

A Reliability Study of the Indiana Risk Assessment System

FINAL REPORT

(submitted 11/21/19)

Prepared for the State of Indiana Office of Court Services

Prepared by the University of Cincinnati Corrections Institute (UCCI)

Sarah M. Manchak, Ph.D.

Damon Petrich, M.A.

Eric Willoughby, M.S.

# Table of Contents

Introduction and Overview.....	1
Methodology .....	3
Participants.....	3
<i>Participant Selection</i> .....	3
<i>Participant Recruitment and Enrollment</i> .....	5
<i>Demographic Characteristics of Enrolled Participants</i> .....	8
Measures .....	16
<i>Creation of Study Vignettes</i> .....	17
<i>Establishment of Criterion Scores</i> .....	17
Results: Opinion Survey .....	18
Results: Reliability .....	24
Explanation of Research Findings and Instructions for Interpretation .....	24
IRAS Reliability Results .....	27
<i>Community Supervision Tool (CST)</i> .....	27
<i>Supplemental Reentry Tool (SRT)</i> .....	30
<i>Prison Intake Tool (PIT)</i> .....	33
<i>Pretrial Assessment Tool (PAT)</i> .....	36
IYAS Reliability Results.....	38
<i>Disposition Tool (DIS)</i> .....	38
<i>Detention Tool (DET)</i> .....	41
<i>Diversion Tool (DIV)</i> .....	43
<i>Residential Tool (RES)</i> .....	45
Reentry Tool (RET) .....	48
Conclusions & Recommendations.....	51
References.....	52

## List of Tables

Table 1. Recruitment Targets for Each Tool.....	4
Table 2. Enrollment Figures by Risk Assessment Tool.....	7
Table 3. Community Supervision Tool- Enrolled Participant Demographic Information.....	8
Table 4. Prison Intake Tool- Enrolled Participant Demographic Information.....	9
Table 5. Pretrial Assessment Tool- Enrolled Participant Demographic Information.....	10
Table 6. Supplemental Reentry Tool- Enrolled Participant Demographic Information.....	11
Table 7. Disposition Tool- Enrolled Participant Demographic Information.....	12
Table 8. Detention Tool- Enrolled Participant Demographic Information.....	13
Table 9. Diversion Tool- Enrolled Participant Demographic Information.....	14
Table 10. Residential Tool- Enrolled Participant Demographic Information.....	15
Table 11. Reentry Tool- Enrolled Participant Demographic Information.....	16
Table 12. Community Supervision Tool-Agreement Ratings.....	19
Table 13. Prison Intake Tool-Agreement Ratings.....	19
Table 14. Pretrial Assessment Tool-Agreement Ratings.....	20
Table 15. Supplemental Reentry Tool-Agreement Ratings.....	20
Table 16. Disposition Tool-Agreement Ratings.....	21
Table 17. Detention Tool-Agreement Ratings.....	21
Table 18. Diversion Tool-Agreement Ratings.....	22
Table 19. Residential Tool-Agreement Ratings.....	22
Table 20. Reentry Tool-Agreement Ratings.....	23
Table 21. CST Average Deviation from Criterion Scores; M (SD).....	27
Table 22. CST Average Criterion ICC across 3 Vignettes; M (SD).....	28
Table 23. Correlation of CST Criterion ICC Total Score with Rater Characteristics.....	28
Table 24. CST Inter-rater Reliability ICC (CI Low, CI High).....	29
Table 25. SRT Average Deviation from Criterion Scores; M (SD).....	30
Table 26. SRT Average Criterion ICC across 3 Vignettes; M (SD).....	31
Table 27. Correlation of SRT Criterion ICC Total Score with Rater Characteristics.....	31
Table 28. SRT Inter-rater Reliability ICC (CI Low, CI High).....	32
Table 29. PIT Average Deviation from Criterion Scores; M (SD).....	33
Table 30. PIT Average Criterion ICC across 3 Vignettes; M (SD).....	33
Table 31. Correlation of PIT Criterion ICC Total Score with Rater Characteristics.....	34
Table 32. PIT Inter-rater Reliability ICC (CI Low, CI High).....	34
Table 33. PAT Average Deviation from Criterion Scores; M (SD).....	36
Table 34. PAT Criterion ICC; M (SD).....	36
Table 35. Correlation of PAT Criterion ICC Total Score with Rater Characteristics.....	36
Table 36. PAT Interrater Reliability ICC (CI Low, CI High).....	37
Table 37. DIS Average Deviation from Criterion Scores; M (SD).....	38
Table 38. DIS Criterion ICC; M (SD).....	39
Table 39. Correlation of DIS Criterion ICC Total Score with Rater Characteristics.....	39
Table 40. DIS Interrater Reliability.....	40
Table 41. DET Average Deviation from Criterion Scores; M (SD).....	41
Table 42. DET Criterion ICC; M (SD).....	41
Table 43. Correlation of DET Criterion ICC Total Score with Rater Characteristics.....	41
Table 44. DET Interrater Reliability ICC (CI Low, CI High).....	42
Table 45. DIV Average Deviation from Criterion Scores; M (SD).....	43
Table 46. DIV Criterion ICC; M (SD).....	43
Table 47. Correlation of DIV Criterion ICC Total Score with Rater Characteristics.....	43
Table 48. DIV Interrater Reliability ICC (CI Low, CI High).....	44
Table 49. RES Average Deviation from Criterion Scores; M (SD).....	45

Table 50. RES Average Criterion ICC across 3 Vignettes; M (SD).....	46
Table 51. Correlation of RES Criterion ICC Total Score with Rater Characteristics .....	46
Table 52. RES Inter-rater Reliability ICC (CI Low, CI High).....	47
Table 53. RET Average Deviation from Criterion Scores; M (SD) .....	48
Table 54. RET Average Criterion ICC across 3 Vignettes; M (SD)* .....	49
Table 55. Correlation of RET Criterion ICC Total Score with Rater Characteristics .....	49
Table 56. RET Inter-rater Reliability ICC (CI Low, CI High).....	50

## **Introduction and Overview**

In order for risk assessment tools to be valid, they must be reliably rated by users. That is, risk assessment tools cannot accurately predict offenders' risk for recidivism and correctly classify offenders into appropriate risk classifications if the users are not using and scoring the tool correctly. There are a number of reasons why there may be poor reliability—and therefore improper scoring of a risk assessment tool. These include: (1) lack of training or poor training on the tool, (2) confusion about item criteria or scoring, (3) failure to follow the proper scoring guidelines established in the risk assessment manuals, (4) rater “drift” over time (e.g., relying on heuristics or memory over time rather than strictly adhering to scoring guidelines) (5) lack of substantial information needed to score an item, (6) personal bias(es), (7) distractions or depleted resources in the work environment...to name a few.

Poor reliability results in decreased validity of the risk assessment tools. In turn, this can result in the misallocation of resources, with higher risk offenders not receiving as many services or as much supervision as required because they were misclassified as lower risk, or lower risk offenders receiving too many resources and/or being exposed to high risk people and situations because they were classified at a higher risk level. Reliability can be enhanced with proper training and certification, booster trainings, ongoing supervision, random “spot checks” from supervisors of users' scores, and clear, open communication between users and the tool developers or trainers when any instances of confusion arise while scoring out the instruments.

The current reliability study was conducted at the request of the State of Indiana as part of a larger evaluation and revalidation of the state's risk assessment system. Specifically, researchers evaluated the Indiana Risk Assessment System (IRAS) for adults, which included evaluation of four tools (Community Supervision Tool, Supplemental Reentry Tool, Prison Intake Tool, and the Pretrial Assessment Tool), and the Indiana Youth Assessment System (IYAS) for juveniles, which included evaluation of five tools (Disposition, Diversion, Detention, Residential, and Reentry). As part of the larger re-validation project, it was necessary to determine the extent to which users are accurately rating the risk assessment tools, as compared to established criterion scores. Additionally, it was also necessary to determine how consistently IRAS and IYAS users are rating the tool amongst themselves.

This study is part of a larger project and ongoing partnership between Indiana and the

University of Cincinnati Corrections Institute (UCCI). The study was conducted by Sarah Manchak, Ph.D., Associate Professor in the School of Criminal Justice at the University of Cincinnati<sup>1</sup>. Dr. Manchak has been researching risk assessment for sixteen years, is certified in the administration of several assessment tools, including the Ohio Risk Assessment System (ORAS<sup>2</sup>), and recently led a state-wide reliability study of the ORAS. Dr. Manchak was assisted by Damon Petrich, M.A., a doctoral student in the School of Criminal Justice at the University of Cincinnati and Graduate Research Assistant with UCCI, and Eric Willoughby, M.S., a Junior Research Associate with UCCI who is also a certified trainer of both the Ohio Risk Assessment System and the Ohio Youth Assessment System. Data collection began on Wednesday, February 20, 2019 and closed/ended on Friday, March 22, 2019. Data cleaning and preparation occurred in April and May 2019, and data analysis and report writing took place in June and July 2019.

---

<sup>1</sup> Correspondence concerning this report can be addressed to [sarah.manchak@uc.edu](mailto:sarah.manchak@uc.edu)

<sup>2</sup> The IRAS and IYAS are based off of the ORAS and OYAS, respectively.

# Methodology

## Participants

All participants were currently active users of one or more of the risk assessment tools in the IRAS or IYAS. Additionally, all participants were, in fact, certified or authorized users of the tool for which they were randomly selected to participate, as verified by self-report or administrative data. All participants were employed by an agency within the state of Indiana, and all were competent to give consent to participate in the study.

### *Participant Selection*

Researchers were initially provided with a full list of IRAS and IYAS users who completed assessments between August 1, 2017 and August 24, 2018. This list also indicated which assessments were completed by each user during this time, and how many assessments for each tool were completed for each user. It was from this list that participants were selected for recruitment.

Probability (i.e., random) sampling was employed in the present study. To increase the confidence with which practitioners, policy makers, and other stakeholders have in the results, an enrollment goal of 20% of all registered IRAS/IYAS users was set. Because researchers anticipated potential issues in response rate due to staff turn-over (from the 2017-2018 list) and, to a lesser extent, staff refusal to participate, 25% of users for each tool were initially recruited into the study. Ultimately, the study sought to assess reliability of one out of every five IRAS or IYAS users within the state of Indiana, and random sampling from the wider possible pool of participants ensured the findings are generalizable to the other 80% of users.

Using the list of certified or authorized users provided by the state of Indiana, the researcher assigned users to be selected for participation in the reliability assessment for one and only one of the state assessment tools. In large part, this assignment process was dictated by the total number of assessments they completed for each tool in the 8/1/17-8/24/18 period. In most cases, individuals were assigned to the tool which they completed most often, if they completed assessments with more than one tool. In some cases, individuals were assigned to the tool for which they completed the second or third highest number of assessments in the prior year. Users who completed less than 20 assessments of that tool in the prior year (less than approximately 2

assessments/month) were excluded from the selection pool of each tool. This was done in large part because it was assumed that these users are not regular assessors or users of the tool and therefore would not be representative of the “average” user.

Once assigned to a tool, participants were listed in a random order (i.e., not ordered alphabetically or by frequency of assessments) and assigned a numeric value/code. Researchers then used a random number generator from the internet to select participants for each tool until the total number of participants was equivalent to 25% of the total assigned sample for each tool. Table 1 below reports starting sample sizes and recruitment targets. The final column on the right reports the total number of people for each tool who were ultimately contacted and recruited for participation.

*Table 1. Recruitment Targets for Each Tool*

	<b>Total Starting Sample Size</b>	<b>Total Users with 20+ Assessments</b>	<b>Total Assigned</b>	<b>20% of users w/ 20+ Assessments (Target N)</b>	<b>25% of Users with 20+ Assessments (Total Recruited)</b>
<b>Adult Tools</b>					
Community Supervision*	1585	1411	1369	282	353
Prison Intake Pretrial Assessment Supplemental Reentry*	266	182	139	36	46
	589	229	170	47	58
	339	231	131	46	58
<b>Youth Tools</b>					
Disposition*	473	362	159	72	91
Detention	366	180	97	36	45
Diversions	399	260	109	52	65
Residential	341	94	36	19	24
Reentry	255	66	56	13	17
<b>FULL SAMPLE</b>	<b>4613</b>	<b>3015</b>	<b>2266</b>	<b>603</b>	<b>757</b>



## ***Participant Recruitment and Enrollment***

Prior to the study, each agency supervisor was notified by Michelle Goodman (Indiana Office of Court Services) about the reliability study and was asked to give their staff a “heads up” about the upcoming study and encourage staff participation. Ms. Goodman also provided the list of names randomly selected for recruitment, so that supervisors could encourage these individuals directly to participate. During this process, Ms. Goodman was able to determine many of the initially selected users who had retired, relocated, left the agency, or switched positions/responsibilities and thus were no longer conducting risk assessments. She provided these names to the researchers, who then removed the name from the list of possible participants and replaced it with another randomly selected name from the assigned pool of participants. A similar approach was employed for “bounce-back” emails once the initial recruitment email went out to users. That is, if an email bounced back—typically indicating the person was no longer an employee with the agency—the person was marked as ineligible and a new person was randomly selected for recruitment.

Once the study began, participants were recruited by the researchers directly through email. A brief overview of the study goals and timeline was provided in the body of the email, along with encouragement to use the IYAS/IRAS user manuals when scoring the vignettes. The email also included the link to participate in the on-line study and a unique access code, which was used internally by researchers to track participation and inform follow-up recruitment efforts. In an accompanying attachment was a PDF file that included the 3 study vignettes and accompanying collateral material. Reminder emails were sent two weeks and three weeks after this initial email to all those who had still not responded to recruitment efforts by those dates.

Table 2 reports the recruitment and enrollment outcomes for all those who were eligible to participate. When an individual responded directly to recruitment efforts and expressed their wishes to *not* participate, they were recorded as an “active refusal”. Anyone else for whom the researcher had a working email (i.e., it did not bounce back) and who did not participate but did not otherwise respond to recruitment efforts was marked as a “non-responder”. The raw recruitment, refusal, and enrollment numbers with accompanying refusal/non-response and participation rates are reported below in Table 2. Although most participants were considered “enrolled” in the study, there were a few people who participated who only completed the demographics section and did not score any vignettes ( $n = 13$  across all tools). These individuals were removed from the study for all analyses. Thus, the total enrolled column reflects all those

who participated and scored at least one vignette. The final column provides the percent-to-goal, which is the total enrolled divided by the total goal sample size for each tool (which was 20% of all users with 20+ assessments who were assigned to each tool). As shown in Table 2, there was a 63% or higher participation rate across all tools, with an average participation rate of 74%. The highest participation rate was observed for the IYAS Diversion Tool (DIV), and the lowest participation rate was observed for the IRAS Prison Intake Tool (PIT). Enrollment goals for the study were met at a rate of 75% or higher across all tools, again with the DIV tool showing the highest target met rates and the PIT showing the lowest.

*Table 2. Enrollment Figures by Risk Assessment Tool*

	<b>Total Recruited</b>	<b>Non-Responders</b>	<b>Active Refusals</b>	<b>Refusal / Non-response Rate</b>	<b>Total Participants</b>	<b>Participation Rate</b>	<b>Total Enrolled</b>	<b>% to Target Enrollment Goal</b>
<b>Adult Tools</b>								
Community Supervision	353	73	12	24%	268	76%	264	94%
Prison Intake	46	17	0	37%	29	63%	27	75%
Pretrial Assessment	58	16	1	29%	41	71%	39	83%
Supplemental Reentry	58	21	0	36%	37	64%	37	80%
<b>Youth Tools</b>								
Disposition	91	19	2	23%	70	77%	68	94%
Detention	45	12	0	27%	33	73%	33	92%
Diversion	65	10	4	21.5%	51	78.5%	50	96%
Residential	24	6	0	25%	18	75%	16	84%
Reentry	17	5	0	29%	12	71%	12	92%
<b>Total Sample</b>	<b>757</b>	<b>179</b>	<b>19</b>	<b>26%</b>	<b>559</b>	<b>74%</b>	<b>546</b>	<b>90.5%</b>

***Demographic Characteristics of Enrolled Participants***

Tables 3 through 11 include the demographic information of the enrolled participants for each tool.

*Table 3. Community Supervision Tool- Enrolled Participant Demographic Information*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
Gender			263
Male	34.50		
Female	65.20		
Highest degree earned			264
High School/GED	7.20		
Associate's	2.70		
Bachelor's	74.60		
Master's	14.80		
Doctoral	0.80		
Most-used instrument in past year			264
IRAS-CST	84.80		
IRAS-CSST	13.60		
IRAS-PAT	1.50		
Supervisor reviews CST assessments			263
No	74.60		
Yes	25.00		
Supervisor gives feedback on CST assessments			63
No	3.80		
Yes	20.10		
Years of service at current agency	10.91	8.61	263
Years of service in corrections	13.32	9.72	261
Years of experience conducting risk assessments	9.38	7.13	262
Average number of assessments per month	9.96	7.96	244

*Note:* CST = Community Supervision Tool; CSST = Community Supervision Screening Tool; PAT = Pretrial Assessment Tool

*Table 4. Prison Intake Tool- Enrolled Participant Demographic Information*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
Gender			26
Male	18.50		
Female	77.80		
Highest degree earned			27
High School/GED	29.60		
Associate's	11.10		
Bachelor's	51.90		
Master's	7.40		
Most-used instrument in past year			26
IRAS-CST	3.70		
IRAS-PIT	55.60		
IRAS-SRT	37.00		
Supervisor reviews PIT assessments			27
No	70.40		
Yes	29.60		
Supervisor gives feedback on PIT assessments			8
No	3.70		
Yes	25.90		
Years of service at current agency	11.05	9.34	27
Years of service in corrections	11.98	9.04	27
Years of experience conducting risk assessments	5.61	4.47	27
Average number of assessments per month	8.00	11.34	27

*Note:* CST = Community Supervision Tool; PIT = Prison Intake Tool; SRT = Supplemental Reentry Tool

*Table 5. Pretrial Assessment Tool- Enrolled Participant Demographic Information*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
Gender			39
Male	38.50		
Female	61.50		
Highest degree earned			39
High School/GED	7.70		
Associate's	2.60		
Bachelor's	74.40		
Master's	15.40		
Most-used instrument in past year			39
IRAS-CST	46.20		
IRAS-CSST	2.60		
IRAS-PAT	48.70		
IRAS-SRT	2.60		
Supervisor reviews PAT assessments			38
No	66.70		
Yes	30.80		
Supervisor gives feedback on PAT assessments			11
No	5.10		
Yes	23.10		
Years of service at current agency	7.77	6.56	39
Years of service in corrections	9.60	7.33	39
Years of experience conducting risk assessments	6.42	4.66	39
Average number of assessments per month	27.51	51.66	35

*Note:* CST = Community Supervision Tool; CSST = Community Supervision Screening Tool; PAT = Pretrial Assessment Tool; SRT = Supplemental Reentry Tool

*Table 6. Supplemental Reentry Tool- Enrolled Participant Demographic Information*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
Gender			37
Male	45.90		
Female	54.10		
Highest degree earned			37
High School/GED	24.30		
Associate's	16.20		
Bachelor's	48.60		
Master's	10.80		
Most-used instrument in past year			36
IRAS-PIT	18.90		
IRAS-SRT	78.40		
Supervisor reviews SRT assessments			37
No	51.40		
Yes	48.60		
Supervisor gives feedback on SRT assessments			17
No	13.50		
Yes	32.40		
Years of service at current agency	12.26	9.27	37
Years of service in corrections	14.36	9.98	37
Years of experience conducting risk assessments	5.19	4.15	37
Average number of assessments per month	4.31	3.48	36

*Note:* PIT = Prison Intake Tool; SRT = Supplemental Reentry Tool

*Table 7. Disposition Tool- Enrolled Participant Demographic Information*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
Gender			68
Male	29.40		
Female	70.60		
Highest degree earned			68
Bachelor's	79.40		
Master's	20.60		
Most-used instrument in past year			68
IYAS-DIV	10.30		
IYAS-DIS	86.80		
IYAS-DET	1.50		
IYAS-RES	1.50		
Supervisor reviews DIS assessments			68
No	61.80		
Yes	38.20		
Supervisor gives feedback on DIS assessments			26
No	7.40		
Yes	30.90		
Years of service at current agency	14.63	9.09	68
Years of service in corrections	15.63	9.76	67
Years of experience conducting risk assessments	11.27	7.29	66
Average number of assessments per month	6.15	6.93	64

*Note:* DIV = Diversion Tool; DIS = Disposition Tool; DET = Detention Tool; RES = Residential Tool



*Table 8. Detention Tool- Enrolled Participant Demographic Information*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
Gender			33
Male	42.40		
Female	57.60		
Highest degree earned			33
High School/GED	3.00		
Bachelor's	69.70		
Master's	27.30		
Most-used instrument in past year			32
IYAS-DIV	15.20		
IYAS-DIS	30.30		
IYAS-DET	51.50		
Supervisor reviews DET assessments			33
No	54.50		
Yes	45.50		
Supervisor gives feedback on DET assessments			15
No	15.20		
Yes	30.30		
Years of service at current agency	11.84	9.18	33
Years of service in corrections	12.77	9.95	33
Years of experience conducting risk assessments	7.66	5.57	33
Average number of assessments per month	5.13	5.01	29

*Note:* DIV = Diversion Tool; DIS = Disposition Tool; DET = Detention Tool

*Table 9. Diversion Tool- Enrolled Participant Demographic Information*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
Gender			50
Male	20.00		
Female	80.00		
Highest degree earned			50
Bachelor's	82.00		
Master's	18.00		
Most-used instrument in past year			49
IYAS-DIV	62.00		
IYAS-DIS	32.00		
IYAS-RET	4.00		
Supervisor reviews DIV assessments			50
No	66.00		
Yes	34.00		
Supervisor gives feedback on DIV assessments			17
No	6.00		
Yes	28.00		
Years of service at current agency	12.41	8.71	50
Years of service in corrections	14.71	8.48	45
Years of experience conducting risk assessments	10.06	6.74	48
Average number of assessments per month	9.19	9.10	48

*Note:* DIV = Diversion Tool; DIS = Disposition Tool; RET = Reentry Tool

*Table 10. Residential Tool- Enrolled Participant Demographic Information*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
Gender			16
Male	31.30		
Female	68.80		
Highest degree earned			16
Bachelor's	68.80		
Master's	31.30		
Most-used instrument in past year			15
IYAS-DIS	62.50		
IYAS-RES	25.00		
IYAS-RET	6.30		
Supervisor reviews RES assessments			16
No	75.00		
Yes	25.00		
Supervisor gives feedback on RES assessments			4
No	12.50		
Yes	12.50		
Years of service at current agency	12.13	6.71	16
Years of service in corrections	11.51	8.66	15
Years of experience conducting risk assessments	7.67	2.65	16
Average number of assessments per month	3.62	5.53	13

*Note:* DIS = Disposition Tool; RES = Residential Tool; RET = Reentry Tool

*Table 11. Reentry Tool- Enrolled Participant Demographic Information*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
Gender			12
Male	50.00		
Female	50.00		
Highest degree earned			12
Bachelor's	41.70		
Master's	58.30		
Most-used instrument in past year			12
IYAS-DIS	33.30		
IYAS-RET	67.00		
Supervisor reviews RET assessments			12
No	58.30		
Yes	41.70		
Supervisor gives feedback on RET assessments			5
No	16.70		
Yes	25.00		
Years of service at current agency	13.92	7.23	12
Years of service in corrections	16.91	6.85	11
Years of experience conducting risk assessments	9.42	5.33	12
Average number of assessments per month	2.27	1.68	11

*Note:* DIS = Disposition Tool; RET = Reentry Tool

## **Measures**

The present study was administered on-line through a web-based survey software called Qualtrics. Participants were given a link to the survey that corresponded to the tool to which they were assigned. Upon opening the site, participants read the study consent form and agreed to participate. They were first asked to provide basic demographic information about themselves and their professional background (reported above). Next, participants were asked six items assessing their opinion about risk assessment practices and their experiences with assessments. Finally, participants were given three separate vignettes and asked to score all the items of the risk assessment tool to which they were assigned for each vignette. Domain scores and total scores were later computed by researchers using participants' raw item-level data, to ensure participants would not otherwise be penalized for any arithmetic errors when computing domain and total scores.

### ***Creation of Study Vignettes***

The first step in this reliability study was the creation of three unique vignettes for each of the nine risk assessment tools, for a total of 27 different vignettes. Each vignette was carefully crafted to include sufficient information with which to score the assessment. Each vignette included a narrative about the offender's life and index offense and also included some type of collateral information—in some cases, multiple sources were included—e.g., commentary from a known other (family, friend, boss, landlord), self-report tool, file review.

Once the vignettes were drafted, they were shared with certified trainers or master trainers from the State of Indiana, all of whom were IRAS/IYAS trained and/or certified as master trainers of the risk assessment system(s) in their state. These representatives reviewed each of the vignette drafts for coherence with Indiana law and local supervision and assessment practices more generally. Edits were suggested by these representatives, and the vignettes were subsequently finalized by the research team.

### ***Establishment of Criterion Scores***

To establish “criterion” scores to which all study participant's scores would be compared, the vignettes were sent to multiple individuals who were trained in the ORAS/OYAS, one of whom was also certified to train others on these tools. These individuals independently scored each vignette and submitted their scores to the evaluator. The evaluator examined these independently-rated scores and used them to establish the final criterion score. This approach was employed so as to avoid placing too much weight on any one person's ratings.

## **Results: Opinion Survey**

Participants were asked to rate a series of six questions using a Likert-type scale, where higher ratings indicated higher agreement with the statement (e.g., 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree). Results for participants are reported by tool below in Tables 12 through 20. In general, across all tools, users indicate that risk assessments are being used appropriately in their agency and feel confident in using them. Users are less supportive of having more oversight or trainings.

*Table 12. Community Supervision Tool-Agreement Ratings*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
IRAS assessment information is used by me or my agency to determine offenders' supervision levels	4.23	1.26	263
IRAS assessment information is used by me or my agency to determine appropriate interventions based on offenders' risk levels	4.08	1.21	263
I feel confident in my ability to score the IRAS- CST accurately	4.19	1.10	263
The use of the IRAS has made my job easier	3.29	0.98	262
IRAS booster trainings should be offered more frequently	3.22	1.03	263
Supervisors should offer more oversight/provide feedback to those using IRAS assessments	2.92	0.97	263

*Note:* CST = Community Supervision Tool

*Table 13. Prison Intake Tool-Agreement Ratings*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
IRAS assessment information is used by me or my agency to determine offenders' supervision levels	2.93	1.39	27
IRAS assessment information is used by me or my agency to determine appropriate interventions based on offenders' risk levels	4.04	1.26	27
I feel confident in my ability to score the IRAS-PIT accurately	4.19	1.18	27
The use of the IRAS has made my job easier	3.35	1.02	26
IRAS booster trainings should be offered more frequently	2.93	0.87	27
Supervisors should offer more oversight/provide feedback to those using IRAS assessments	3.07	0.68	27

*Note:* PIT = Prison Intake Tool

*Table 14. Pretrial Assessment Tool-Agreement Ratings*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
IRAS assessment information is used by me or my agency to determine offenders' supervision levels	4.23	1.37	39
IRAS assessment information is used by me or my agency to determine appropriate interventions based on offenders' risk levels	4.18	1.30	39
I feel confident in my ability to score the IRAS- PAT accurately	4.46	1.05	39
The use of the IRAS has made my job easier	3.54	1.10	39
IRAS booster trainings should be offered more frequently	3.23	0.87	39
Supervisors should offer more oversight/provide feedback to those using IRAS assessments	2.92	0.74	39

*Note:* PAT = Pretrial Assessment Tool

*Table 15. Supplemental Reentry Tool-Agreement Ratings*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
IRAS assessment information is used by me or my agency to determine offenders' supervision levels	3.41	1.17	37
IRAS assessment information is used by me or my agency to determine appropriate interventions based on offenders' risk levels	3.95	1.08	37
I feel confident in my ability to score the IRAS- SRT accurately	4.08	0.72	37
The use of the IRAS has made my job easier	3.30	0.85	37
IRAS booster trainings should be offered more frequently	3.05	1.05	37
Supervisors should offer more oversight/provide feedback to those using IRAS assessments	2.92	0.98	37

*Note:* SRT = Supplemental Reentry Tool



*Table 16. Disposition Tool-Agreement Ratings*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
IYAS assessment information is used by me or my agency to determine offenders' supervision levels	4.29	1.02	68
IYAS assessment information is used by me or my agency to determine appropriate interventions based on offenders' risk levels	4.12	1.06	68
I feel confident in my ability to score the IYAS-DIS accurately	4.26	0.92	68
The use of the IYAS has made my job easier	3.18	1.18	68
IYAS booster trainings should be offered more frequently	3.12	1.03	68
Supervisors should offer more oversight/provide feedback to those using IYAS assessments	2.84	1.09	68

*Note:* DIS = Disposition Tool

*Table 17. Detention Tool-Agreement Ratings*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
IYAS assessment information is used by me or my agency to determine offenders' supervision levels	4.09	1.18	33
IYAS assessment information is used by me or my agency to determine appropriate interventions based on offenders' risk levels	4.00	1.09	33
I feel confident in my ability to score the IYAS- DET accurately	4.36	0.93	33
The use of the IYAS has made my job easier	3.21	0.86	33
IYAS booster trainings should be offered more frequently	3.24	0.90	33
Supervisors should offer more oversight/provide feedback to those using IYAS assessments	3.06	0.66	33

*Note:* DET = Detention Tool

*Table 18. Diversion Tool-Agreement Ratings*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
IYAS assessment information is used by me or my agency to determine offenders' supervision levels	4.28	1.01	50
IYAS assessment information is used by me or my agency to determine appropriate interventions based on offenders' risk levels	4.14	1.03	50
I feel confident in my ability to score the IYAS- DIV accurately	4.44	0.76	50
The use of the IYAS has made my job easier	3.20	0.95	50
IYAS booster trainings should be offered more frequently	2.96	0.95	50
Supervisors should offer more oversight/provide feedback to those using IYAS assessments	2.60	0.93	50

*Note:* DIV = Diversion Tool

*Table 19. Residential Tool-Agreement Ratings*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
IYAS assessment information is used by me or my agency to determine offenders' supervision levels	4.38	0.72	16
IYAS assessment information is used by me or my agency to determine appropriate interventions based on offenders' risk levels	4.06	0.85	16
I feel confident in my ability to score the IYAS- RES accurately	4.38	0.50	16
The use of the IYAS has made my job easier	3.06	1.06	16
IYAS booster trainings should be offered more frequently	3.06	1.06	16
Supervisors should offer more oversight/provide feedback to those using IYAS assessments	2.75	1.00	16

*Note:* RES = Residential Tool

*Table 20. Reentry Tool-Agreement Ratings*

	<b>% / Mean</b>	<b>SD</b>	<b>Valid N</b>
IYAS assessment information is used by me or my agency to determine offenders' supervision levels	3.75	1.42	12
IYAS assessment information is used by me or my agency to determine appropriate interventions based on offenders' risk levels	3.75	1.22	12
I feel confident in my ability to score the IYAS- RET accurately	3.58	1.38	12
The use of the IYAS has made my job easier	3.00	1.13	12
IYAS booster trainings should be offered more frequently	2.92	1.00	12
Supervisors should offer more oversight/provide feedback to those using IYAS assessments	2.50	1.00	12

*Note:* RET = Reentry Tool

## **Results: Reliability**

### **Explanation of Research Findings and Instructions for Interpretation**

The results for each tool are presented separately below in a series of four tables per tool. Each table is accompanied by a brief narrative summary and commentary on/interpretation of the findings.

The first table for each tool reports the average (mean, or “*M*”) and standard deviation (“*SD*”) units (i.e., risk assessment points) by which participants on that specific tool deviated from the true, or criterion score at the total score level. If the tool has individual domain scores, results for the amount of deviation for each domain are also reported. This “deviation” index offers the raw value estimation by which IRAS/IYAS users depart from the criterion. Because each user’s individual rating was subtracted from the criterion score, scores can be both positive and negative. Negative values indicate that the average users’ score was higher than the criterion, whereas positive values indicate that the average users’ score was lower than the criterion.

Because the total score and each individual domain score differ in terms of their range of possible values, it is not useful to compare domain to domain per se. Rather, it is insightful to examine the results within each domain individually and within the total score individually, to get a sense of how far off users are from the value they should be scoring. This index can be useful in identifying what domain may be problematic for users to score, particularly if a trend is observed across multiple vignettes within each tool. This index is also useful in instances where the intraclass correlation coefficient or “ICC” (reported in Table 2 for each tool) may be unreliable for one reason or another (see details under description of Table 2).

The second table for each tool reports the average amount by which users agree with the criterion scores, as measured with the ICC. ICC values range from 0 to 1.0, and higher values indicate better agreement to the criterion score. ICC values of 0.50 indicate a level of agreement that is no better or worse than chance, and therefore any ICC value between 0.00-0.50 is considered “poor” agreement; 0.51-0.75 is considered “moderate” agreement; 0.76-0.90 is considered “good” agreement; 0.91-1.00 is considered “excellent” agreement (Koo & Li, 2016). ICC values were computed by examining each user’s score within each domain, and, separately, the total score across all three vignettes. In order to compute the ICC, therefore, users had to complete/score at least two of the three vignettes. This is mentioned because there are a couple

of instances (reported by tool below) where the user/participant only scored one of the three vignettes; in these cases, the user was excluded from the analyses.

In some instances when negative covariation is observed (i.e., when scores are correlated but in the opposite—or wrong direction; e.g., Case 1 scores Vignette A as 10, Vignette B as 20, and Vignette C as 25 and Case 2 scores Vignette A as 25, Vignette B as 20 and Vignette C as 10), a negative ICC value is computed; however, any observed negative ICC values were recoded to a zero value before subsequent analyses were performed. This is an acceptable strategy for data analysis with ICC values, because it still accurately reflects poor or no agreement without disproportionately skewing the data—and therefore the results—of subsequent analyses (see Bartko, 1976).

Given that this study employed random selection of a sample of users, and given that the same set of users for each tool also all rated the same vignettes, “two-way random” ICC and “absolute agreement” criteria were employed for these analyses. Additionally, because the focus is on individual users’ scores as they compare to a criterion score, the “single measures” results were recorded and analyzed. In Table 2, the mean (*M*) and standard deviation (*SD*) ICC-to-criterion values are reported at the total score level and, when relevant, for individual domain scores for each tool.

The third table for each tool reports the correlation (i.e., association) of users’ ICC-to-criterion values of each tool’s total score with (a) years of experience administering risk assessments, (b) years of experience administering the specific tool that is the focus of the analysis, and (c) the total estimated number of assessments completed per month within the past year, of that particular assessment tool being examined. In general, positive correlations would be anticipated across the board, where people who have more experience with the tool and with risk assessment in general will perform better overall.

The correlation coefficient “*r*” ranges from -1.0 to +1.0, where values closer to -1.0 and +1.0 indicate a stronger association, and values closer to 0 indicate a weaker association. Cohen’s (1988) effect size conventions for correlations are 0.10 = “small”, 0.30 = “medium”, and 0.50 = “large”. Negative correlation coefficients indicate that as experience increases, performance decreases. If and when this is observed in the results, it should be cause for concern. Nevertheless, it should be noted that in all but one tool (IYAS-Reentry), no significant correlations were observed with these indices. As such, it is still insightful to examine the direction (+ vs. -) and strength of the relationships observed.

The fourth table for each tool presents the inter-rater reliability estimates using the ICC. For these analyses, the “two-way random” and “consistency” criteria were employed in the analysis. Results are presented for the total score, as well as for any domain scores, if domains exist within a given tool. These results indicate how well users on a specific tool agree with one another. This index is examined and reported, because it is important to determine whether users are consistent with one another. Ideally, users of the same tool are rating the same case in the same manner. It is important to note, however, that users can be consistent with one another but be inaccurate—i.e., “reliably wrong”. For this reason, it is important to examine interrater reliability and interpret the findings alongside reliability-to-criterion scores. An ideal scenario for the results would be to observe a high interrater ICC and high reliability-to-criterion ICC.

In Table 4, the ICC value is reported, along with the lower and upper value ICCs in parenthesis, to indicate the full range of scores obtained. The “Single Measures” value is the most useful for the purposes of this report. This indicates the average amount by which a randomly selected user agrees/disagrees with other IRAS/IYAS users. The “Average Measures” is also reported. This value reflects the average of all users’ interrater reliability. It will almost always be a higher value than reported in the “Single Measures”. Because this value increases when more users have higher ICC values, it provides a more inflated interpretation of agreement between users. Thus, it is recommended that the Single Measures be of focus in interpretation in this table.

## IRAS Reliability Results

### *Community Supervision Tool (CST)*

Table 21. CST Average Deviation from Criterion Scores; *M (SD)*

	Joe	Julia	Remy
Criminal History Range = 0-8	0.14 (1.06)	-0.23 (0.53)	0.98 (0.62)
Education, Employment, and Financial Situation Range = 0-6	0.05 (0.49)	0.10 (0.62)	0.10 (0.33)
Family and Social Support Range = 0-5	-0.02 (0.65)	0.14 (0.51)	0.99 (0.33)
Neighborhood Problems Range = 0-3	-0.02 (0.16)	0.03 (0.24)	0.06 (0.25)
Substance Use Range = 0-6	-1.03 (0.91)	-0.70 (0.96)	-1.20 (0.80)
Peer Associations Range = 0-8	-1.65 (1.15)	-0.58 (0.62)	0.23 (0.59)
Criminal Attitudes and Behavioral Patterns Range = 0-13	0.36 (1.44)	-0.33 (1.01)	0.28 (1.04)
Total Score Range = 49	-2.14 (2.62)	-1.56 (1.97)	1.43 (1.71)

The highest total score deviation was observed with the “Joe” vignette. Additionally, on at least two of three vignettes, users showed deviation from the criterion score by more than a point on the “Substance Use” domain, suggesting that this may be more challenging for users to score. Particular attention to this domain and the items that comprise it in future trainings is suggested. Given the full range of scores possible at the total score level (0-49), the average deviation at this level does not appear too concerning. Most users seem to be arriving at or near the criterion total score, likely ultimately placing the hypothetical offenders being scored in the same risk assessment category.

*Table 22. CST Average Criterion ICC across 3 Vignettes; M (SD)*

Criminal History	0.85 (0.15)
Education, Employment, and Financial Situation	0.95 (0.09)
Family and Social Support	0.11 (0.22)
Neighborhood Problems	0.99 (0.04)
Substance Use	0.70 (0.22)
Peer Associations	0.73 (0.22)
Criminal Attitudes and Behavioral Patterns	0.68 (0.23)
Total Score	0.93 (0.08)

*Note: 5 cases were excluded from analysis due to incomplete scores*

CST users demonstrate excellent agreement with criterion for the total score, with 93% of users meeting good to excellent agreement. There was moderate agreement with the criterion score in the “Substance Use”, “Peer Associations”, and “Criminal Attitudes and Behavioral Patterns” domains; good agreement with criterion scores in the “Criminal History” domain; and excellent agreement with criterion in the “Education, Employment, and Financial Situation” and “Neighborhood Problems” domains. The only domain that appeared problematic was the “Family and Social Support” domain. However, upon closer inspection of the raw data, it was determined that there was very low intra-individual variability across the users’ scores for the three vignettes. That is, users rated all 3 vignettes with the same “Family and Social Support” domain score. Because the ICC considers within user and between user variability, the ICC was adversely impacted here. Rather than interpreting the ICC value for this domain, it is recommended that attention be focused on the CST Table 1 output for the “Family and Social Support” domain.

*Table 23. Correlation of CST Criterion ICC Total Score with Rater Characteristics*

	<i>r</i>
Years working with risk assessment tools	0.09
Years working with CST	0.07
Total # of CST assessments completed per month, past year	-0.00



Small, non-significant associations are observed between years of experience and performance on the CST assessment tool.

*Table 24. CST Inter-rater Reliability ICC (CI Low, CI High)*

	Single Measures	Average Measures
Criminal History	0.31 (0.08, 1.00)	0.99 (0.96, 1.00)
Education, Employment, and Financial Situation	0.67 (0.28, 1.00)	0.99 (0.99, 1.00)
Family and Social Support	0.03 (0.00, 0.97)	0.89 (0.46, 1.00)
Neighborhood Problems	0.99 (0.95, 1.00)	1.00 (1.00, 1.00)
Substance Use	0.82 (0.47, 1.00)	0.99 (0.99, 1.00)
Peer Association	0.71 (0.32, 1.00)	0.99 (0.99, 1.00)
Criminal Attitudes and Behavioral Patterns	0.74 (0.36, 1.00)	0.99 (0.99, 1.00)
Total Score	0.89 (0.61, 1.00)	1.00 (0.99, 1.00)

*Note: 5 cases were excluded from analysis due to incomplete scores*

Poor interrater reliability was observed for the “Criminal History” domain. This finding is especially concerning because these items—more than any other items on the IRAS-CST—are the most verifiable and therefore should generally show the highest reliability to criterion and interrater reliability across the board. Good interrater reliability was observed for the “Education/Employment”, “Substance Use”, Peer Associations”, and “Criminal Attitudes and Behavioral Patterns” domains, as well as the total score; Excellent interrater reliability was observed for the “Neighborhood Problems” domain. As also observed in Table 2, the “Family and Social Support” domain ICC value was adversely impacted by low within subject variance (i.e., many users incorrectly rated all three vignettes with the same score in this domain). In this instance, it would be appropriate to interpret the “Average Measures” value, which suggests good interrater reliability.

**Supplemental Reentry Tool (SRT)**

Table 25. SRT Average Deviation from Criterion Scores; *M (SD)*

	Alex	Chad	Mary
Criminal History Range = 0-13	-0.05 (0.62)	-0.03 (0.73)	-0.54 (1.04)
Education, Employment, and Social Support Range = 0-9	1.89 (0.70)	0.00 (0.88)	0.32 (1.13)
Substance Use and Mental Health Range = 0-4	-0.30 (0.74)	0.21 (0.41)	0.38 (0/68)
Criminal Attitudes and Behavioral Patterns Range = 0-19	4.14 (2.70)	2.54 (3.21)	0.81 (1.73)
Total Score Range = 0-45	5.68 (3.09)	2.73 (3.66)	0.97 (2.80)

The highest total score level deviation was observed with the “Alex” vignette. Additionally, on at least two of three vignettes, users showed deviation from the criterion score by more than a point on the “Attitudes” domain, suggesting that this may be more challenging for users to score.

Particular attention to this domain and the items that comprise it in future trainings is suggested.

Given the full range of scores possible for the total score (0-49), the average deviation for the total score is somewhat concerning. In fact, it is much higher than observed on any other tool in this report. Additionally, the average total score deviation amount (particularly with the “Alex” vignette) could potentially place an offender in a different risk category, resulting in of offenders to supervision and resources.

*Table 26. SRT Average Criterion ICC across 3 Vignettes; M (SD)*

Criminal History	0.99 (0.05)
Education, Employment, and Social Support	0.81 (0.18)
Substance Use and Mental Health	0.61 (0.43)
Criminal Attitudes and Behavioral Patterns	0.77 (0.20)
Total Score	0.88 (0.10)

SRT users demonstrate good overall agreement to criterion for the total score, with 92% of users meeting good to excellent agreement with criterion scores. SRT users demonstrate moderate agreement with the criterion score in the “Substance Use and Mental Health” domain; good agreement with criterion scores in the “Criminal Attitudes and Behavioral Patterns” and “Education, Employment, and Social Support” domains; and excellent agreement with criterion in the “Criminal History” domain.

*Table 27. Correlation of SRT Criterion ICC Total Score with Rater Characteristics*

	<i>R</i>
Years working with risk assessment tools	-0.12
Years working with SRT	-0.20
Total # of SRT assessments completed per month, past year	0.18

Small, non-significant associations are observed between years of experience and performance on the SRT assessment tool.

Table 28. SRT Inter-rater Reliability ICC (CI Low, CI High)

	Single Measures	Average Measures
Criminal History	0.99 (0.95, 1.00)	1.00 (0.99, 1.00)
Education, Employment, and Social Support	0.81 (0.52, 1.00)	0.99 (0.98, 1.00)
Substance Use and Mental Health	0.29 (0.08, 0.94)	0.94, (0.75, 1.00)
Criminal Attitudes and Behavioral Patterns	0.78 (0.47, 1.00)	0.99 (0.97, 1.00)
Total Score	0.89 (0.68, 1.00)	0.99 (0.99, 1.00)

Poor interrater reliability was observed for the “Substance Use and Mental Health” domain, and this is consistent with the slightly lower (compared to other domains) reliability-to-criterion *ICC* observed in Table 2; good interrater reliability was observed for the “Education, Employment, and Social Support”, and “Criminal Attitudes and Behavioral Patterns” domains, as well as the total score; excellent interrater reliability was observed for the “Criminal History” domain.

***Prison Intake Tool (PIT)***

*Table 29. PIT Average Deviation from Criterion Scores; M (SD)*

	Cameron	Erin	Rachel
Criminal History Range = 0-11	0.17 (1.23)	0.07 (1.25)	-0.82 (0.94)
Education, Employment, and Financial Situation Range = 0-7	0.70 (1.51)	0.14 (1.16)	-0.11 (1.03)
Family and Social Support Range = 0-6	-0.07 (1.08)	0.17 (1.07)	-1.03 (0.88)
Substance Use and Mental Health Range = 0-5	-0.83 (0.70)	-1.21 (0.82)	0.07 (0.71)
Criminal Attitudes and Behavioral Patterns Range = 0-11	0.10 (1.52)	-0.79 (1.90)	-0.61 (1.55)
Total Score Range = 0-40	0.07 (4.08)	-1.62 (4.60)	-2.50 (2.71)

The highest total score deviation was observed with the “Rachel” vignette. Results suggest that minor improvements could be made in the “Substance Use and Mental Health” and “Criminal Attitudes and Behavioral Patterns” domains. Overall, no major trend-level concerns were noted in the domains or with the total score.

*Table 30. PIT Average Criterion ICC across 3 Vignettes; M (SD)*

Criminal History	0.88 (0.27)
Education, Employment, and Financial Situation	0.87 (0.27)
Family and Social Support	0.40 (0.39)
Substance Use and Mental Health	0.52 (0.19)
Criminal Attitudes and Behavioral Patterns	0.86 (0.23)
Total Score	0.92 (0.19)

*Note: 1 case was excluded from analysis due to incomplete scores*

PIT users demonstrate excellent overall agreement to criterion for the total score, with 96% of users meeting good to excellent agreement. PIT users demonstrate moderate agreement with the criterion score in the “Substance Use and Mental Health” domain and good agreement with criterion scores in the “Criminal History”, “Education, Employment, and Financial Situation”, and “Criminal Attitudes and Behavioral Patterns” domains. Poor agreement with criterion was observed for the “Family and Social Support” domain, which is consistent with the interrater reliability findings reported in PIT Table 4. It is recommended that more attention be given to the “Family and Social Support” and “Substance Use and Mental Health” domains and the items that comprise them during trainings.

*Table 31. Correlation of PIT Criterion ICC Total Score with Rater Characteristics*

	<i>R</i>
Years working with risk assessment tools	0.20
Years working with PIT	0.29
Total # of PIT assessments completed per month, past year	0.01

Small, non-significant associations are observed between years of experience and performance on the PIT assessment tool.

*Table 32. PIT Inter-rater Reliability ICC (CI Low, CI High)*

	Single Measures	Average Measures
Criminal History	0.50 (0.12, 1.00)	0.96 (0.79, 1.00)
Education, Employment, and Financial Situation	0.78 (0.36, 1.00)	0.99 (0.94, 1.00)
Family and Social Support	-0.02 (-0.04, 0.95)	-0.98 (-10.23, 1.00)
Substance Use and Mental Health	0.60 (0.19, 1.00)	0.98 (0.86, 1.00)
Criminal Attitudes and Behavioral Patterns	0.58 (0.17, 1.00)	0.97 (0.85, 1.00)
Total Score	0.54 (0.15, 1.00)	0.97 (0.82, 1.00)

*Note: 1 case was excluded from analysis due to incomplete scores*

Moderate (but low moderate) interrater reliability was observed for the “Criminal History”, “Substance Use and Mental Health”, and “Criminal Attitudes and Behavioral Patterns” domains, and the total score; good interrater reliability was observed for the “Education, Employment, and Financial Situation” domain; poor interrater reliability was observed for the “Family and Social Support” domain. Upon closer examination of the raw data and as indicated by the negative *ICC* values, there is a lot of variability—including negative covariation—in the “Family and Social Support” domain. This domain appears to be particularly problematic for users, given its low interrater reliability and reliability to the criterion score. Thus, it is suggested that this domain and the items that comprise it also be given special added attention in future trainings.

***Pretrial Assessment Tool (PAT)***

*Table 33. PAT Average Deviation from Criterion Scores; M (SD)*

	Emily	Juan	Leon
Total Score <sup>a</sup> Range = 0-9	-1.05 (0.69)	0.03 (0.28)	-0.49 (0.60)

<sup>a</sup> Because there are no scale scores for the PAT, results are reported only for the total score.

The highest total score deviation was observed with the “Emily” vignette. Results suggest that minor improvements could be made in the rating of the tool, but deviation values do not pose a major concern.

*Table 34. PAT Criterion ICC; M (SD)*

Total Score	0.94 (0.06)
-------------	-------------

PAT users demonstrate excellent reliability to the criterion score, and 97% of users falling in good to excellent reliability classifications.

*Table 35. Correlation of PAT Criterion ICC Total Score with Rater Characteristics*

	<i>r</i>
Years working with risk assessment tools	0.16
Years working with PAT	0.15
Total # of PAT assessments completed per month, past year	0.06

Small, non-significant associations are observed between years of experience and performance on the PAT assessment tool



*Table 36. PAT Interrater Reliability ICC (CI Low, CI High)*

	Single Measures	Average Measures
Total Score <sup>a</sup>	0.92 (0.75, 1.00)	0.99 (0.99, 1.00)

*<sup>a</sup> Because there are no scale scores for the PAT, results are reported only for the total score.*

Excellent interrater reliability is observed for the PAT.

## IYAS Reliability Results

### *Disposition Tool (DIS)*

Table 37. DIS Average Deviation from Criterion Scores; *M (SD)*

	Laura	Jack	Caleb
Juvenile Justice History Range = 0-3	0.04 (0.40)	-0.09 (0.34)	-0.08 (0.37)
Family and Living Arrangements Range = 0-6	-1.21 (0.80)	0.08 (1.02)	-0.38 (0.85)
Peers and Social Support Network Range = 0-6	0.03 (0.34)	0.26 (0.69)	0.21 (0.77)
Education and Employment Range = 0-4	0.63 (0.60)	0.00 (0.31)	0.21 (0.45)
Prosocial Skills Range = 0-3	0.28 (0.54)	-0.15 (0.48)	0.29 (0.73)
Substance Abuse, Mental Health, & Personality Range = 0-6	0.24 (0.52)	0.49 (0.56)	0.30 (0.66)
Values, Beliefs, & Attitudes Range = 0-5	0.37 (0.60)	-0.21 (0.41)	-0.22 (0.41)
Total Score Range = 0-33	0.38 (1.64)	0.39 (2.23)	1.79 (2.56)

The highest total score deviation was observed with the “Caleb” vignette. Results suggest that minor improvements could be made in the rating of the tool, but deviation values do not pose a major concern. Even so, trends are observed across the vignettes in almost every domain except for the “Juvenile Justice History” domain. This suggests that Disposition Tool users across the state may benefit from some additional tune-up trainings and/or reminders to adhere to scoring criteria and make use of their scoring guide whenever possible.

*Table 38. DIS Criterion ICC; M (SD)*

Juvenile Justice History	0.84 (0.30)
Family and Living Arrangements	0.52 (0.19)
Peers and Social Support Network	0.92 (0.16)
Education and Employment	0.82 (0.20)
Prosocial Skills	0.92 (0.19)
Substance Abuse, Mental Health, & Personality	0.63 (0.31)
Values, Beliefs, & Attitudes	0.89 (0.14)
Total Score	0.95 (0.13)

*Note: 2 cases were excluded from analysis due to incomplete scores*

DIS users demonstrate excellent overall agreement to criterion for the total score, with 96% of users meeting good to excellent agreement. DIS users demonstrate moderate agreement with the criterion score in the “Family & Living Arrangements” and “Substance Use, Mental Health, & Personality” domains; good agreement with criterion scores in the “Juvenile Justice History”, “Education and Employment”, and “Values, Beliefs, & Attitudes” domains; and excellent agreement was observed for “Peers and Social Support Network” and “Prosocial Skills” domains. Minor improvements are recommended for the “Family & Living Arrangements” and “Substance Use, Mental Health, & Personality” domains, but no major concerns arise from these results.

*Table 39. Correlation of DIS Criterion ICC Total Score with Rater Characteristics*

	<i>r</i>
Years working with risk assessment tools	0.15
Years working with DIS	0.17
Total # of DIS assessments completed per month, past year	-0.01

Small, non-significant associations are observed between years of experience and performance on the DIS assessment tool.

Table 40. DIS Interrater Reliability

	Single Measures	Average Measures
Juvenile Justice History	0.82 (0.48, 1.00)	0.99 (0.98, 1.00)
Family and Living Arrangements	0.79 (0.41, 1.00)	0.99 (0.98, 1.00)
Peers and Social Support Network	0.94 (0.75, 1.00)	0.99 (0.99, 1.00)
Education and Employment	0.79 (0.42, 1.00)	0.99 (0.98, 1.00)
Prosocial Skills	0.90 (0.63, 1.00)	0.99 (0.99, 1.00)
Substance Abuse, Mental Health, & Personality	0.65 (0.25, 1.00)	0.99 (0.96, 1.00)
Values, Beliefs, & Attitudes	0.96 (0.83, 1.00)	0.99 (0.99, 1.00)
Total Score	0.96 (0.83, 1.00)	0.99 (0.99, 1.00)

*Note: 2 cases were excluded from analysis due to incomplete scores*

Moderate interrater reliability was observed for the “Substance Abuse, Mental Health, & Personality” domain; good interrater reliability was observed for the “Juvenile Justice History”, “Family & Living Arrangements”, “Education & Employment”, and “Prosocial Skills” domains; excellent interrater reliability was observed for “Peers & Social Support Network” and the “Values, Beliefs, & Attitudes” domains, and for the total score. Given the slightly lower *ICC* values in the “Substance Abuse, Mental Health, & Personality” domain for both interrater reliability and reliability-to-criterion indices, it is recommended that more attention be given in trainings to this domain and the items that comprise it.

***Detention Tool (DET)***

*Table 41. DET Average Deviation from Criterion Scores; M (SD)*

	Anakin	Brendan	Melonie
Total Score <sup>a</sup> Range = 0-7	0.69 (-0.42)	0.00 (0.00)	0.34 (0.57)

<sup>a</sup> Because there are no scale scores for the DET, results are reported only for the total score.

The vignette with the largest average total raw score deviation is “Anakin”. Given the range of possible scores on the DET, an average deviation of less than one point is not too concerning.

*Table 42. DET Criterion ICC; M (SD)*

Total Score	0.94 (0.12)
-------------	-------------

DET users demonstrate excellent reliability to the criterion score, and 88% of users meet good to excellent levels of reliability to criterion.

*Table 43. Correlation of DET Criterion ICC Total Score with Rater Characteristics*

	<b><i>r</i></b>
Years working with risk assessment tools	0.22
Years working with DET	0.19
Total # of DET assessments completed per month, past year	0.34

Small non-significant associations are observed between years of experience and performance on the DET assessment tool; a medium non-significant association is observed between the criterion score and the total # of assessments completed per month in the past year.

Table 44. DET Interrater Reliability ICC (CI Low, CI High)

	Single Measures	Average Measures
Total Score <sup>a</sup>	0.56 (0.23, 0.98)	0.98 (0.91, 1.00)

<sup>a</sup> Because there are no scale scores for the DET, results are reported only for the total score.

There is moderate interrater reliability on the DET. Though it is rare to have lower interrater reliability but high reliability to criterion, such a result suggests that there are likely a few users whose scoring is quite disparate from the rest of the group, which effectively brings the *ICC* value down for interrater reliability. An examination of the raw data suggests that this is, in fact, the case. Specifically, there were three users who performed poorly on scoring the “Anakin” vignette. When these users were removed from the analyses, the interrater reliability increased to  $ICC = 0.85 (0.59, 1.00)$ , which is considered good interrater reliability.

***Diversion Tool (DIV)***

*Table 45. DIV Average Deviation from Criterion Scores; M (SD)*

	Bart	Dina	John
Total Score <sup>a</sup> Range = 0-7	0.02 (0.24)	0.38 (0.77)	0.34 (0.52)

<sup>a</sup> Because there are no scale scores for the DIV, only total score results are reported.

The vignette with the greatest raw total score deviation was “Dina”. Given the total possible range of scores on this tool, the average deviation of less than one-half of one point is not concerning.

*Table 46. DIV Criterion ICC; M (SD)*

Total Score	0.94 (0.11)
-------------	-------------

DIV users’ reliability to criterion is considered excellent, and 96% of users meet good to excellent standards for reliability to criterion.

*Table 47. Correlation of DIV Criterion ICC Total Score with Rater Characteristics*

	<i>r</i>
Years working with risk assessment tools	-0.11
Years working with DIV	0.05
Total # of DIV assessments completed per month, past year	0.11

Small, non-significant associations are observed between years of experience and performance on the DIV assessment tool.

*Table 48. DIV Interrater Reliability ICC (CI Low, CI High)*

	Single Measures	Average Measures
Total Score <sup>a</sup>	0.90 (0.71, 1.00)	0.99 (0.99, 1.00)

*<sup>a</sup> Because there are no scale scores for the DIV, only total score results are reported.*

Interrater reliability on the DIV is considered good; even the lower bound of scores falls within moderate reliability range.



**Residential Tool (RES)**

*Table 49. RES Average Deviation from Criterion Scores; M (SD)*

	Alberto	Danielle	Ryan
Juvenile Justice History Range = 0-4	-0.13 (0.34)	0.00 (0.00)	-0.06 (0.25)
Family and Living Arrangements Range = 0-3	-0.56 (0.96)	0.00 (0.00)	1.00 (0.82)
Peers and Social Support Network Range = 0-7	0.00 (0.00)	-0.75 (0.45)	-0.25 (0.68)
Education and Employment Range = 0-3	0.19 (0.40)	0.00 (0.00)	0.00 (0.37)
Prosocial Skills Range = 0-4	-0.44 (0.81)	-0.38 (0.62)	0.38 (0.50)
Substance Use, Mental Health, and Personality Range = 0-8	-0.06 (0.44)	-0.19 (1.05)	0.38 (0.62)
Values, Beliefs, and Attitudes Range = 0-5	-0.25 (0.45)	-0.06 (0.25)	0.06 (0.25)
Total Score Range = 0-34	-1.25 (2.34)	-1.38 (1.50)	1.50 (1.71)

The vignette with the largest average total raw score deviation is “Ryan”. Given the range of possible scores on the RES, the average total score deviation values do not pose any serious concern. There are some minor trends observed across several vignettes with respect to higher relative deviation values in the “Family and Living Arrangements”, “Peers and Social Support Network”, “Prosocial Skills”, and “Substance Use, Mental Health, and Personality” domains. It is recommended that these domains and the items that comprise them be given some added attention in future trainings.

*Table 50. RES Average Criterion ICC across 3 Vignettes; M (SD)*

Juvenile Justice History	0.98 (0.07)
Family and Living Arrangements	0.73 (0.29)
Peers and Social Support Network	0.81 (1.15)
Education and Employment	0.95 (0.08)
Prosocial Skills	0.91 (0.12)
Substance Use, Mental Health, and Personality	0.95 (0.08)
Values, Beliefs, and Attitudes	0.99 (0.01)
Total Score	0.96 (0.05)

RES users demonstrate excellent overall agreement to criterion for the total score, with 100% of users meeting good to excellent agreement to criterion. RES users show good reliability to criterion scores in the “Family and Living Arrangements” and “Peers and Social Support Network” domains and excellent reliability to criterion in the remaining domains.

*Table 51. Correlation of RES Criterion ICC Total Score with Rater Characteristics*

	<i>R</i>
Years working with risk assessment tools	-0.03
Years working with RES	-0.21
Total # of RES assessments completed per month, past year	-0.03

Small non-significant but negative associations are observed between years of experience and performance on the RES assessment tool.

Table 52. RES Inter-rater Reliability ICC (CI Low, CI High)

	Single Measures	Average Measures
Juvenile Justice History	0.96 (0.86, 1.00)	0.99 (0.99, 1.00)
Family and Living Arrangements	0.66 (0.29, 1.00)	0.97 (0.87, 1.00)
Peers and Social Support Network	0.35 (0.07, 0.96)	0.90 (0.56, 1.00)
Education and Employment	0.91 (0.71, 1.00)	0.99 (0.98, 1.00)
Prosocial Skills	0.89 (0.65, 1.00)	0.99 (0.97, 1.00)
Substance Use, Mental Health, and Personality	0.95 (0.81, 1.00)	0.99 (0.99, 1.00)
Values, Beliefs, and Attitudes	0.99 (0.96, 1.00)	0.99 (0.99, 1.00)
Total Score	0.97 (0.88, 1.00)	0.99 (0.99, 1.00)

Interrater reliability on the RES was poor for the “Peers and Social Support Network” domain; moderate for “Family and Living Arrangements”; good for “Prosocial Skills”; and excellent for “Juvenile Justice History”, “Education and Employment”, “Substance Use, Mental Health, & Personality” and “Values, Beliefs, & Attitudes” domains and the total score.

***Reentry Tool (RET)***

*Table 53. RET Average Deviation from Criterion Scores; M (SD)*

	Steve	Lionel	Bernie
Juvenile Justice History Range = 0-7	-0.08 (0.29)	0.45 (0.93)	0.73 (1.27)
Family and Living Arrangements Range = 0-4	0.00 (0.43)	0.09 (0.94)	-0.45 (1.13)
Peers and Social Support Network Range = 0-9	-0.08 (0.29)	0.64 (2.01)	2.00 (2.10)
Education and Employment Range = 0-4	0.00 (0.00)	0.36 (0.92)	0.36 (0.67)
Prosocial Skills Range = 0-4	-0.17 (0.58)	1.36 (1.36)	0.36 (1.63)
Substance Use, Mental Health, and Personality Range = 0-7	-0.08 (0.29)	0.55 (1.69)	0.00 (1.18)
Values, Beliefs, and Attitudes Range = 0-7	-0.58 (1.00)	0.27 (1.79)	1.55 (2.07)
Total Score Range = 0-42	-1.00 (1.71)	3.73 (7.56)	4.55 (8.10)

*Note: Initial inspection of the data revealed one case in which the participant did not seem to actually properly or reliably score the vignettes, as all scores entered were “0” for every item across all three vignettes; thus, this case was removed from all analyses.*

The largest total score deviation was observed with the “Bernie” vignette. The total point deviation could, in theory, likely place someone in a different risk category—a finding that is cause for concern. Additionally, larger deviations were found across two of the three vignettes in “Juvenile Justice History”, a domain that is the most verifiable of the domains to rate. Additional notable deviation trends are observed across vignettes in the “Peers and Social Support Network”, “Education and Employment”, “Prosocial Skills”, and “Values, Beliefs, & Attitudes” domains. Taken together, these findings suggest that RET users could benefit from ongoing

booster trainings and supervision of scores, to ensure they are rating items consistent with scoring guidelines.

*Table 54. RET Average Criterion ICC across 3 Vignettes; M (SD)\**

Juvenile Justice History	0.97 (0.03)
Family and Living Arrangements	0.61 (0.20)
Peers and Social Support Network	0.93 (0.08)
Education and Employment	0.90 (0.11)
Prosocial Skills	0.74 (0.33)
Substance Use, Mental Health, and Personality	0.96 (0.07)
Values, Beliefs, and Attitudes	0.92 (0.10)
Total Score	0.98 (0.02)

RET users demonstrate excellent overall agreement to criterion for the total score, with 91% of users meeting good to excellent agreement to criterion. Moderate reliability to criterion scores was observed in the “Family and Living Arrangements” and “Prosocial Skills” domain; good reliability was observed in the “Education and Employment” domain; and excellent reliability to criterion scores was observed for the remaining domain scores.

*Table 55. Correlation of RET Criterion ICC Total Score with Rater Characteristics*

	<i>r</i>
Years working with risk assessment tools	-0.70*
Years working with RET	-0.16
Total # of RET assessments completed per month, past year	0.31

\*  $p < .05$

A small but nonsignificant negative association was observed between years of experience working with the RET and reliability to criterion scores; a medium positive but non-significant correlation was observed between the total # of RET assessments completed per month in the past year and reliability to criterion scores; and a large negative statistically significant association was found between the years working with risk assessment tools in general and criterion scores.

*Table 56. RET Inter-rater Reliability ICC (CI Low, CI High)*

	Single Measures	Average Measures
Juvenile Justice History	0.85 (0.54, 1.00)	0.98 (0.93, 1.00)
Family and Living Arrangements	0.42 (0.08, 0.97)	0.89 (0.50, 1.00)
Peers and Social Support Network	0.84 (0.53, 1.00)	0.98 (0.92, 1.00)
Education and Employment	0.71 (0.32, 0.99)	0.96 (0.84, 1.00)
Prosocial Skills	0.62 (0.23, 0.99)	0.95 (0.77, 1.00)
Substance Use, Mental Health, and Personality	0.83 (0.50, 1.00)	0.98 (0.92, 1.00)
Values, Beliefs, and Attitudes	0.82 (0.48, 0.99)	0.98 (0.91, 1.00)
Total Score	0.88 (0.61, 1.00)	0.99 (0.95, 1.00)

Poor interrater reliability was observed for the “Family and Living Arrangements” domain; moderate interrater reliability was observed for the “Prosocial Skills” domain; and good interrater reliability was observed for the remaining domains and total score.

## **Conclusions & Recommendations**

The Indiana Risk Assessment Reliability Study was a successful undertaking. There was a strong partnership between the state and the University of Cincinnati and excellent leadership and coordination from within the state of Indiana (vis-à-vis Michelle Goodman). These conditions were ideal for ensuring outstanding participation rates. Additionally, the results are sound, and the study design allows the findings to generalize to the larger body of active IYAS and IRAS users within the state of Indiana.

The results from this reliability study were quite promising. Certified and authorized users of the IRAS and IYAS tools in the State of Indiana are able to rate the tool in a reliable and accurate manner. When users have access to all the information necessary to arrive at a specific score for each item—as was the case in this study—they do quite well. As long as users are diligently working to gather accurate and complete information on the individuals they assess day-to-day during the course of their work, these findings are expected to translate to “real-world” assessment practices.

By and large, no glaring or serious areas of concern were observed for any of the risk assessment tools at the level of total score, though improvements could still be made to increase reliability in specific domains (which differs by tool). Ongoing booster trainings are recommended for all users on a semi-yearly to yearly basis. It is also recommended that line staff supervisors take a more active role in monitoring users’ accuracy in rating the tool. This can come in the form of scheduled observations of interviews and scoring sessions and/or random “spot checks” of completed assessments to determine their accuracy. In every training and supervision contact, it is important to consistently and emphatically emphasize the ongoing use and frequent consultation of the IRAS/IYAS scoring guide, so that subjectivity and heuristics do not creep into the scoring practices, ultimately reducing reliability in assessment.

## References

- Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 83(5), 762–765.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.
- Koo, T., & Li, M. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15 (2), 155-163.