornell DG, Lovegrove PJ, & Baly MW (2014). Invalid survey response patterns among middle school students. Psychological Assessment, 26, 277–287. 10.1037/a0034808. [PubMed: 24219702]
13. Cornell D, Klein J, Konold T, & Huang F (2012). Effects of validity screening items on adolescent survey data. Psychological Assessment, 24, 21–35. 10.1037/a0024824. [PubMed: 21823800]
14. Furlong MJ, Fullchange A, & Dowdy E (2017). Effects of mischievous responding on universal mental health screening: I love rum raisin ice cream, really I do! School Psychology Quarterly, 32, 320–335. 10.1037/spq0000168. [PubMed: 27441548]
15. Jia Y, Konold TR, Cornell D, & Huang F (2018). The impact of validity screening on associations between self-reports of bullying victimization and student outcomes. Educational and Psychological Measurement, 78, 80–102. 10.1177/0013164416671767. [PubMed: 29795948]
16. Laajasalo T, Aronen ET, Saukkonen S, Salmi V, Aaltonen M, & Kivivuori J (2016). To tell or not to tell? Psychopathic traits and response integrity in youth delinquency surveys: Psychopathic traits and response integrity. Criminal Behaviour and Mental Health, 26, 81–93. 10.1002/cbm.1940. [PubMed: 25382604]
17. Barry CT, Lui JHL, & Anderson AC (2017). Adolescent narcissism, aggression, and prosocial behavior: The relevance of socially desirable responding. Journal of Personality Assessment, 99, 46–55. 10.1080/00223891.2016.1193812. [PubMed: 27362301]
18. Pechorro P, Ayala-Nunes L, Oliveira JP, Nunes C, & Gonqalves RA (2016). Psychometric properties of the socially desirable response set-5 among incarcerated male and female juvenile offenders. International Journal of Law and Psychiatry, 49, 17–21. 10.1016/j.ijlp.2016.05.003. [PubMed: 27210577]

19. Stokes J, Pogge D, Wecksell B, & Zaccario M (2011). Parent-child discrepancies in report of psychopathology: The contributions of response bias and parenting stress. Journal of Personality Assessment, 93, 527–536. 10.1080/00223891.2011.594131. [PubMed: 21859293]
20. Wardell JD, Rogers ML, Simms LJ, Jackson KM, & Read JP (2014). Point and click, carefully: Investigating inconsistent response styles in middle school and college students involved in web-based longitudinal substance use research. Assessment, 21, 427–442. 10.1177/1073191113505681. [PubMed: 24092819]
21. Boström P, Johnels JÅ, Thorson M, & Broberg M (2016). Subjective mental health, peer relations, family, and school environment in adolescents with intellectual developmental disorder: A first report of a new questionnaire administered on tablet PCs. Journal of Mental Health Research in Intellectual Disabilities, 9, 207–231. 10.1080/19315864.2016.1186254.
22. Hopfenbeck TN, & Maul A (2011). Examining evidence for the validity of pisa learning strategy scales based on student response processes. International Journal of Testing, 11, 95–121. 10.1080/15305058.2010.529977.
23. Cushing CC, Marker AM, Bejarano CM, Crick CJ, & Huffhines LP (2017). Latent variable mixture modeling of ecological momentary assessment data: Implications for screening and adolescent mood profiles. Journal of Child and Family Studies, 26, 1565–1572. 10.1007/s10826-017-0689-5.
24. Bevans KB, Riley AW, & Forrest CB (2010). Development of the healthy pathways child-report scales. Quality of Life Research, 19, 1195–1214. 10.1007/s11136-010-9687-4. [PubMed: 20563886]
25. Bethell CD, Read D, Stein REK, Blumberg SJ, Wells N, & Newacheck PW (2002). Identifying children with special health care needs: Development and evaluation of a short screening instrument. Ambulatory Pediatrics, 2, 38–48. [PubMed: 11888437]
26. Bevans KB, Riley AW, & Forrest CB (2012). Development of the healthy pathways parent-report scales. Quality of Life Research, 21, 1755–1770. 10.1007/s11136-012-0111-0. [PubMed: 22298201]
27. Huang JL, Curran PG, Keeney J, Poposki EM, & DeShon RP (2012). Detecting and deterring insufficient effort responding to surveys. Journal of Business and Psychology, 27, 99–114. 10.1007/s10869-011-9231-8.
28. Yentes R, Wilhelm F (2018) Careless: Procedures for computing indices of careless responding. R package version 1.1.3. 2018
29. Tendeiro J, Meijer R, & Niessen S (2016). PerFit: An R package for person-fit analysis in IRT. Journal of Statistical Software, 74, 1–27. 10.18637/jss.v074.i05.
30. Rosenberg J, Beymer P, Anderson D, & Schmidt J (2018). Tidylpa: An R Package to easily carry out latent profile analysis (lpa) using open-source or commercial software. Journal of Open Source Software, 3(30), 978 10.21105/joss.00978.
31. Celeux G, & Soromenho G (1996). An entropy criterion for assessing the number of clusters in a mixture model. Journal of Classification, 13, 195–212.
32. Marsh HW, Ludtke O, Trautwein U, & Morin AJ (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person-and variable centered approaches to theoretical models of self-concept. Structural Equation Modeling, 16, 191–225.
33. Tein JY, Coxe S, & Cham H (2013). Statistical power to detect the correct number of classes in latent profile analysis. Structural Equation Modeling, 20, 640–657. [PubMed: 24489457]

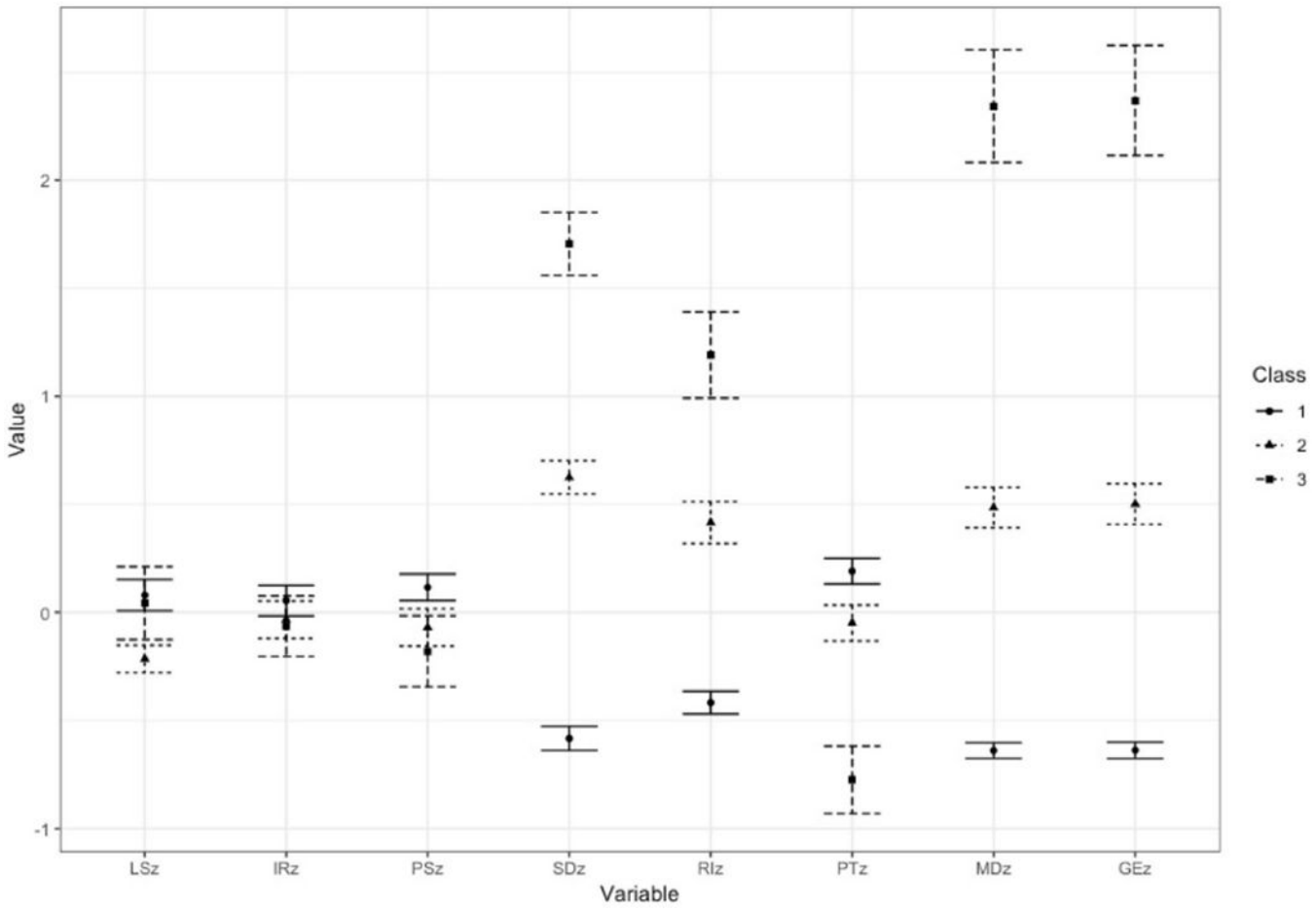

Fig. 1.

Standardized data quality metrics by LPA class. *LS* longstring, *IR* resampled individual reliability, *PS* psychometric synonyms, *SD* inter-item standard deviation, *RI* reverse-coded item inconsistency, *PT* person-total correlation, *MD* Mahalanobis distance, *GE* polytomous Guttman errors

Notes: LS = longstring, IR = resampled individual reliability, PS = psychometric synonyms, SD = inter-item standard deviation, RI = reverse-coded item inconsistency, PT = person-total correlation, MD = Mahalanobis distance, GE = polytomous Guttman errors

Table 1

Person-reported outcome data quality indicators

Data quality measures	Description	Abstracted from all studies		Abstracted from studies involving children	
		# of unique measures, *n*(%)[a]	Frequency of use, *n*(%)[b]	# of unique measures, *n*(%)[c]	Frequency of use, *n*(%)[d]
Direct indicators					
Bogus items	Questions with obvious answers. Designed with the assumptior that there is only one correct answer.	13 (9.8%)	22 (4.5%)	1 (11.1%)	1 (4.3%)
Instructed items	Items that direct the respondent to select a specific correct response. Instructions range from simple to complex.	11 (8.3%)	17 (3.4%)	–	–
Self-report indicators	Direct measures that ask respondents to rate their own participation in a survey (e.g., honesty, attention, diligence, effort, motivation, interest, seriousness, faking).	20 (15.0%)	42 (8.5%)	2 (22.2%)	7 (30.4%)
Content recall	Respondents asked to recall specific words, phrases, or concepts that were used in the survey.	2 (1.5%)	2 (0.4%)	–	–
Inconsistency with other measures of similar construct	Inconsistency between scores on self-report and other (often, more objective) measures of the same construct (e.g., physiological measure).	3 (2.3%)	3 (0.6%)	–	–
Observation of test-taking behavior	Observed demonstration of non-compliant or incompatible test-taking behaviors (e.g., skipping instructions, unfocused eye gaze).	5 (3.8%)	5 (1.0%)	1 (11.1%)	1 (4.3%)
Response time	The time it takes respondents to answer an item or set of items.	1 (0.8%)	19 (3.8%)	–	–
Unlikely virtues/social desirability	Items that assess the tendency to present oneself in an overly positive way.	51 (38.3%)	135 (27.3%)	3 (33.3%)	3 (13%)
Unlikely symptoms	Items that assess the tendency to present oneself in an overly negative way (e.g., over-reporting symptoms).	27 (20.3%)	80 (16.2%)	2 (22.2%)	2 (8.7%)
Archival indicators					
Skipped/omitted items	Number of items with missing response.	–	12 (2.4%)	–	–
Illogical responding	Responses across multiple items reflect an impossible, illogical, or highly implausible situation (e.g., “How old are you?” (10 years old) and “Do you have a driver’s license?” (yes)).	–	2 (0.4%)	–	–
Invariability (longstring)	The maximum number of consecutive items with the same response.	–	17 (3.4%)	–	2 (8.7%)
Synonyms/antonyms	Within-person correlations across item pairs with strong positive associations (synonyms) or strong inverse associations (antonyms). Pairs can be related semantically (identified before looking at response data) or psychometrically (identified by finding highly correlated item pairs in response data).	–	69 (14.0%)	–	3 (13%)
Direct item repetition	Inconsistency in response to the same (identical) item administered twice in the same survey.	–	2 (0.4%)	–	–

Data quality measures	Description	Abstracted from all studies		Abstracted from studies involving children	
		# of unique measures, *n*(%)[a]	Frequency of use, *n*(%)[b]	# of unique measures, *n*(%)[c]	Frequency of use, *n*(%)[d]
Even-odd consistency/ individual reliability	Within-person correlation across subscales formed by - even-odd split of a unidimensional scale or the average of within-person correlations derived from randomly formed split-halves resampled many times	–	12 (2.4%)	–	–
Inter-item standard deviation	Within-individual standard deviation; measures how much an - individual strays from their own personal midpoint across a set of scale items	–	3 (0.6%)	–	–
Person-total correlation	Correlation between a person's item responses and the mean of - all others' item responses	–	1 (0.2%)	–	–
Reverse-coded item inconsistency	The difference between a person's response to reverse-coded - items and non-reverse-coded items from the same unidimensional scale, (e.g., I am bored with my schoolwork; I am interested in my schoolwork)	–	4 (0.8%)	–	1 (4.3%)
Polytomous Guttman errors	The frequency with which a person endorses a higher response - category for a more difficult item than for a less difficult item. In its polytomous form, Guttman errors also account for the magnitude of this discrepancy (the distance between expected and actual response categories)	–	2 (0.4%)	–	–
Item response theory (IRT) person fit statistics	Methods for assessing the likelihood of an item-score vector - given known IRT-based item parameters, such as the relative difficulty of items	–	9 (1.8%)	–	1 (4.3%)
Univariate outlier statistics	Basic descriptive statistics used to characterize one value from - a distribution relative to other values in that distribution (e.g., a value that exists 1.5 standard deviations from the mean)	–	23 (4.7%)	–	2 (8.7%)
Mahalanobis distance	Assesses statistically unlikely response patterns. Multivariate - distance between response vector and the vector of the sample mean. Unlike univariate outlier analyses, Mahalanobis distance considers the patterns of responses across a series of items	–	13 (2.6%)	–	–

[a] Percentage of the 133 unique direct quality indicators

[b] Percentage of the 494 observed instances in which a data quality measure was applied

[c] Percentage of the 9 unique direct quality indicators

[d] Percentage of the 23 observed instances in which a data quality measure was applied

Table 2

Archival data quality measures descriptive statistics and intercorrelations

Measure	*M* (*SD*)	Range	Intercorrelations (Pearson *r*)						
			LS	IR	PS	SD	RS	PT	MD
Longstring (LS)	5.30 (2.55)	2–21	–						
Resampled individual reliability (IR)	0.42 (0.31)	− 0.88–1.00	0.20	–					
Psychometric synonyms (PS)	0.42 (0.27)	− 0.47–1.00	0.12	0.50	–				
Inter-item standard deviation (SD)	0.93 (0.23)	0.27–1.70	− 0.28	− 0.16	− 0.05	–			
Reverse-coded item inconsistency (RI)	0.90 (0.46)	0.00–3.06	0.06	− 0.05	− 0.01	0.57	–		
Person-total correlation (PT)	0.42 (0.16)	− 0.07–0.82	− 0.09	− 0.05	0.32	0.05	− 0.04	–	
Mahalanobis distance (MD)	48.35 (24.65)	8.62–184.25	− 0.09	0.01	− 0.12	0.84	0.46	− 0.32	–
Polytomous Guttman errors (GE)	116.40 (68.08)	1–492	− 0.10	− 0.08	− 0.12	0.85	0.52	− 0.33	0.95

Qual Life Res. Author manuscript; available in PMC 2021 May 01.

Table 3

Unstandardized data quality indicators: descriptive statistics by class

	Longstring (LS)			Resampled Individual reliability (IR)			Psychometric synonyms (PS)		
	M	**SD**	**Range**	**M**	**SD**	**Range**	**M**	**SD**	**Range**
Class 1	5.58	2.67	2–18	0.43	0.32	− 0.86–1.00	0.44	0.27	− 0.44–1.00
Class 2	4.80	2.18	2–17	0.41	0.3	− 0.88–0.89	0.39	0.27	− 0.35–0.97
Class 3	5.48	2.89	2–21	0.40	0.28	− 0.35–0.83	0.36	0.26	− 0.47–0.88
F (post hoc)	13.78 *** (1 > 2)			2.83			21.09 *** (1 > 2,3)		
	Inter-item standard deviation (SD)			**Reverse-coded item inconsistency (RI)**			**Person-total correlation (PT)**		
	M	**SD**	**Range**	**M**	**SD**	**Range**	**M**	**SD**	**Range**
Class 1	0.77	0.14	0.27–1.13	0.70	0.32	0.04–1.95	0.45	0.15	− 0.03–0.82
Class 2	1.08	0.13	0.68–1.45	1.09	0.46	0.00–3.06	0.41	0.16	− 0.07–0.76
Class 3	1.35	0.13	1.07–1.70	1.44	0.51	0.13–2.77	0.29	0.15	− 0.06–0.69
F (post hoc)	3344 **** (1 < 2 < 3)			697.10 *** (1 < 2 < 3)			121.40 *** (1 > 2 > 3)		
	Mahalanobis distance (MD)			**Polytomous Guttman errors (GE)**					
	M	**SD**	**Range**	**M**	**SD**	**Range**			
Class 1	32.17	8.95	8.62–62.93	71.33	24.03	1–149			
Class 2	60.44	11.88	36.30–96.98	150.27	31.79	88–257			
Class 3	106.62	21.23	70.07–184.25	277.79	55.46	191–492			
F (post hoc)	5700 **** (1 < 2 < 3)			6034 **** (1 < 2 < 3)					

Post hoc class comparisons with Bonferroni correction,

*** $p < 0.001$,

**** $p < 0.0001$

Table 4

Child age, gender, health status, and school performance by data quality indicator classes

	All	Data quality indicator classes		
		Class 1	Class 2	Class 3
Child age,[a] *M*(*SD*)	9.6 (1.0)	9.6 (0.9)	9.6 (1.0)	9.3 (0.9)**
Male gender,[a] (ref = girls), %	48.2	46.6	47.3	63.2****
Special healthcare needs (≥ 1),[b] %	35.5	29.5	40.2***	57.4****
Medication use	28.6	24.8	31.1*	45.2****
Healthcare or educational services	14.3	11.0	16.5**	28.9****
Functional limitations	7.0	3.8	10.3****	15.4****
Special therapy (e.g., PT, OT, speech)	3.1	1.3	4.3**	10.5****
Developmental/behavioral treatment or counseling	12.9	9.6	16.1**	22.9****
Chronic health or neurodevelopmental condition (≥ 1),[b] %	25.5	20.7	28.7**	45.2****
Asthma	11.2	10.1	11.6	16.8*
Epilepsy	0.5	0.8	0.0	0.0
Attention deficit hyperactivity disorder	12.0	9.1	14.3**	22.4****
Learning disability	6.8	3.7	9.0***	20.0****
Speech impairment or delay	2.5	1.4	3.6*	7.5***
School performance, *M*(*SD*)				
Reading standardized test score[c]	100.4 (14.9)	104.1 (14.1)	96.5 (14.5)****	91.4 (14.3)****
Math standardized test score[c]	100.7 (14.6)	104.3 (13.5)	97.4 (14.3)****	90.7 (15.7)****
Parent-reported academic performance[b]	3.8 (0.9)	4.0 (0.9)	3.6 (0.9)***	3.3 (0.9)****

For class analyses, class 1 was the reference category

[a]Based on child-report: All (n = 1780), class 1 (n = 998), class 2 (n = 637), class 3 (n = 145)

[b]Parent-report: All (n = 1221), class 1 (n = 708), class 2 (n = 413), class 3 (n = 100)

[c]school records: All (n = 1764), class 1 (n = 990), class 2 (n = 631), class 3 (n = 143)

*$p < 0.05$

***$p < 0.001$

****$p < 0.0001$