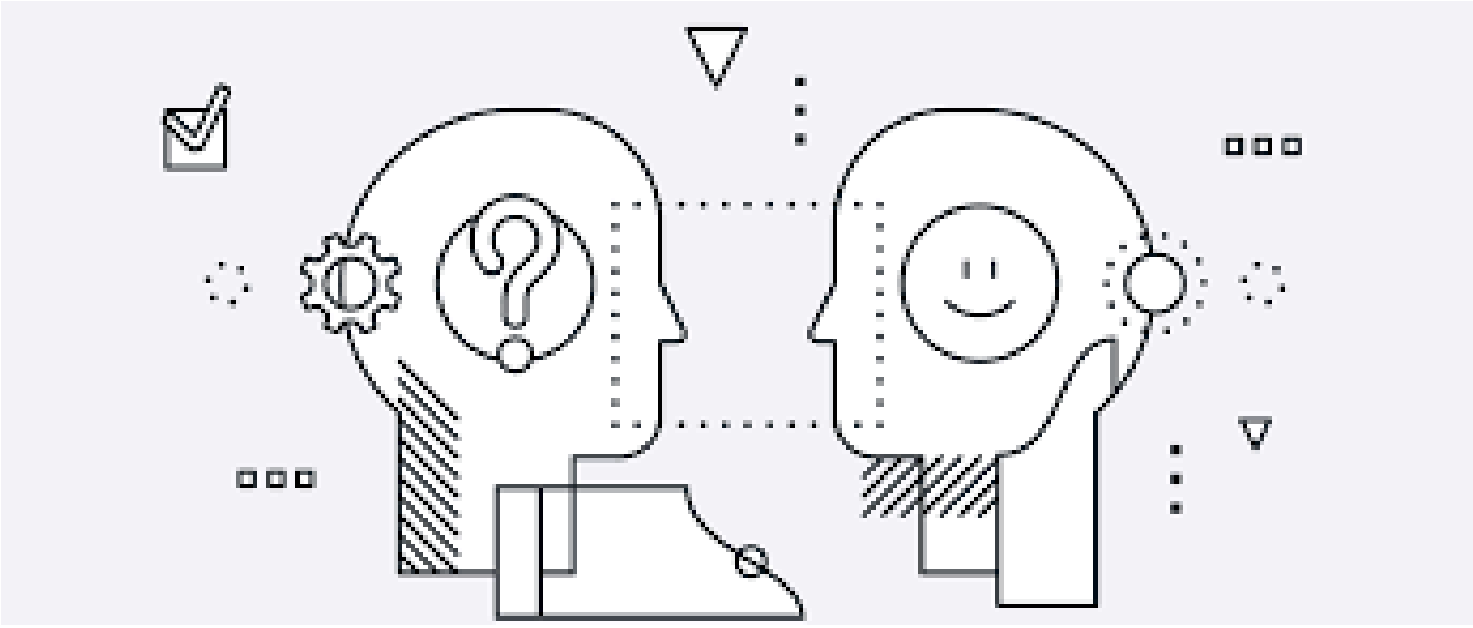
Review of Psychometric Testing of the National Core Indicators – Intellectual and Developmental Disabilities In-Person Survey



Contents

Acronyms and Abbreviations

ADL	Activities of Daily Living
ANOVA	Analysis of Variance
BI	Background Information (section of the NCI-IDD IPS)
CC	Choice and Control
CFA	Confirmatory Factor Analysis
CI	Community Inclusion
CMS	Centers for Medicare & Medicaid Services
DD	Developmental Disabilities
DIF	Differential Item Functioning
FMS	Full Measure Submission
HCBS	Home and Community-Based Services
HLR	Human and Legal Rights
HSRI	Human Services Research Institute
IDD	Intellectual and Developmental Disabilities
IDM	Instrument Derived Measure
IPS	In-Person Survey
IRR	Inter-Rater Reliability
IRT	Item Response Theory
IUR	Inter-Unit Reliability
KFF	Kaiser Family Foundation
LTSS	Long-Term Services and Supports
MCO	Managed Care Organization
NASDDDS	National Association of State Directors of Developmental Disabilities Services
NCI	National Core Indicators
NCI-IDD	National Core Indicators — Intellectual and Developmental Disabilities
NHIS	National Health Interview Survey
NQF	National Quality Forum
ODESA	Online Data Entry System and Analysis
PCA	Principal Component Analysis
PCP	Person-Centered Planning
PQM	Partnership for Quality Measurement
SIB	Self-Injurious Behavior

PAGE INTENTIONALLY LEFT BLANK

Introduction

This document reviews the validity, reliability, and overall measurement quality of the National Core Indicators – Intellectual and Developmental Disabilities (NCI-IDD) In-Person Survey (IPS). The IPS has served as a cornerstone instrument for evaluating state developmental disability service systems and outcomes for adults with IDD since the late 1990s. Given how the data are used in policy, quality improvement, and research, regular evaluation of the instrument's psychometric properties is essential.

The review synthesizes evidence from internal validation studies, external peer-reviewed publications, and ongoing quality assurance activities. To keep the focus on results, descriptive material about the survey itself, its purpose, instrument, data collection, and data validation, has been moved to Appendix A. Readers seeking that context should consult the appendix; the main text begins with a brief framing of validity and reliability and proceeds directly to the psychometric evidence.

1. Brief Introduction to Validity and Reliability

Understanding Psychometrics

Psychometrics is the scientific field concerned with measuring psychological attributes, behaviors, and outcomes. In simple terms, psychometrics helps determine whether a measurement tool, such as a survey or assessment, actually measures what it claims to measure and does so consistently. For the NCI-IDD IPS, psychometrics ensures that questions about community inclusion, choice and control, or service satisfaction truly capture these aspects of people's lives in a reliable and accurate way.

Validity: Does the Survey Measure What It Claims to Measure?

Validity refers to the accuracy of measurement, whether a survey actually measures the construct it is intended to measure. For the NCI-IDD IPS, validity evidence helps us understand whether questions about community participation reflect real community inclusion, or whether satisfaction ratings truly capture people's feelings about their services.

Content Validity

Content validity examines whether survey questions adequately cover all important aspects of what we want to measure. For example, if we want to measure community inclusion, content validity assesses whether questions cover the full range of community activities and experiences important for people with IDD. Content validity is typically supported by stakeholder review and cognitive testing of items with people from the target population.

Construct Validity

Construct validity evaluates whether the survey measures the theoretical concepts ("constructs") it is designed to assess. The two main statistical tools used to support construct validity in NCI work are correlation analysis and factor analysis. Correlation analysis examines whether items related in theory also correlate empirically. Factor analysis (exploratory and confirmatory) examines whether items group together in ways consistent with theoretical constructs such as self-determination, community inclusion, or quality of life.

Face Validity

Face validity addresses whether measures appear appropriate and meaningful to users, including people with IDD, family members, advocates, providers, and policymakers. Face

validity is supported when items are developed with stakeholder input and when the resulting measures are recognized as relevant indicators of service quality. Inclusion of NCI-IDD measures in peer-reviewed literature also provides ongoing evidence of face validity, as the scientific community treats them as valid representations of the constructs they purport to assess.

Criterion Validity

Criterion validity assesses how well survey results relate to other measures or outcomes that should theoretically be connected. With NCI data, this is typically pursued by examining correlates and contributors, for example, whether people who report higher levels of choice and control also show higher scores on related self-determination indicators, or whether state-level scores align with other indicators of system performance, rather than through strict criterion validation against a gold standard.

Reliability: Does the Survey Produce Consistent Results?

Reliability refers to the consistency of measurement results. A reliable instrument produces similar results when measuring the same thing under similar conditions. In practical terms, reliability addresses questions like: “Would different interviewers get similar results when surveying the same person?” or “Would the same person give similar responses if asked the same questions at different times when their circumstances haven’t changed?”

Inter-Rater Reliability

Inter-rater reliability measures the degree to which different surveyors obtain consistent results when assessing the same individuals. This is particularly important for the NCI-IDD IPS because data collection involves multiple surveyors across different states and settings. Inter-rater reliability is typically tested using *agreement statistics* such as percent agreement and Cohen’s kappa, which corrects for chance agreement. Following Landis and Koch (1977), kappa values of 0.61–0.80 indicate substantial agreement and values above 0.80 indicate almost perfect agreement.

Internal Consistency Reliability

Internal consistency reliability examines whether items intended to measure the same concept produce consistent results. It is most often quantified using Cronbach’s alpha and the Spearman-Brown coefficient, with corrected item-total correlations used to evaluate the contribution of each item. Inter-item correlation analysis (typically Pearson correlations) supports this work by identifying items that hang together as expected.

Test-Retest Reliability

Test-retest reliability evaluates whether the survey produces stable results over time when the underlying conditions being measured have not changed. It is assessed through correlation analysis of repeated measurements with the same respondents.

Why Validity and Reliability Matter

Validity and reliability work together as fundamental quality indicators. Valid measures ensure that data reflect what researchers claim to be studying; reliable measures ensure that those data are consistent across administrations. Together, they establish the credibility of the instrument and the data it generates, supporting appropriate statistical inference and meaningful generalization beyond the specific sample studied. For an instrument like the NCI-IDD IPS that informs state-level policy and resource allocation, this foundation is essential.

Validity and Reliability in the Context of IDD

Research with individuals with IDD presents distinct challenges. Cognitive and communication diversity, the frequent necessity of proxy reporting, and the potential for response biases such as acquiescence can all complicate measurement. These challenges are not insurmountable: a person-centered, flexible, and culturally sensitive approach, supported by cognitive testing, carefully designed proxy protocols, and surveyor training, can produce high-quality data that reflects the experiences of individuals with IDD. The psychometric evidence for the NCI-IDD IPS, summarized in the next section, has been built using exactly this kind of approach.

2. Psychometric Analyses of the NCI-IDD IPS

Psychometric evaluation of the NCI-IDD IPS reflects an ongoing commitment to scientific rigor and continuous quality improvement. The testing program has integrated multiple types of evidence, inter-rater reliability studies, scale development analyses, cognitive testing, pilot studies of new modalities, and external review for measure endorsement, to provide a comprehensive picture of the instrument's measurement properties.

This section organizes the evidence into three groups: (1) reliability and validity work conducted during survey development and refinement; (2) the psychometric work conducted for National Quality Forum (NQF) endorsement of 14 NCI-IDD measures in 2022; and (3) other published psychometric studies. A summary table at the end of the section consolidates the key findings across studies.

2.1 Validity and Reliability Tests During Survey Development

2.1.1 Inter-Rater Reliability Tests

Inter-rater reliability testing has been a cornerstone of NCI-IDD IPS evaluation since the survey's inception. These studies are commonly summarized using Cohen's kappa, which corrects for agreement expected by chance. Following Landis and Koch (1977), values above 0.60 are considered substantial and values above 0.80 are considered almost perfect. The studies described below, conducted between 1997 and 2010 and documented by Smith and Ashbaugh (2001) and through subsequent NCI internal reports, provide an empirical foundation for the survey's administration. Detailed numeric results for each study are shown in the summary table at the end of this section.

1997 Pilot Test (n = 30)

Two trained surveyors independently interviewed the same 30 adults with IDD in a member state. Inter-rater agreement averaged 93% across all items, strong initial evidence that the instrument could be administered consistently across surveyors and an enabler of broader multi-site rollout.

1998 Inter-Rater Reliability Study (n = 25)

Building on the 1997 pilot, a more rigorous paired-interviewer study with enhanced training again achieved 93% agreement and an average kappa of 0.794, substantial agreement under the Landis-Koch convention. The consistency between 1997 and 1998 results indicated that

high reliability was a property of the instrument and procedures, not an artifact of a single sample.

1999 Reliability Test (n = 27)

The 1999 paired-interviewer study found 92% agreement, maintaining the high reliability observed in earlier rounds. The cumulative pattern across three studies provided strong evidence that data could be collected reliably across an expanding network of states.

2008 Post-Revision Reliability Testing (n = 16)

Following significant 2008 revisions to question wording and content, a paired-interviewer study reported an average kappa of 0.90 across rater pairs, in the “almost perfect” range. Although the sample was smaller than in earlier studies, it was sufficient to confirm that revisions had not compromised, and may have enhanced, reliability.

2010 Shadowing Study (six surveyors, 20 interviews)

The 2010 study used a shadowing methodology in which six trained state surveyors conducted 20 interviews while NCI team members independently recorded responses. This approach reduced participant burden and reflected real-world field conditions. Kappa scores ranged from 0.82 to 0.95 (average 0.89), with overall percent agreement of 80%, supporting confidence in the quality of routine data collection.

Background Section Inter-Rater Reliability

The Background Information (BI) section, which collects demographic and service-related information from administrative records and proxy respondents, has also been evaluated. A dedicated study by the University of Minnesota found that BI data extracted from state administrative sources demonstrated reliability rates of 88–96% across participating states, with items related to employment, volunteering, managed care plans, funding sources, and specific health conditions showing particularly strong reliability (Tichá, 2017).

2.1.2 Routine Tests for Multi-Item Scale Development

The IPS includes several multi-item scales designed to measure complex constructs that cannot be assessed through single questions. Their development follows established psychometric procedures adapted for the IDD population.

Inter-Item Correlations

Pearson correlations between all pairs of items within a proposed scale are calculated and inspected. Items with weak correlations ($r < 0.30$) with other scale items are flagged as candidates for revision or relocation. Moderate correlations ($r = 0.30–0.70$) are considered optimal: items measure related aspects of the same construct without being redundant. Inter-item correlation analyses across NCI-IDD scales have generally supported the conceptual organization of items into their intended scales, with most items showing the expected pattern of relationships.

Exploratory Factor Analysis (PCA)

Principal Component Analysis (PCA) is used to examine the underlying structure of items and validate the conceptual organization of scales. Factors with eigenvalues above 1.0 are typically retained, with scree plot inspection and parallel analysis used as additional criteria. Varimax rotation is commonly applied to enhance interpretability. Loadings above 0.40 are considered meaningful. Across NCI-IDD scale development work, PCA results have generally aligned with theoretical expectations, with items loading on their intended factors and the extracted factors corresponding to constructs such as community inclusion, self-determination, and quality of life.

2.1.3 Cognitive Pretest of Newly Added Person-Centered Planning (PCP) Items

In 2018–19, in connection with the addition of Person-Centered Planning (PCP) questions to the survey, a cognitive test was conducted with 10 IDD service users from member states using a 35-item test instrument. Cognitive testing involves administering survey items to individuals from the target population along with follow-up probes to assess comprehension and interpretation. The test results provided information about how well typical respondents understood the new PCP questions, whether the intent of each question was interpreted similarly across respondents, and whether respondents could provide consistent and relevant responses.

Following exploratory analysis, principal component analysis was used to construct multi-item scales, and confirmatory factor analysis (CFA) tested the joint factor structure. Internal consistency was evaluated using Cronbach's alpha and the Spearman-Brown coefficient, supplemented by corrected item-total correlations. Five multi-item scales were developed:

- CC-4: Life Decisions (Choice and Control)
- HLR-1: Respect for Personal Space (Human and Legal Rights)
- CI-3: Transportation (Community Inclusion)
- CI-4: Community Inclusion Activities
- PCP-5: Satisfaction with Community Inclusion Participation

Cognitive testing confirmed that the new PCP questions were generally well understood by the target population. Minor revisions to wording were made based on participant feedback to enhance clarity and consistent interpretation.

2.1.4 Pilot Study of Remote (Videoconference) Administration

During the COVID-19 pandemic, the NCI team piloted survey administration via videoconference to test the feasibility and equivalence of remote data collection. A subset of participants completed the survey via videoconference while a matched comparison group completed traditional in-person surveys. Preliminary results indicated no significant differences in response patterns between videoconference and in-person administration for most sections, supporting the equivalence of the two modes. The pilot has important implications for reaching

individuals in rural areas, those with transportation barriers, and during public health emergencies.

2.1.5 Non-Responder (Selection Bias) Analysis

In 2020, the NCI team conducted a non-responder study with two member states, comparing characteristics of respondents to sample individuals who did not respond during the 2018–19 survey cycle. Background information for non-responders was provided by states from administrative records. Because most non-respondents were people who declined participation rather than who agreed and then dropped out, this analysis is best described as an assessment of selection bias rather than attrition bias. The study included 429 respondents and 363 non-respondents from State 1, and 2,137 respondents and 1,702 non-respondents from State 2.

State 1 achieved a 54% response rate; among non-responses with documented reasons (n = 247), 57% could not be reached, 41% were caregiver refusals, and 2% were self-refusals. State 2 achieved a 56% response rate; among 1,702 non-responses, 60% could not be reached, 26% were refusals by others, 9% were self-refusals, and 9% were cancellations or no-shows after initial agreement.

State-specific patterns. In State 1, responders were significantly older (44 vs. 35 years), more likely to have mild or moderate intellectual disability (70% vs. 55%), more likely to use spoken communication (80% vs. 74%), more likely to live in group settings (36% vs. 6%) or foster homes (15% vs. 3%) than in parents' homes (35% vs. 77%), more likely to have guardians (76% vs. 71%), more likely to receive behavioral supports (53% vs. 42%), and less likely to use self-directed supports (22% vs. 63%). No significant differences were found by gender, race, or funding source. In State 2, responders were significantly older (44 vs. 39 years), more likely to live in group settings (54% vs. 39%) than in parents' homes (25% vs. 36%), more likely to have Medicaid HCBS funding (92% vs. 84%), and less likely to use self-directed supports (6% vs. 11%). No significant differences were found in gender, race, level of intellectual disability, communication mode, or legal status.

Across both states, inability to contact sampled individuals was the leading cause of non-response. Older people and those living in group settings were consistently overrepresented; people in parents' homes and those receiving self-directed supports were underrepresented. These findings argue for maintaining up-to-date contact information, intentional outreach to younger and more independent service users, and the use of videoconference administration in rural and hard-to-reach areas.

2.2 Testing for NQF Measure Endorsement

In 2022, the National Quality Forum (NQF) endorsed 14 NCI-IDD measures for assessing the quality of long-term services and supports for people with IDD. Endorsement followed scientific review by the NQF Scientific Methods Panel, consensus panel analysis, and a public comment

period — representing significant external validation of the IPS's scientific rigor and utility for quality measurement.

The 14 NQF-Endorsed Measures

The endorsed measures span four domains. Person-Centered Planning and Coordination (5 measures): PCP-1 Employment Goal in Service Plan; PCP-2 Service Plan Includes Important Things; PCP-3 Independence Goal in Service Plan; PCP-4 Supported to Learn New Things; PCP-5 Satisfaction with Community Inclusion. Community Inclusion (4 measures): CI-1 Does Not Feel Lonely Often; CI-2 Has Friends Outside Staff/Family; CI-3 Adequate Transportation; CI-4 Engages in Activities Outside Home. Choice and Control (4 measures): CC-1 Chose or Could Request to Change Staff; CC-2 Could Change Case Manager; CC-3 Can Stay Home When Others Go Out; CC-4 Makes Choices in Life Decisions. Human and Legal Rights (1 measure): HLR-1 Personal Space Respected.

Reliability Testing for NQF Endorsement

Background Information data abstracted from state administrative records were tested for inter-abstractor reliability by having multiple individuals abstract the same records. The remaining endorsed measures draw primarily on items from the Community Inclusion, Choices, and Rights sections, which demonstrated 90% or higher agreement in the 2010 inter-rater reliability study, providing strong confidence in the quality of data underlying the endorsed measures.

Performance Measure Score Reliability

Analysis of Variance (ANOVA). ANOVA tested whether between-state variance significantly exceeded within-state variance for each measure. Between-state variation was significantly larger than within-state variation for all 14 measures ($p < 0.001$), indicating that the measures successfully differentiate among state service systems.

Inter-Unit Reliability (IUR). IUR was calculated for each measure as a complementary reliability metric, expressing the proportion of between-state variance remaining after removing within-state variance. The IUR for the 14 measures ranged from 0.75 to 0.98. A minimum IUR of 0.7 is recommended as acceptable for clinical performance measures; all NCI measures exceeded that threshold, and for several measures (CC-1 Chose Staff, CC-4 Life Decisions Scale, PCP-5 Satisfaction with Community Inclusion Scale) practically the entire variation was between states, with IUR values approaching 1.0.

In interpreting IUR values, it is important to note that the metric was originally developed for clinical performance measurement, where scores reflect substantial variability in social determinants of health across measurement units. For HCBS measures with states as the unit of analysis, legal and policy differences across states contribute to within-state homogeneity and thereby raise the IUR. The high IUR values therefore underscore the policy relevance of these measures: differences in scores are primarily driven by state policies, practices, and service-system characteristics, making the measures particularly useful for guiding system-level quality improvement.

Validity Evidence for NQF Endorsement

Construct validity. The measures were evaluated against their intended constructs (e.g., person-centered planning, community inclusion, choice and control, rights). Evidence came from the theoretical framework underlying measure development, the empirical clustering of items as expected, and patterns of correlations among measures aligned with theoretical predictions.

Face validity. The IPS was developed with extensive user input, and the endorsed measures reflect outcomes that users, people with IDD, family members, advocates, providers, and policymakers identified as important indicators of quality services.

Criterion validity. State-level analyses demonstrated that the measures distinguish among states' performance in ways that align with other indicators of service-system quality. States that perform well on NCI measures tend to have policies and practices recognized as supporting person-centered, community-integrated services.

NQF endorsement is not a one-time event; the NCI program is committed to ongoing psychometric monitoring and periodic revalidation as new data become available. The 2022 endorsement positions NCI measures for potential use in federal quality reporting requirements, value-based payment programs, and other accountability initiatives, and validates the broader IPS instrument and the stakeholder-driven development process behind it.

2.3 Ongoing Performance Measure Testing (2024–25 Cycle)

As part of the maintenance and continued endorsement of NCI-IDD measures, the team conducted accountable-entity-level reliability and validity testing for a set of derived measures using 2024–25 NCI-IDD IPS data. This section summarizes the testing methodology and results for four representative measures, ADL Goal (PCP-3), Satisfaction with Community Inclusion (PCP-5), Social Connectedness (CI-1), and Has Friends (CI-2), prepared for the Spring 2026 PQM submission cycle. These analyses follow the inter-unit reliability (IUR) approach described in Section 2.2 and extend the evidence base with current data and convergent validity testing against hypothesized correlates.

2.3.1 Methodology

Reliability. For each measure, the NCI team fit an intercept-only hierarchical logistic model (linear for the multi-item scale) with the measure as the dependent variable and state as the level-2 random effect, partitioning variance into within- and between-state components (He et al., 2019). The IUR was calculated as the proportion of total variance attributable to differences between states, providing a signal-to-noise ratio for state-level performance comparisons. States with sample sizes below 20 were excluded from analysis. Decile-stratified IURs were also reported (Tables 2a and 2b of each FMS submission) to assess reliability across the distribution of state sample sizes and reliability scores.

Validity. Convergent and discriminant validity were assessed by correlating each measure with hypothesized state-level correlates drawn from the NCI-IPS and external sources (e.g., the KFF State Health Facts Database for HCBS spending). Hypotheses specified the expected direction of association, the expected strength of the association, and the theoretical rationale linking the correlate to the measure. Pearson or Spearman rank correlations were reported depending on distributional properties, with one-tailed significance tests aligned with the directional hypotheses.

2.3.2 Performance Score Distributions

Table 2 summarizes the overall performance scores and the range of state-level scores across the four measures. Performance gaps are substantial for each measure, indicating meaningful room for quality improvement across participating states.

Table 2. Performance Score Distributions, 2024–25 NCI-IDD IPS

Measure	Domain	Overall Mean	Min (State)	Max (State)	States / Respondents
ADL Goal (PCP-3)	PCP	76.7%	16.2%	96.6%	34 states / 1,838 respondents
Satisfaction with Community Inclusion (PCP-5)	PCP	60.6%	45%	86%	39 states (multi-item scale)
Social Connectedness (CI-1)	CI	87.3%	81%	98%	39 states
Has Friends (CI-2)	CI	75.6%	44%	91%	39 states

Note. Performance gap analyses reveal that for ADL Goal (PCP-3) half of the participating states score below 80%, meaning at least 20% of participants who need help with ADLs and want to learn to perform them more independently do not have this goal in their service plan. For Satisfaction with Community Inclusion (PCP-5), 39% of survey participants reported they would like to either increase or decrease their current level of community participation. For Has Friends (CI-2), the 47-point spread between states (44% to 91%) underscores the considerable variability in providers' efforts to support clients in establishing relationships outside the immediate circle of family and paid staff.

2.3.3 Inter-Unit Reliability (IUR) Results

Table 3 reports the overall IUR for each of the four measures and the range of IURs across deciles of state sample size. All four measures meet or exceed the 0.7 IUR threshold recommended for accountable-entity-level performance measurement, indicating that observed differences in state scores reflect genuine differences in service-system performance rather than random measurement noise. The multi-item Satisfaction with Community Inclusion scale shows the highest overall IUR (0.944), reflecting the additional precision afforded by multiple-

item measurement. The single-item Social Connectedness measure shows the lowest overall IUR (0.774), with the smallest-sample decile dipping below threshold — a pattern consistent with the lower precision of binary single-item measures in small-sample states.

Table 3. Inter-Unit Reliability (IUR) Results, 2024–25 NCI-IDD IPS

Measure	Overall IUR	Min IUR (lowest decile)	Max IUR (highest decile)	Interpretation
ADL Goal (PCP-3)	0.882	0.729	0.961	All deciles above 0.7 threshold
Satisfaction with Community Inclusion (PCP-5)	0.944	0.915	0.969	All deciles approach 0.95; very strong signal
Social Connectedness (CI-1)	0.774	0.380	0.978	Above threshold overall; lowest decile below threshold
Has Friends (CI-2)	0.952	0.784	0.996	All deciles above 0.78; very strong signal

Note. IUR represents the proportion of total variance attributable to differences between states (signal) versus within-state variation (noise). Values ≥ 0.7 are conventionally considered acceptable for accountable-entity-level performance measurement (He et al., 2019). For Satisfaction with Community Inclusion, an IUR > 0.9 may in clinical contexts indicate unaddressed case-mix differences (Hartman et al., 2024); however, where the unit of analysis is a state agency operating within state-specific policy and funding environments, as is the case here, high IUR more plausibly reflects substantive policy-driven differences across states rather than case-mix confounding.

2.3.4 Convergent and Discriminant Validity

Each measure was tested against a set of hypothesized correlates drawn from the NCI-IPS and external sources. Table 4 summarizes the strongest correlations observed for each measure, along with the direction predicted by the underlying theoretical model. Patterns of association generally supported the hypothesized construct relationships.

Table 4. Validity Correlations with Hypothesized Correlates, 2024–25 NCI-IDD IPS

Measure	Hypothesized Correlate	N States	Correlation (p)	Expected / Direction
Satisfaction with Community Inclusion (PCP-5)	Free Time Choice (most proximate)	39	0.411 (Spearman, p = 0.006)	Strongest — supported

Measure	Hypothesized Correlate	N States	Correlation (p)	Expected / Direction
Satisfaction with Community Inclusion (PCP-5)	Free Time Decisions	39	0.442 (Spearman, p = 0.002)	Strong — supported
Satisfaction with Community Inclusion (PCP-5)	Community Inclusion Scale (discriminant)	39	0.114 (Spearman, p = 0.099)	Weak — supported (discriminant)
Satisfaction with Community Inclusion (PCP-5)	HCBS Spending per enrollee	26	0.254 (Pearson, p = 0.105)	Positive — directionally supported
Social Connectedness (CI-1)	Satisfaction with Community Inclusion (proximate)	39	0.692 (Pearson, p < 0.001)	Strongest — supported
Social Connectedness (CI-1)	Has Friends	39	0.201 (Pearson, p = 0.110)	Positive — directionally supported
Social Connectedness (CI-1)	Community Inclusion Scale	39	0.156 (Pearson, p = 0.172)	Positive — directionally supported
Has Friends (CI-2)	Often Lonely (inverse)	39	−0.201 (Pearson, p = 0.110)	Negative — supported
Has Friends (CI-2)	Community Inclusion Scale	39	0.246 (Spearman, p = 0.065)	Positive — supported
Has Friends (CI-2)	Per Capita HCBS Funding	26	0.127 (Pearson, p = 0.268)	Positive — directionally supported
Has Friends (CI-2)	Ages 65+ (discriminant)	39	−0.149 (Spearman, p = 0.182)	Negative — supported (discriminant)

Note. Correlations reported using either Pearson or Spearman rank coefficients based on the distributional properties of each correlate. One-tailed significance tests were used, consistent with the directional hypotheses. State-level N is limited to 26–39 depending on the correlate (HCBS Spending data is available for fewer states), which reduces statistical power for detecting significant associations; the directional consistency of the observed correlations, however, supports the construct validity of each measure.

2.3.5 Interpretation

Reliability. All four measures meet the 0.7 IUR threshold at the overall level, with three of the four exceeding 0.88. The Satisfaction with Community Inclusion scale (IUR = 0.944) and Has

Friends (IUR = 0.952) show particularly strong signal-to-noise ratios, indicating that observed state-to-state differences predominantly reflect genuine differences in service-system performance. The Social Connectedness measure (overall IUR = 0.774) is reliable in aggregate but loses precision in small-sample states (lowest decile IUR = 0.380), suggesting that minimum sample size requirements should be observed when interpreting state-level scores on this measure.

Validity. Convergent validity was supported for each measure: Satisfaction with Community Inclusion correlated most strongly with the conceptually proximate Free Time Choice (a person-centered planning item), and only weakly with the Community Inclusion Scale (a frequency-of-activity measure), supporting the distinction between these constructs. Social Connectedness correlated strongly with Satisfaction with Community Inclusion ($r = 0.69$, $p < 0.001$) and showed the expected positive associations with Has Friends and community engagement. Has Friends showed the expected negative correlation with Often Lonely and the expected positive associations with HCBS funding indicators, supporting its construct validity as an indicator of service-system support for relationship-building. Where correlations did not reach conventional significance thresholds, the directional consistency of associations and the small state-level sample sizes (26–39) suggest power limitations rather than absence of association, a caveat the NCI team acknowledges explicitly in the FMS submissions.

Together, these 2024–25 results provide updated empirical support for the continued endorsement of these NCI-IDD measures and demonstrate that the inter-unit reliability properties established at the time of NQF endorsement in 2022 have been maintained with current data.

References for Section 2.3. He, K., Kalbfleisch, J. D., Yang, Y., Fei, Z., Kim, S., Kang, J., & Li, Y. (2019). Inter-Unit Reliability for Quality Measure Testing. *Journal of Hospital Administration*, 8(2), 1–6. Hartman et al. (2024), cited in the source FMS submissions for context on IUR interpretation. State-level HCBS spending and enrollment data drawn from the KFF State Health Facts Database (2020/2022 estimates).

2.4 Other Published Psychometric Studies

Inclusion of NCI-IDD measures in peer-reviewed journals and applied research outputs provides ongoing evidence of face validity, demonstrating that the measures are considered valid by the scientific community for assessing the constructs of interest. This section groups the published work into three categories: (a) factorial and construct validity studies developing multi-item measurement models, (b) reliability and validation work on the Background Information section, and (c) external research analyses and applied/regional studies that have used NCI-IDD measures with results that are consistent with expected construct relationships.

Factorial and Construct Validity Studies

Factorial Validity of a Four-Factor Model of Personal Opportunities (Prohn et al., 2022)

Prohn, Dinora, Bogenschutz, Broda, and Lineberry (2022) evaluated the factorial validity of a four-factor model of personal opportunities using national 2017–2018 NCI-IPS data from 25,549 adults with IDD across 35 states and the District of Columbia (mean age 42; 59% male, 41% female; 67% White, 16% Black, 10% Latinx). The study responded to NQF's call for standardized measures to monitor HCBS quality. The four factors derived from NCI-IPS items were:

- Privacy Rights (2 items: key to home; lockable bedroom door)
- Everyday Choice (3 items: choice over schedule, free time, purchases)
- Community Participation (4 items: shopping, errands, entertainment, eating out)
- Expanded Friendships (8 items: participation with friends and having non-staff/family friends)

All four factors showed strong construct validity and significant inter-factor correlations. Privacy Rights and Everyday Choice were strongly correlated, reflecting shared environmental and cultural factors in residential settings. The authors concluded that personal opportunity factors derived from NCI-IPS data are reliable and valid indicators of HCBS quality.

Wellness Indicators Measure for Adults with IDD (Bogenschutz et al., 2021)

Bogenschutz, Broda, Dinora, Prohn, and Lineberry (2021) developed and validated a three-factor wellness indicators measure using national and state-level NCI-IPS data (national n = 25,477; Virginia pilot n = 809; demographics mirrored national trends). The three wellness domains, drawn from the IPS Background section, were:

- Cardiovascular Health (4 items): cardiovascular disease, diabetes, high blood pressure, high cholesterol
- Mental Health (4 items): mood disorder, anxiety disorder, psychotic disorder, other psychiatric diagnosis
- Behavioral Wellness (6 items): behavioral challenges, behavior-modifying medication, behavior plan, support needs for self-injurious, disruptive, and destructive behaviors

All three dimensions were significantly correlated but distinct, supporting a multidimensional structure. The study established the IPS wellness variables as a valid and reliable wellness model for adults with IDD receiving HCBS.

Reliability and Validation of the Background Information (BI) Section

BI Inter-Abstractor Reliability Study (Tichá, 2017)

As referenced in Section 2.1.1, Tichá (2017) reported on a dedicated reliability evaluation of the Background Information section of the NCI-IDD IPS — the section that collects demographic, disability, service, and health information from administrative records and knowledgeable proxy respondents rather than through direct interview. The study assessed inter-abstractor reliability by having multiple individuals abstract the same records and compared their entries item by item. Reliability rates ranged from 88% to 96% agreement across the participating states. Items related to employment, volunteering, managed care plan enrollment, funding sources, and specific health conditions showed particularly strong reliability. Importantly, the findings supported the use of state administrative data as a reasonably accurate source for the variables aggregated in the BI section, while also identifying items with comparatively lower agreement (typically open-text or judgment-based fields) where the NCI team has since prioritized state-level data cleaning protocols and the discrepancy-flag procedures described in Appendix A.4.

The Tichá (2017) findings provide a complementary line of evidence to the inter-rater reliability work described in Section 2.1.1: where the 1997–2010 studies and the 2010 shadowing study evaluated reliability of the surveyor-administered components of the instrument (Sections I and II), the BI study evaluated reliability of the administrative-records-derived component. Together, these results support reliability inferences across the three structurally distinct parts of the instrument.

External Research Analyses Using NCI-IDD Measures

Quality Monitoring Utility of the 14 NQF-Endorsed Measures (Bradley & Hiersteiner, 2022)

Bradley and Hiersteiner (2022), writing in Frontiers in Rehabilitation Sciences, evaluated the utility and applicability of the 14 NQF-endorsed NCI-IDD measures for state and national quality monitoring. The study surveyed state IDD program directors on the relevance of each of the 14 measures to their states' policy and quality-improvement priorities, organized by NQF framework domain (Person-Centered Planning and Coordination; Community Inclusion; Choice and Control; Human and Legal Rights). State managers ranked indicators for high, medium, or low utility for quality monitoring; results showed that the great majority of indicators were ranked of "high utility" by most states. Indicators in the Choice and Control domain (e.g., "the proportion of people who report making choices in life decisions") achieved the strongest endorsement, with 100% of responding states rating them as high-utility. The Bradley-Hiersteiner paper provides face-validity evidence at the level of system-level decision-makers, complementing the more granular construct- and convergent-validity evidence from the FMS submissions reviewed in Section 2.3.

Effects of Person-Centered Planning on Health and Well-Being (Isvan, Bonardi & Hiersteiner, 2023)

Isvan, Bonardi, and Hiersteiner (2023), in the Journal of Intellectual Disability Research, conducted a multilevel analysis linking NCI-IDD IPS responses with state administrative data to

test whether person-centered planning practices captured in the IPS are associated with better health and well-being outcomes for adults with IDD. The analysis used hierarchical models to account for clustering within states and tested whether respondents reporting more person-centered planning practices reported better outcomes across multiple domains, controlling for individual- and state-level covariates. The observed associations were in the predicted direction and statistically significant for several outcomes, providing convergent validity evidence for the PCP-related items in the IPS as indicators of a construct (person-centered care) that has measurable downstream effects. This pattern of findings has direct relevance to the FMS submission for the Satisfaction with Community Inclusion measure (PCP-5), where person-centered planning items were the strongest correlates (Section 2.3.4).

Characteristics and Outcomes of Older Adults with IDD (Bradley et al., 2020)

Bradley, Hiersteiner, Li, Bonardi, and Vegas (2020), in the Developmental Disabilities Network Journal, used the 2018–19 NCI-IDD IPS to characterize the experiences of adults with IDD aged 55 and over, with selected comparisons to the general population as measured by the National Health Interview Survey (NHIS). The study found that older adults with IDD were more socially isolated, had smaller social networks than their younger peers, had less access to transportation, and were more likely than the general older-adult population to report vision and hearing limitations, mobility needs, and mood or anxiety disorders. Beyond the substantive findings, the comparison with NHIS supports the convergent validity of the demographic, health, and social-network items in the BI section and Sections I/II: NCI-IDD IPS measures detected the expected age-related patterns (greater isolation among older adults; lower employment in older age groups; higher rates of sensory and mobility limitations) at magnitudes consistent with the literature on aging in the general population, while also revealing IDD-specific disparities. This pattern of “known-groups” differentiation is one form of construct-validity evidence.

Self-Injurious Behavior Support Needs (Bradley et al., 2018)

Bradley, Hiersteiner, Rotholz, Maloney, Li, Bonardi, and Bershadsky (2018), in the Journal of Intellectual Disability Research, analyzed NCI-IDD IPS data to identify personal characteristics and outcomes among individuals with developmental disabilities who need support for self-injurious behavior. The study used IPS background-section variables to define the population of interest and Sections I and II variables to compare outcomes between people with and without identified self-injurious behavior support needs. The patterns of association observed — including the expected relationships between support needs, residential setting, communication ability, and outcomes — provided convergent validity evidence for the BI items used to identify the population and for the outcome measures examined.

Life Experience and Outcomes of Adults on the Autism Spectrum (Hiersteiner et al., 2017)

Hiersteiner, Bradley, Ne’eman, Bershadsky, and Bonardi (2017), in the journal Inclusion, used the NCI-IDD IPS to compare outcomes of adults on the autism spectrum to outcomes of adults with other developmental disabilities. The comparison provided known-groups evidence for the IPS: where the literature predicts higher rates of certain conditions and lower rates of community engagement among adults with autism compared to other DD groups, the IPS

detected those patterns. The study also documented variation within the autism-spectrum group by age, residential setting, and communication mode, supporting the responsiveness of the instrument to within-group heterogeneity.

Applied and Regional Analyses

State-Level Reports and Regional Analyses

In addition to peer-reviewed work, the NCI-IDD program publishes annual national reports and member states publish state-specific and regional analyses that document patterns of variation across service systems. Recent national reports include the 2021–2022 report (27 states; n = 13,559 adults) and the 2022–2023 report (33 states; n > 25,000 adults). These reports describe distributions and cross-state patterns of the indicators reviewed in this document and have been used by state IDD agencies and federal partners (CMS, ACL) to inform HCBS Final Settings Rule implementation and HCBS Access Rule work. Although descriptive rather than evaluative in design, these state-level reports contribute to face-validity evidence by demonstrating that the instrument detects expected patterns of variation across service systems and over time. Regional and state-specific analyses have also been conducted (for example, California Regional Center reports drawing on state-supplemented IPS data, and state quality-improvement initiatives that use NCI-IDD data to track progress against state policy goals); these are useful primarily for confirming that the instrument supports the granular within-state analyses needed for state quality improvement.

Doctoral and Master's-Level Research

The NCI-IDD IPS is also used in graduate research that contributes psychometric evidence. The team is aware of doctoral dissertations and master's theses that have used IPS data for secondary analyses on topics including community inclusion, employment outcomes, self-determination, choice and control, and disparities by race, ethnicity, age, and residential setting. A consolidated bibliography of dissertations and theses using NCI-IDD data would be a useful addition to a future revision of this review; the present review focuses on peer-reviewed work and authoritative reports to keep the evidence base verifiable.

Note. This list is intended to be representative rather than exhaustive. The published literature drawing on NCI-IDD IPS data has grown substantially since 2017, and additional published work on specific indicators (employment, person-centered planning, transportation access, health and wellness outcomes, racial and ethnic disparities) is referenced in the bibliography of each annual NCI-IDD national report.

2.5 Summary Table of Psychometric Evidence

Table 5 consolidates the major psychometric studies reviewed in this section, including study type, sample size, statistical approach, and key findings. Detailed narrative descriptions appear in the corresponding subsections above.

Study / Source	Year(s)	Type of Evidence	Sample / Scope	Statistic(s)	Key Result(s)
1997 Pilot Test	1997	Inter-rater reliability	30 adults, 1 state	% agreement	93% agreement across all items
1998 IRR Study	1998	Inter-rater reliability	25 adults, paired interviewers	% agreement; Cohen's kappa	93% agreement; mean kappa = 0.794 (substantial)
1999 Reliability Test	1999	Inter-rater reliability	27 adults	% agreement	92% agreement (consistent with prior years)
2008 Post-Revision Test	2008	Inter-rater reliability	16 adults	Cohen's kappa	Mean kappa = 0.90 (almost perfect)
2010 Shadowing Study	2010	Inter-rater reliability under field conditions	20 interviews, 6 surveyors	Cohen's kappa; % agreement	Kappa range 0.82–0.95 (mean 0.89); 80% overall agreement
BI Reliability (Tichá)	2017	Inter-rater reliability of Background Info	Multiple states	% agreement	88–96% agreement; strongest for employment, volunteering, managed care, funding sources, specific health items
PCP Cognitive Pretest	2018–19	Cognitive testing; scale development	10 service users; 35 items	PCA, CFA, Cronbach's alpha, Spearman-Brown, item-total r	5 multi-item scales developed (CC-4, HLR-1, CI-3, CI-4, PCP-5); items understood after minor wording revisions
Remote Administration Pilot	2020–21	Mode equivalence	Subset of pandemic-era surveys	Comparison of response patterns	No significant differences between videoconference and in-person for most sections
Non-Responder Study	2020	Selection bias assessment	State 1: 429 R / 363 NR;	z-tests, chi-square, t-tests	Older adults and group-setting residents

Study / Source	Year(s)	Type of Evidence	Sample / Scope	Statistic(s)	Key Result(s)
			State 2: 2,137 R / 1,702 NR		overrepresented; parents'-home and self-directed underrepresented
NQF Endorsement — ANOVA	2022	Performance measure reliability	14 endorsed measures, all states	ANOVA on between- vs. within-state variance	Between-state variance significantly larger than within-state for all 14 measures ($p < 0.001$)
NQF Endorsement — IUR	2022	Inter-Unit Reliability	14 endorsed measures	IUR (signal-to-noise ratio)	IUR range 0.75–0.98; CC-1, CC-4, and PCP-5 approached 1.0
FMS — ADL Goal (PCP-3)	2024–25	IUR + convergent validity	34 states / 1,838 respondents	IUR; correlations	IUR = 0.882; mean score 76.7% (range 16–97%); all deciles ≥ 0.73
FMS — Satisfaction w/ Comm. Inclusion (PCP-5)	2024–25	IUR + convergent validity	39 states; multi-item scale	IUR; Spearman/Pearson correlations	IUR = 0.944; mean 61% (range 45–86%); strongest correlate Free Time Choice ($\rho = 0.41$, $p = 0.006$)
FMS — Social Connectedness (CI-1)	2024–25	IUR + convergent validity	39 states (single item)	IUR; Pearson correlations	IUR = 0.774; mean 87% (range 81–98%); strong correlation with Satisfaction with Community Inclusion ($r = 0.69$, $p < 0.001$)
FMS — Has Friends (CI-2)	2024–25	IUR + convergent/discriminant validity	39 states	IUR; Pearson/Spearman correlations	IUR = 0.952; mean 76% (range 44–91%); expected negative correlation with Often Lonely and Ages 65+

Study / Source	Year(s)	Type of Evidence	Sample / Scope	Statistic(s)	Key Result(s)
Prohn et al. (2022)	2022	Factorial validity	N = 25,549; 35 states + DC	Confirmatory factor analysis	4-factor model of personal opportunities supported; strong construct validity
Bogenschutz et al. (2021)	2021	Construct validity (wellness)	N = 25,477 + Virginia pilot n = 809	Confirmatory factor analysis	3-factor wellness model (cardiovascular, mental health, behavioral) supported
Bradley & Hiersteiner (2022)	2022	Face validity / utility for quality monitoring	Survey of state IDD program directors; 14 NQF-endorsed measures	State-rated utility (high/medium/low)	Most measures rated high-utility by majority of states; Choice and Control measures strongest (100% high-utility)
Isvan, Bonardi & Hiersteiner (2023)	2023	Convergent/predictive validity of PCP measures	Linked NCI-IPS and administrative data, multi-state	Multilevel models	Person-centered planning practices significantly associated with better health and well-being outcomes
Bradley et al. (2020) — Older Adults	2018–19 data	Known-groups / construct validity	NCI-IPS subsample aged 55+; comparison with NHIS	Descriptive comparison	Expected age-related patterns detected; older adults with IDD more isolated, more sensory and mobility needs than NHIS peers
Bradley et al. (2018) — SIB Support Needs	2018	Convergent validity (BI × outcomes)	Multi-state NCI-IPS	Group comparisons	Expected associations between SIB support needs, residential setting, communication, and outcomes

Study / Source	Year(s)	Type of Evidence	Sample / Scope	Statistic(s)	Key Result(s)
Hiersteiner et al. (2017) — Autism	2017	Known-groups validity	Autism-spectrum subsample compared to other DD groups in NCI-IPS	Group comparisons	Predicted differences in community engagement, health, and support needs detected

Note. Cell entries reflect the most representative summary statistic reported in each study. R = respondents; NR = non-respondents.

2.6 Important Considerations for State-Specific Analyses

The psychometric evidence above is based on national or multi-state analyses. State-specific analyses do not always replicate these results because of variation in surveyor training and experience, sampling approaches, state-specific procedural adaptations, the extent of state-level quality assurance, the demographic and clinical characteristics of state populations, and sample size. States are encouraged to conduct their own psychometric analyses when using NCI-IDD data for state-level decision-making and to interpret results in the context of their specific circumstances and data quality indicators.

3. Future Directions

The NCI-IDD team is committed to continuous improvement of the survey's psychometric properties through ongoing research, methodological innovation, and responsiveness to emerging needs. This section outlines planned and recommended priorities.

3.1 Ongoing Psychometric Monitoring

The NCI team will continue periodic inter-rater reliability studies, factor analyses to confirm scale structure, and assessment of internal consistency for multi-item scales. Data quality indicators will be tracked across waves to identify trends and emerging issues, and ongoing collaboration with states will support continuous improvement of data collection and quality assurance procedures.

3.2 Survey Modality Research

Telephone Pilot Testing

A pilot test of telephone administration is planned. Key questions include whether the survey can be administered effectively by telephone with people with IDD, whether results are comparable to in-person administration, what participant and surveyor experiences look like, and what procedural adaptations are required. Comprehensive psychometric testing for telephone administration will include reliability and validity studies, inter-rater reliability for telephone surveyors, comparison of response patterns across modes, and assessment of systematic bias. Successful development of telephone administration could expand access for difficult-to-reach individuals and reduce travel costs.

Cross-Modality Equivalence

Building on the videoconference pilot and the planned telephone pilot, future research should systematically examine equivalence of results across in-person, telephone, videoconference, and possibly online or tablet-based administration. Studies should identify best practices for adapting administration to each modality while maintaining psychometric integrity, and examine accessibility for individuals with sensory impairments, motor limitations, communication challenges, and varying technology access.

3.3 Family Survey Psychometric Development

The NCI-IDD Family Survey, which collects information from family members of people with IDD, requires additional psychometric development to establish it as a scientifically rigorous companion to the IPS. Priorities include test-retest reliability studies and internal consistency

analyses for multi-item scales, construct and criterion validity studies, factor analyses to confirm structure, and cognitive testing with family respondents. Particular attention is needed for measures of family satisfaction, family support needs, family-provider partnership, and family outcomes.

3.4 Responsiveness and Cultural Competence

Future psychometric work should be stratified by characteristics that may affect survey performance, including primary language (separate analyses for English and other languages, and assessment of whether translated versions and interpretation services preserve psychometric properties), race and ethnicity (to identify differential item functioning or measurement bias), and alternate modes of communication (augmentative and alternative communication, sign language, and other methods). Translation, cultural adaptation, cognitive testing with diverse groups, and enhanced surveyor training in cultural competence will be needed alongside formal differential item functioning (DIF) analyses.

3.5 Strengthening Background Information (BI)

Improving the accuracy and reliability of BI data is a priority. Planned activities include validation studies comparing BI data with other authoritative sources (medical records, service plans, billing data), discrepancy and error-pattern analyses across administrative systems, development of standardized protocols for state extraction and verification, and regular audits with feedback to states.

3.6 Additional Planned Validity Studies

Criterion validity. Future research should link IPS measures to objective outcomes such as health records, employment data, service utilization patterns, and other validated quality indicators, asking whether NCI measures predict important outcomes and identify individuals or groups at risk for poor outcomes.

Longitudinal studies. Test-retest studies with extended intervals can assess measure stability for constructs that should be relatively stable, while pre-post studies around major service-system changes can assess sensitivity to change. Tracking individuals over multiple waves can detect meaningful change in circumstances and outcomes.

3.7 Integration of New Statistical Methods

Item Response Theory (IRT) can assess item functioning at a more granular level, identify optimal item characteristics, and support shorter forms or adaptive testing. *Machine learning*

can identify patterns in response data, predict data quality issues, and support automated anomaly detection. *Network analysis* can examine relationships among constructs and identify central or peripheral elements in the conceptual structure of quality of life and service quality. *Bayesian methods* can provide more nuanced analysis when sample sizes are limited or prior information from previous studies is informative.

3.8 Payment and Incentive Effects

An important area for future research is how payment and incentives affect response rates, response patterns, and data quality. Experimental or quasi-experimental designs comparing incentive structures (no incentive, small, larger, different types) could inform state policies on participant compensation while maintaining data integrity.

3.9 Collaboration and Sustainability

The NCI-IDD team welcomes collaboration with academic researchers, state agencies interested in supplementary validation studies, advocacy organizations representing diverse IDD populations, and international researchers working on comparable instruments. Long-term sustainability of psychometric monitoring will require dedicated resources for ongoing research and validation, systematic documentation of psychometric evidence and methods, regular training for NCI team members and state partners, integration of psychometric monitoring into routine survey operations, and a standing psychometric advisory committee to guide ongoing work.

The future directions outlined here demonstrate a continued commitment to scientific rigor and to ensuring that the IPS remains responsive to the evolving needs of people with IDD, their families, providers, policymakers, and researchers.

Appendix A. The NCI-IDD In-Person Survey

This appendix provides background on the NCI-IDD IPS itself, its purpose, instrument, data collection process, and data validation procedures, to support readers who want fuller context for the psychometric evidence presented in the main text. Material here is descriptive rather than evaluative; readers focused on measurement evidence can consult Section 2.

A.1 Purpose of the Survey

The IDD population includes individuals with intellectual disability, autism spectrum disorders, cerebral palsy, epilepsy, fetal alcohol spectrum disorders, fragile X syndrome, and other related conditions. This population is notably diverse, with individuals having varying levels of support needs, communication abilities, living situations, and life experiences. Some live independently with minimal support; others require extensive assistance with daily activities and decision-making. People with IDD often require ongoing supports, residential services, day programs, employment assistance, healthcare coordination, transportation, assistive technology, and various therapeutic interventions, to participate fully in community life, maintain health and safety, achieve personal goals, and exercise their rights.

The NCI-IDD IPS was developed to provide standardized, person-centered measurement of the quality of services and supports provided to adults with IDD. Its primary purpose is to capture the experiences, outcomes, and satisfaction of individuals receiving IDD services from the perspective of the recipients themselves whenever possible. It moves beyond traditional administrative measures of service utilization and costs to focus on the experiences and outcomes that matter most to people with disabilities and their families.

How the Data Respond to States' and Researchers' Needs

Filling information gaps. Administrative data systems typically capture services delivered, costs incurred, and basic demographics, but provide limited insight into whether services are achieving their intended outcomes from the perspective of recipients. The IPS takes a person-centered perspective by directly assessing individuals' experiences with services, their satisfaction, and their achievement of community living outcomes.

Measuring key quality constructs. The IPS measures person-centered care, self-determination, community inclusion, and quality of life, multifaceted concepts that are central to contemporary disability policy but are difficult to assess through administrative data alone.

Supporting equity work. With intentional sampling, the IPS can collect data needed to identify outcome disparities across race, ethnicity, age, disability type, and geography, and to support state strategies for addressing them.

Adapting to evolving priorities. Each participating state can add a limited number of state-specific questions to address particular policy priorities, while core content is held constant for

cross-state and trend analyses. A COVID supplement was developed during the pandemic and retired in 2022–23 as priorities shifted; Person-Centered Planning items were added to support measurement of HCBS quality.

Special Considerations for Survey Research with IDD Populations

Conducting survey research with IDD populations requires careful attention to factors that distinguish this work from research with general populations: communication and cognitive considerations that shape question design, response formats, and interview techniques; proxy response protocols for individuals who cannot participate directly; informed consent procedures adapted for different communication abilities and cognitive levels; sampling and representativeness challenges arising from limited sampling frames; cultural and linguistic diversity within the IDD population; and ethical attention to the protection of vulnerable participants and minimization of burden. These considerations are operationalized in the survey instrument and the data collection procedures described below.

A.2 The Survey Instrument

The IPS is organized into distinct sections, each designed to capture different aspects of individuals' experiences while accommodating diverse communication abilities and support needs.

Background Information (BI) Section

The BI section collects essential demographic, disability, and service information about survey participants. This information is typically gathered from administrative records, case files, and knowledgeable staff rather than through direct interview. The BI section includes:

- Basic demographics (age, gender, race/ethnicity)
- Disability characteristics, health conditions, and support needs
- Residential setting and living arrangements
- Employment status and other activities
- Service utilization patterns and funding sources

BI data provides context for interpreting survey responses, enables subgroup analyses to identify disparities, supports sampling and weighting procedures, and allows comparison of survey participants with the broader IDD population. The BI section also represents an important example of data linkage, aggregating in one dataset information from multiple sources.

Section I — Person Receiving Services

Section I consists of questions asked directly of the individual with IDD whenever possible. Proxy responses are not allowed for Section I questions, reflecting the commitment to capturing authentic self-reported experiences and perspectives. The prohibition on proxy responses rests on both philosophical and methodological grounds: it affirms the right of individuals to speak for themselves, and research has shown that proxy respondents may have systematically different perspectives than the individuals themselves on subjective matters such as satisfaction, preferences, and quality of life. A series of standardized Proxy Determination questions are included to help the surveyor determine whether the person can answer Section I.

Section I covers community inclusion and belonging, employment, choice and control, relationships, satisfaction with services and supports, and other topics best answered by the individual.

Section II — Background and Other Information

Section II covers content for which proxy responses are appropriate and may be answered by a knowledgeable proxy when needed. It complements Section I by capturing information that may be more accurately reported by someone with detailed knowledge of the individual's services, health, and circumstances.

A.3 Data Collection Process

Sampling

Sampling procedures for the NCI-IDD IPS vary across participating states. States typically work from administrative lists of individuals receiving IDD services to define their sampling frame, and use stratification (for example, by region, residential setting, or service type) to support meaningful within-state comparisons. Specific sampling designs, frame definitions, and stratification variables are determined at the state level, and readers should consult state-specific documentation for details. Sample size calculations are computed to support state-level estimation; participating states are required to aim for a minimum sample size that achieves a 5% margin of error and 95% confidence based on the sample frame, with consideration of expected response rates, design effects from stratification and clustering, and analytical needs for subgroup comparisons.

Surveyor Training

Surveyors receive standardized training before conducting interviews. Training covers the survey instrument, administration procedures, communication strategies for interviewing people with IDD, proxy determination, informed consent, and quality assurance practices. Training includes practice interviews and standardized scripts, and procedures for handling difficult interview situations. Refresher training and ongoing quality assurance support consistent administration across surveyors and states.

Contacting Participants and Informed Consent

Potential survey participants are contacted at the state level. Initial contact procedures typically involve reaching out to individuals through their service providers or residential settings, which helps ensure that contact attempts are appropriate and that individuals receive necessary support to understand and respond to invitations. Informed consent processes are adapted to accommodate different communication abilities and cognitive levels and may involve additional time, simplified language, visual aids, or alternative communication methods. The consent process covers the purpose of the survey, what participation involves, how information will be used and protected, the voluntary nature of participation, and the right to refuse or withdraw at any time. Where individuals do not have the capacity to provide informed consent independently, legally authorized representatives may provide consent while still involving the individual to the greatest extent possible.

Surveyor Assessment of Section I Cognition

Before beginning Section I questions, surveyors conduct a brief assessment to determine whether the individual can participate meaningfully. This assessment is supported by the standardized Proxy Determination questions included in the survey, and considers the individual's ability to understand and respond to simple questions, the coherence and relevance of responses, attention and engagement, communication ability and method, and apparent comprehension of the survey purpose. Surveyors are trained to make this determination using standardized criteria while remaining sensitive to different communication styles and abilities. If an individual is determined unable to participate in Section I, the surveyor proceeds directly to Section II with an appropriate proxy respondent.

Section II Validation Flag

For individuals who participate in Section II through proxy respondents, surveyors assign validation flags to indicate their assessment of the quality and reliability of the proxy responses. Flags consider the proxy's knowledge of the individual, the quality and consistency of responses, evidence of bias or limitations in proxy knowledge, and completeness of responses. The validation flag system documents data quality concerns and supports appropriate use of proxy-provided information in analysis.

Note. The precise interpretation of the surveyor invalid flag, in particular, whether it applies when any single question appears not understood or only when no questions in the section appear understood, should be confirmed against current administration documentation.

A.4 Data Validation

ODESA and Real-Time Logic Checks

The NCI-IDD survey has a validation process to ensure the accuracy, completeness, and consistency of collected data. Many of these procedures are integrated into the Online Data

Entry System and Analysis (ODESA) platform, which provides real-time validation and quality control. ODESA logic checks are implemented during data entry to identify potential errors or inconsistencies as data are entered. They include range checks (responses within valid values), logic checks (consistency across related questions), completeness checks (missing required items), and pattern checks (unusual patterns suggesting data entry errors). Inconsistent responses are flagged in real time so surveyors and data entry staff can correct errors immediately. ODESA also generates data quality reports for state review, highlighting issues such as high rates of missing data, unusual response patterns, or deviations from expected distributions.

Data Cleaning and State-Level Review

Beyond automated checks, the NCI team conducts systematic data cleaning in collaboration with states. This includes confirming validity flags with states during cleaning, verifying flagged cases, and resolving discrepancies between BI data sources and self-reported information. States play a critical role by reviewing and verifying specific data elements that are particularly prone to error or inconsistency:

Wage data. States review reported wages for individuals in community employment to identify and correct implausible values or data entry errors. This includes checking for wages outside reasonable ranges, missing decimal points, or confusion between hourly and annual wages.

Race/ethnicity data. States verify race and ethnicity information to ensure accurate demographic representation and identify missing or miscoded data, supporting analyses of disparities and equity.

Living arrangement and self-direction combinations. States review cases where individuals are reported to live in specific residential settings (e.g., agency-operated group homes) and also use self-directed supports, since some combinations may indicate data entry errors or warrant clarification.

Other critical BI fields. States conduct targeted reviews of additional BI fields based on identified patterns of missing or inconsistent data, including disability type, level of intellectual disability, guardianship status, funding sources, and service utilization. For some demographic variables, data are collected in both the BI section (administrative records) and through self-report in the interview; discrepancies are flagged for state review as part of data quality assurance.

Decisions to Exclude Records

Not all collected responses are included in final analytic datasets. The NCI team applies systematic procedures to decide which records to exclude, with the validity flag confirmed with states during cleaning rather than relied on as an automatic exclusion.

Section I and Section II invalid flags. At the end of Section I, the surveyor indicates whether the respondent appeared to understand and answer questions consistently. Where the surveyor's response is negative, Section I data are excluded from analysis after confirmation with the state.

If Section I data are excluded based on this assessment, Section II data are also excluded for that case unless a proxy respondent was used for Section II.

Other reasons for exclusion. Records may be excluded for incomplete surveys with insufficient data to calculate measures, surveys conducted outside the specified data collection period, duplicate records identified during validation, cases where informed consent was not properly documented, or surveys that do not meet minimum data quality standards.

All exclusion decisions are documented with clear rationales, and the impact of exclusions on sample representativeness is evaluated and reported.

Treatment of Missing Data and National Reporting Standards

The NCI-IDD survey uses several strategies for handling missing data. Item-level missing data patterns are analyzed to determine whether they are random or systematic; systematic patterns may indicate problems with question design, administration, or data entry that require attention. At the point of analysis, data may be excluded if state-level sample sizes are too small to support reliable estimation. For Background Information, if a state is missing more than 25% of data in the BI section, that state's data for affected measures are marked with an asterisk in national reports. Health-related data are typically presented twice — once with and once without missing data included in the denominator — to support transparent interpretation. The extent and patterns of missing data are documented in technical notes and data quality reports.

Ongoing Quality Assurance

Quality assurance includes regular monitoring of data quality indicators across states, feedback to states on issues identified during validation, continuous refinement of data collection and validation procedures, and documentation of data quality metrics in annual reports.

References

Bogenschutz, M., Broda, M., Lineberry, S., Dinora, P., & Prohn, S. (2021). Testing a wellness indicators measure for people with intellectual and developmental disabilities. Developmental Disabilities Network Journal, 1(2), 1–24.

Bradley, V. J., & Hiersteiner, D. (2022). Quality monitoring of intellectual and developmental disabilities systems in the US: Assessing the utility and applicability of selected National Core Indicators to national and state priorities. Frontiers in Rehabilitation Sciences, 3, 960996. https://doi.org/10.3389/fresc.2022.960996

Bradley, V., Hiersteiner, D., Li, H., Bonardi, A., & Vegas, L. (2020). What do NCI data tell us about the characteristics and outcomes of older adults with IDD? Developmental Disabilities Network Journal, 1(1), Article 6. https://doi.org/10.26077/esw0-2h31

Bradley, V., Hiersteiner, D., Rotholz, D., Maloney, J., Li, H., Bonardi, A., & Bershadsky, J. (2018). Personal characteristics and outcomes of individuals with developmental disabilities who need support for self-injurious behaviour. Journal of Intellectual Disability Research. https://doi.org/10.1111/jir.12518

Friedman, C. (2023). Medicaid Home and Community Based Services Waivers for People with Intellectual and Developmental Disabilities. AAIDD. https://www.aaidd.org/docs/default-source/prepressarticles/medicaid-home-and-community-based-services-waivers-for-people-with-intellectual-and-developmental-disabilities.pdf

Hartman et al. (2024). [Cited in NCI-IDD FMS submissions for Spring 2026 PQM cycle regarding IUR interpretation when values exceed 0.9. Full citation to be verified from source.]

He, K., Kalbfleisch, J. D., Yang, Y., Fei, Z., Kim, S., Kang, J., & Li, Y. (2019). Inter-Unit Reliability for Quality Measure Testing. Journal of Hospital Administration, 8(2), 1–6. https://doi.org/10.5430/jha.v8n2p1

Hiersteiner, D., Bradley, V., Ne'eman, A., Bershadsky, J., & Bonardi, A. (2017). Putting the research in context: The life experience and outcomes of adults on the autism spectrum. Inclusion, 5(1), 45–59. https://doi.org/10.1352/2326-6988-5.1.45

Isvan, N., Bonardi, A., & Hiersteiner, D. (2023). Effects of person-centered planning and practices on the health and well-being of adults with intellectual and developmental disabilities: a multilevel analysis of linked administrative and survey data. Journal of Intellectual Disability Research. https://doi.org/10.1111/jir.13015

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

National Quality Forum. (2016). Quality in Home and Community Based Services to Support Community Living: Addressing Gaps in Performance Measurement. Washington, DC.

Prohn, S., Dinora, P., Bogenschutz, M., Broda, M., & Lineberry, S. (2022). Measuring four personal opportunities for adults with intellectual and developmental disabilities. Inclusion, 10(1), 19–34. https://doi.org/10.1352/2326-6988-10.1.19

Smith, G., & Ashbaugh, J. (2001). National Core Indicators Project: Phase II consumer survey technical report. Cambridge, MA: Human Services Research Institute.

Tichá, R. (2017). National Core Indicators (NCI) background study [PowerPoint presentation]. University of Minnesota, Institute on Community Integration. https://legacy.nationalcoreindicators.org/upload/presentation/Renata_NCI_background_study_2017.pdf