PPCJI Pennsylvania Partnership for Criminal Justice Improvement

Interrater Reliability of Risk/Needs Instruments

Prepared by Carey Group for the Pennsylvania Partnership for Criminal Justice Improvement

January 2023

Table of Contents

Introduction

Actuarial risk/needs assessments have become routine across the United States justice system and are the standard for informing decision making. Such tools use specific, measurable variables correlated with behavior to predict a person's likelihood of committing a future illegal act and to identify the factors underlying their violations of the law. There is no question that, when properly administered, these tools help the system identify which people will benefit the most from specific interventions, the degree of supervision each person requires, and how best to structure services to meet their needs. In so doing, they assure the community and victims that their safety and well-being are paramount to the administration of justice, and they help the system better manage its limited resources.

These positive outcomes are significantly diminished if decisions are based on inaccurate assessment results. Leaders must therefore take the time to ensure that their risk/needs instrument does what it is supposed to do (predictive validity) and that staff complete and interpret the assessment correctly (interrater reliability).

Predictive Validity

Predictive validity is the extent to which a score on an instrument properly measures what it is supposed to measure. For example, if a tool is used to predict the likelihood of rearrest, does it do so accurately? What is the frequency of false negatives (i.e., predictions of rearrest when rearrests do not occur) or false positives (i.e., predictions of no rearrests when rearrests occur)? Risk/needs instruments have to be validated, a process that involves using the instrument on a population, measuring recidivism rates across various time frames, and comparing recidivism rates to the assessment results.

Ideally, researchers other than the original developers should validate an instrument during its initial development. The validation should be conducted using a sample population in the jurisdiction where the instrument will be used, and the instrument should be revalidated regularly. Ongoing validation may identify issues that need to be addressed. For example, if an instrument starts to perform below expectations, it may be because the tool is no longer valid for the intended population, or there may be issues with interrater reliability (IRR), which measures how consistently different raters (practitioners) score the same person using the same instrument. Even though validity and IRR are distinct, they strongly influence each other. As an example, if a tool is not accurately validated, this may result in issues with IRR. Likewise, if there are issues with IRR, this will most likely impact the tool’s validity.

Rigorous validation is crucial to ensuring buy-in to, and confidence in, the instrument. The developers and the jurisdiction need to be able to explain to staff, stakeholders, and the community why a specific tool was selected and how the results are utilized. Having this knowledge is paramount when faced with questions about a tool's failure to predict an unwanted outcome.

Interrater Reliability

Even though raters' scoring of an assessment instrument does not need to be identical, it is important that there is IRR and an agreement on an acceptable measure of variation. A lack of consistency in scoring will significantly impact the instrument's reliability and its ability to accurately predict the likelihood of recidivism. The repercussions of inconsistency are significant. If a person at low risk of recidivism is scored as high risk, there is the potential of increasing their risk level by placing them into programs that take them away from their positive support systems or by exposing them to people who are higher risk. In addition, since most people at low risk are self-correcting, valuable resources may be wasted if people at low risk are given programming that they do not need. Conversely, if a person at high risk is scored as low risk, they may not receive the appropriate level of supervision and programming they require to remain law-abiding.

Numerous factors may contribute to scoring inconsistencies:

1. **Training:** Staff knowledge and skills can erode over time—impacting the accuracy of the assessment—if initial training is not reinforced with regular booster training.
2. **Tool Experience:** A staff member's lack of experience using the tool may result in inaccuracies in administration and scoring.
3. **Interviewing Skills:** Most tools include an interview guide. Staff's ability to properly interview a person and draw out the appropriate answers is critical to completing an accurate assessment.
4. **Personal Beliefs or Values:** Even though most instruments use objective questions, many include questions that require subjectivity and that can be influenced by the staff's beliefs and values. Staff may also be tempted to reinterpret scoring rules to change the outcome.
5. **Ambiguity:** Ambiguity in the scoring manual may cause staff members to interpret and score the same information differently. Unfamiliar terminology can further exacerbate discrepancies.
6. **Interpretation:** In addition to being a good interviewer, staff need to be able to cognitively process and interpret the information gathered through the interview and other collateral sources and to determine the appropriate assessment scores.
7. **Data Entry Errors:** There is the potential for unintentional data entry errors and, with paper and pencil assessment tools, for mathematical errors in calculating scores.
8. **Workload:** Staff workloads and time restraints are unfortunate realities in most jurisdictions. If staff feel rushed for time or have competing tasks, it will be more challenging for them to conduct and score the assessment accurately.

IRR Studies

The practice of conducting IRR studies allows agencies to monitor for deviations in scoring and to provide opportunities for correction. Many instruments used today go through rigorous IRR testing during their development. Cohorts of staff complete the assessment and work with the developer to modify wording or scoring criteria so that all staff operate from the same understanding.

IRR testing is needed not only during a tool's development but also during its use in the field. Field IRR studies—the process of conducting IRR testing with practitioners—takes into account the variations that may be unique to a jurisdiction or that occur when instruments become overly routine for staff. They are one aspect of a quality assurance process that identifies areas of drift when it comes to assessments.

In the vast majority of cases, IRR results will provide the jurisdiction with a roadmap for improving assessment quality. For example, IRR results may identify specific questions that staff answer inconsistently. Depending on the issue, question, and staff training, a discussion at a staff meeting or an email could correct the problem. IRR results may also identify possible process or interpretation issues, indicating that staff might benefit from a booster training. In addition, IRR testing may identify issues that must be addressed with specific personnel. In a worst-case scenario, when IRR results are so poor, the jurisdiction might stop using a tool entirely until staff are adequately trained and monitored.

Assessment Experts

Assessments are first completed by "assessment experts" to identify the appropriate answers. Then, staff complete the assessment, and their answers are compared with those of the experts. This process allows a jurisdiction to evaluate if staff are completing the assessment consistently and, more importantly, accurately.

INTERPRETING THE RESULTS OF AN IRR STUDY

IRR studies of assessment instruments rate the percentage of time staff score an item on the assessment in a similar manner. IRR testing results in a numeric value ranging from 0 (everyone disagrees) to 1 (full agreement). The overall goal is to have an IRR score of "1," meaning that everyone scores the instrument exactly the same (100% agreement). The general rule for IRR results is that a score of less than .40 is inadequate, .40–.59 is adequate, .60–.74 is good, and .75 and higher is excellent. That being said, jurisdictions must establish what is acceptable to them. For example, even though a score of .6 is statistically good, a jurisdiction must decide if it is good enough for them given the repercussions of greater inconsistency. Jurisdictions may wish to strive for a score of .75 or higher.

False High IRR Scores

If the majority of staff complete assessments inaccurately, a false high IRR score could be achieved. Deeper analysis is required to ensure that the high score is due to accuracy rather than inaccuracy.

The goal of conducting an IRR study needs to be more than just a number. Jurisdictions should establish questions they aim to answer during the data analysis, for example:

1. How consistently do staff score the same person using the assessment instrument?
    a. What is the preferred rate of consistency (IRR) and how can we improve?
    b. What sections and questions indicated the largest degree of inconsistency? Why did the inconsistencies occur (e.g., personal beliefs, values, biases, etc.)?
    c. What interventions are needed to address inconsistencies in sections and questions?

2. How accurately do staff score the same person using the assessment instrument?
    a. What is the preferred rate of accuracy and how can we improve?
    b. What sections and questions indicated the largest degree of errors? Why did the errors occur (e.g., misunderstanding terminology or definitions, misunderstanding the scoring manual, failing to ask correct questions, etc.)?
    c. What interventions are needed to address errors in sections and questions?

3. What staff or groups of staff had the largest degree of errors?
    a. What appears to be the reason for the errors (e.g., misunderstanding questions, failing to properly interview to gain answers, misinterpreting answers, etc.)?
    b. What interventions are needed to address performance issues?

4. Are there any process or procedural changes that should occur?
    a. Do staff member have the needed information to conduct the assessment?
    b. Is the assessment completed at the appropriate points in the justice system?
    c. Is the assessment completed in locations that are appropriate and conducive to the assessment process?
    d. Are there workload or time constraint issues that interfere with the assessment process?

5. How do staff complete the assessment?
    a. Do staff properly explain to the person who is justice-involved the purpose of the assessment?
    b. Do staff use the interview guide when conducting the assessment?
    c. Do staff appropriately use other reliable sources of information, such as official records and collateral information?
    d. Do staff use interpersonal communication skills that encourage appropriate responses?
    e. Do staff properly explain the results of the assessment?

CONDUCTING AN INTERRATER RELIABILITY STUDY

Even though, at face value, measuring how consistently different raters score the same person using the same assessment instrument appears simple, the process can be somewhat complicated. The recommendation is to use an outside vendor or to partner with a university to conduct an IRR study. Using an independent evaluator adds credibility to the assessment and to the results. At a minimum, each jurisdiction should conduct at least one full independent IRR study on their population.

Several possible questions need to be answered before starting the process:

1. **What is the scope of the sample?** A jurisdiction could conduct an IRR study with all staff, all staff from one or more specific groups (e.g., all staff from the intake unit and/or general supervision unit), or randomly selected staff from one or more groups. Even though a larger number of raters is generally preferable, analyzing results of a larger study could be more problematic if statistical software is unavailable. In any case, the sample would be comprised of staff who will be or are conducting the assessment. If a sample is used, various methods could be employed to identify the sample. A jurisdiction could use a simple randomized method, where everyone has an equal chance of being selected, or a stratified sampling process, where staff are split into specific groups (e.g., by unit, division, or staff classification) and a sample of staff is chosen from each group such that there is equal representation from each group. A jurisdiction could use other methods to determine the sample, such as a convenient sampling method, where the sample consists of staff who are readily available, or voluntary sampling, where the sample consists of staff who volunteer to participate. However, these approaches will impact the statistical validity of the study.
2. **How many IRR studies will be conducted?** It is generally preferable to conduct multiple IRR studies over time to identify drift from recommended assessment practices. In addition, IRR studies can be conducted to identify the benefit of specific continuous quality improvement measures. For example, if results of an IRR study indicated the need for booster training, another IRR study could be conducted after the booster training to assess its benefits. Even though multiple IRR studies are desirable, a jurisdiction needs to evaluate its staff workload and the jurisdiction's capacity to conduct subsequent data analysis before determining how many IRR studies will be conducted.

Not Ready to Conduct an Interrater Reliability Study?

Even though it is highly recommended to conduct IRR studies, there are numerous reasons (e.g., workload, funding, expertise, etc.) why a jurisdiction may elect not to assess IRR. If that is the case, the jurisdiction is encouraged to takes steps such as the following to assess fidelity to the assessment instrument:

- Conduct file reviews to review assessment scores.
- Hold group discussions about the scoring of an individual.
- Conduct booster trainings.
- Have staff complete assessments on files or vignettes where feedback can be provided as a group or individually.

3. **How will the test subject be presented?** A jurisdiction may elect to create a case file with all the necessary documents and information, or they may have all staff watch a video clip of an interview, with possible supporting documents. The video clip might show a vignette with an actor or a recording of an actual client completing an assessment. Tool developers may have video clips available. In addition, the study might determine the IRR for one test subject or for multiple test subjects. The clearer the information for each text subject, the greater the integrity of the IRR study.
4. **What is the focus of the study?** The IRR study can be limited to measuring the consistency of overall scores or it can be expanded to analyze the scores of specific questions, for example, questions with frequent issues or questions where the test subject was rated more than one risk band width apart (i.e., one portion of the sample rated the test subject as low risk while another portion rated them as high risk).

STATISTICAL METHODS FOR CALCULATING THE IRR

Three possible statistical methods for calculating IRR—from simple to more complex—are referenced in the literature. Each method has its strengths and limitations. The method chosen largely depends on the number of raters, type of data, jurisdiction's goals, and jurisdiction's expertise in data analysis.

- **Percentage agreement** reports on the proportion of agreement across different samples, but it does not account for variance among raters or correct for chance agreement. It also becomes cumbersome when used with a larger number of raters.
- **Cohen's Kappa** reports on the proportion of agreement between two raters and corrects for chance agreements. However, this method does not correct or examine other types of variances among raters. Cohen's Kappa should be used only by people who have been trained and are expert in this statistical method.
- **Intraclass correlations** measure the reliability of ratings in clusters of data, yielding a statistic known as an intraclass correlation coefficient (ICC). Different versions of ICCs can be calculated depending on the design of the study. One limitation of using intraclass correlations is that ICCs will be different depending on how they are calculated. Intraclass correlations should be conducted only by people who have been trained and are expert in this statistical method.

The most common statistical method is intraclass correlation; however, it is recommended that multiple methods be used since different conclusions could be drawn on IRR based on the method.

PERCENTAGE AGREEMENT EXAMPLE

The following is a simplified example of the percentage agreement method. Jurisdictions are encouraged to develop a deeper understanding of it prior to using it.

Step 1: Build a table of staff scores. For this example, there are five questions in the assessment tool. Staff members can answer each question with a “0,” indicating that the behavior, belief, or trait is not at all present; “1,” indicating that the behavior, belief, or trait is somewhat present; or “2,” indicating that the behavior, belief, or trait is present. Three staff members have completed the assessment.

Question	Staff 1	Staff 2	Staff 3
1	1	1	1
2	0	1	0
3	2	2	2
4	1	0	1
5	1	2	2

Step 2: Add additional columns for the combinations (pairs) of staff. This allows the evaluator to compare staff with each other. For this example, the three possible combinations are S1/S2, S1/S3, and S2/S3.

Question	Staff 1	Staff 2	Staff 3	S1/S2	S1/S3	S2/S3
1	1	1	1			
2	0	1	0			
3	2	2	2			
4	1	0	1			
5	1	2	2			

Step 3: For each pair, put a "1" for agreement and a "0" for disagreement. Agreement occurs when two staff members answer the questions with the same score. For example, for question 5, S1(1)/S2(2) disagree (0), S1(1)/S3(2) disagree (0), and S2(2)/S3(2) agree (1).

Question	Staff 1	Staff 2	Staff 3	S1/S2	S1/S3	S2/S3
1	1	1	1	1	1	1
2	0	1	0	0	1	0
3	2	2	2	1	1	1
4	1	0	1	0	1	0
5	1	2	2	0	0	1

Step 4: For each question, add up the 1s for all combinations and record the total, as a fraction, in an "Agreement" column.

Question	Staff 1	Staff 2	Staff 3	S1/S2	S1/S3	S2/S3	Agreement
1	1	1	1	1	1	1	3/3
2	0	1	0	0	1	0	1/3
3	2	2	2	1	1	1	3/3
4	1	0	1	0	1	0	1/3
5	1	2	2	0	0	1	1/3

Step 5: Find the mean for the fractions in the "Agreement" column.

Mean = (3/3 + 1/3 + 3/3 + 1/3 + 1/3)/5 = .6

Question	Staff 1	Staff 2	Staff 3	S1/S2	S1/S3	S2/S3	Agreement
1	1	1	1	1	1	1	3/3
2	0	1	0	0	1	0	1/3
3	2	2	2	1	1	1	3/3
4	1	0	1	0	1	0	1/3
5	1	2	2	0	0	1	1/3
							.6

The interrater reliability for this sample is .6.

CONDUCTING A SIMPLE NONSTATISTICAL ANALYSIS

Statistics can be intimidating to many, and some jurisdictions may want to do only a simple, nonstatistical analysis of the data. This type of analysis can be equally, and sometimes more, valuable to a jurisdiction.

For example, a quick review of the data below shows that questions 2, 4, and 5 are potentially problematic since there is inconsistency across staff's scores:

Question	Staff 1	Staff 2	Staff 3
1	1	1	1
2	0	1	0
3	2	2	2
4	1	0	1
5	1	2	2

A deeper dive into each staff's scores, including adding an additional column indicating the answer, may help identify which staff members could benefit from additional coaching or training. In this sample, Staff 1 (who answered three questions—2, 4, and 5—incorrectly) and 3 (who answered two questions—2 and 4—incorrectly) may benefit from more assistance.

Question	Staff 1	Staff 2	Staff 3	Answer
1	1	1	1	1
2	0	1	0	1
3	2	2	2	2
4	1	0	1	0
5	1	2	2	2

Observation

Conducting IRR studies is not the only way to ensure that assessments are administered with fidelity. It is equally important to observe how staff conduct assessments. A jurisdiction should develop an observation tool that supervisors and coaches can use to observe staff members during the assessment process (see appendix A for an example). Observations should focus on the following:

- **Introduction:** It is important that staff properly introduce the assessment, its purpose, and what is expected from the individual. Setting the tone can have an impact on the overall results of the assessment.
- **Communication Skills:** Staff's communication skills—both verbal and nonverbal—can also impact the quality of the assessment. A staff member who uses a warm and genuine approach, displays empathy and respect, listens, and asks prompting questions when there is ambiguity will generally gain more helpful information.
- **Use of Manual:** The majority of assessments include a manual and suggested interview questions. It is essential that staff ask the suggested questions and use the manual to ensure that the assessments are scored correctly. Failure to do so can have a significant impact on the fidelity of the instrument.
- **Sources of Information:** Most assessment manuals indicate what sources of information can and should be used to properly score an instrument. They also include guidelines as to which sources should be considered most reliable when there is disagreement among them. Observations should ensure that staff properly use and weigh available information sources.
- **Closing:** Staff must provide proper closure to the assessment process, including identifying next steps.

A supervisor or coach may also wish to complete their own assessment of the person on supervision and compare their scores to those of the staff they are observing. In this way, the observation serves as an opportunity to review the accuracy of the staff's scoring and to gather information about the reason for any discrepancies.

Conclusion

Validating assessment instruments and conducting IRR studies and observations can appear overwhelming, yet they are crucial aspects of an assessment quality assurance process. If assessments are not completed with fidelity, they have limited to no value.

Validation and IRR studies can provide important information about an assessment tool's strengths and areas in need of improvement—helping ensure the effectiveness and utility of the tool. In addition, both IRR studies and observations can help identify issues related to process (e.g., when and where assessments are conducted, who conducts them, how they are conducted), organizational culture (values and beliefs), potential biases that could undermine the impartiality and fairness created by the use of actuarial assessments, and the performance of individual personnel.

If issues related to assessment are identified, they can often be addressed through feedback, training, coaching, and organizational changes such as increasing assessment capacity (e.g., hiring additional staff). Importantly, the impact of these interventions can subsequently be measured.

Resources

Duwe, G. (2017). *Why inter-rater reliability matters for recidivism risk assessment* (Policy Brief Number 2017-03). Public Safety Risk Assessment Clearinghouse. https://bja.ojp.gov/sites/g/files/xyckuh186/files/media/document/pb-interrater-reliability.pdf

Glen, S. (n.d.). Inter-rater reliability IRR: Definition, calculation. *Statistics How To: Elementary Statistics for the Rest of Us!* https://www.statisticshowto.com/inter-rater-reliability/

Lowenkamp, C. T., Holsinger, A. M., Brusman-Lovins, L., & Latessa, E. J. (2004). Assessing the inter-rater agreement of the Level of Service Inventory Revised. *Federal Probation, 68*(3), 56–65. https://www.uscourts.gov/sites/default/files/68_3_6_0.pdf

Wakeling, H. C., Mann, R. E., & Milner, R. J. (2011). Interrater reliability of risk matrix 2000/s. *International Journal of Offender Therapy and Comparative Criminology, 55*(8), 1324–1337. https://www.researchgate.net/publication/51825856_Interrater_Reliability_of_Risk_Matrix_2000s

Appendix: Assessment Observation Tool

The purpose of this tool is to guide observations of staff who are conducting assessments, evaluations of assessments, and the feedback process. The use of this tool is not intended to replace IRR studies and other continuous quality improvement processes that a jurisdiction has undertaken; rather, it is meant to complement these processes. Use of this tool, which should be incorporated into jurisdictions' policies, will help improve organizational and personal confidence in using the assessment, consistency between raters (i.e., interrater reliability), and accuracy of assessments being completed.

Note that use of this tool is an opportunity to improve jurisdiction outcomes and to support staff's professional development; it should not be used to identify shortcomings and penalize staff for those shortcomings.

Three Sections

This tool includes three sections:

1. **Observation:** Complete this section when observing a staff member conducting an assessment.
2. **Assessment Scoring and Results:** Complete this section when scoring the assessment during an observation or file audit and comparing the staff member's results to your results.
3. **Feedback and Recommendations:** Use this section to record feedback and recommendations for improvement.

Terminology

The following is a list of definitions for terms utilized in this tool:

- **Staff Member's Name:** The name of the staff member completing the risk/needs assessment.
- **Observer/Coach:** The staff member observing or evaluating the results of the risk/needs assessment.
- **Type of Assessment:** The name of the specific risk/needs instrument (e.g., Community Supervision Tool [CST]).
- **Individual's Name:** The name of the person who is being assessed.
- **Identifying Information:** A jurisdiction-specific identifier, usually a number, that can be used in lieu of or in addition to an individual's name.
- **Overall Score:** The total numeric score obtained upon completion of the assessment.
- **Risk Level:** The likelihood (e.g., low, moderate, high)—based on the assessment score—that the individual will violate the law again.

- **Override:** A shift away from the risk level identified by the assessment. Answer “yes” if the staff member changed the risk level based on other available information, and indicate the reason for the override.
- **Level Recommended:** The level of supervision recommended by the staff member.
- **Self-Report:** A form completed by the person being assessed in which they provide information that can assist in the scoring of specific questions (e.g., information about prior substance use, family history, perceptions of the justice system, etc.). The information is used to help score the assessment instrument. If the instrument does not include a self-report component or the jurisdiction does not use self-reports, this section should be left blank.

Assessment Observation Tool

Staff Member's Name: ____

Date of Observation: ____

Observer/Coach: ____

Type of Assessment: ____

Individual's Name: ____

Identifying Information: ____

Overall Score: ________ Risk Level: ___________________________________

Override: No Yes Level Recommended: ____________________________

Override Reason: ____

____

Previous Assessment: No Yes Type of Assessment: _______________________

Date of Assessment: ____

Section 1: Observation

Directions: Complete this section if you are observing a staff member conducting an assessment. Observation can occur while in the same room as the staff member, while observing remotely, or while reviewing an audio or video recording of the staff member conducting the assessment. If you did not observe the staff member conducting the assessment, skip to section 2.

Process		
Performance Measures	**Notes**	**Yes No N/A**
Selected the appropriate assessment tool		
Conducted the assessment at the appropriate interval		
Provided the individual a self-report tool, if applicable, prior to the assessment and used the self-report tool to help score the assessment		

Preparation and Introduction		
Performance Measures	**Notes**	**1 = Advanced 2= Proficient 3 = Developing 4 = Did not demonstrate N/A**
Reviewed the file and other official records prior to the assessment		
Introduced the assessment and its purpose		

Rapport and Communication		
Performance Measures	**Notes**	**1 = Advanced** **2= Proficient** **3 = Developing** **4 = Did not demonstrate** **N/A**
Greeted the individual warmly		
Used good verbal communication skills		
Nonverbal skills (eye contact, facial expressions, posture) conveyed interest and respect		
Used motivational interviewing techniques (e.g., open-ended questions, reflective listening/paraphrasing)		
Exhibited an empathetic and genuine approach		
Reduced tension when necessary		
Used authority appropriately		

Rapport and Communication (Continued)		
Performance Measures	**Notes**	**1 = Advanced** **2= Proficient** **3 = Developing** **4 = Did not demonstrate** **N/A**
Allowed the individual to talk		
Effectively redirected the individual if they veered off topic		
Appropriately handled excessive silence		
Minimized distractions		
Displayed patience and open-mindedness		
Avoided correcting or addressing issues identified through the assessment		
Lightly challenged any inconsistencies		

Conducting the Assessment		
Performance Measures	**Notes**	**1 = Advanced** **2= Proficient** **3 = Developing** **4 = Did not demonstrate** **N/A**
Conducted the assessment in the appropriate length of time		
Used the interview guide and/or open-ended questions as dictated by the tool		
Asked appropriate follow-up questions		
Obtained collateral information to verify information provided		

Scoring and Wrap-Up		
Performance Measures	**Notes**	**1 = Advanced** **2= Proficient** **3 = Developing** **4 = Did not demonstrate** **N/A**
Used the scoring guide to score the assessment		
Scored the assessment accurately		
Appropriately assigned weight to the interview, collateral information, or official records when there was a discrepancy		
Override occurred when appropriate		
Explained the results and next steps		

Section Two: Assessment Scoring and Results

Directions: Complete this section if you are scoring the assessment during an observation or file audit and comparing the staff member’s results to your results. This will help determine the accuracy of the results.

In the table, note any questions where there is a discrepancy between the staff member’s score and your score. (Do not list questions where the scores are the same.) When a difference exists, calculate and record the difference. Add feedback and information clarifying the reason for the discrepancy.

Below the table, list the number of questions with a discrepancy as a fraction of the total number of questions. Add up and record both *your* total cumulative score and the *staff member’s* total cumulative score. Calculate the overall difference in scores.

Type of Review: □ File Audit □ Observed Assessment

Scoring Differences:

Section/Domain	Question	Staff Score	Observer Score	Difference
Observer Feedback:				
Observer Feedback:				
Observer Feedback:				
Observer Feedback:				

Section/Domain	Question	Staff Score	Observer Score	Difference
Observer Feedback:				
Observer Feedback:				
Observer Feedback:				
Observer Feedback:				
Observer Feedback:				
Observer Feedback:				

Number of questions with different scores/Total number of questions: _____/_____

Staff score: __________ Observer score: ______ Variation between scores: __________

Any problematic domains/sections?: No Yes Explain: _____________________________

Additional comments: ____

____

____

Section Three: Feedback and Recommendations

Directions: Provide staff feedback on areas where they are meeting or exceeding expectations so they can be positively reinforced and on areas where further skill development is needed so they can administer assessments more effectively. Complete the professional development plan together.

Summary of areas mastered:

Summary of areas that need improvement:

Professional development plan:

Overall summary/recommendations: (Check all that apply)

- The assessment was completed exceeding expectations.
- The assessment was completed within expectations; no further action is required.
- Minor issues were identified and addressed through verbal feedback; no further action is required.
- Issues were identified; written feedback was provided.
- Issues were identified; it is recommended that the staff member attend a booster training on ____.
- Issues were identified; it is recommended that the staff member undergo another user training.
- It is recommended that an additional observation be scheduled because

  ____

  ____.