Inter-Rater Data Methodology

Sonya Stevens and Gabe Ortiz
Reviewed by the ESA TA Team and Dr. Richard Fiene

Background

The project of establishing Inter-rater reliability of child care licensors across Washington state is taking place to establish a common understanding of the new aligned licensing regulations developed according the Early Start Act (2015). In addition, the system will be able to inform licensing oversight of the consistent use of the monitoring instruments. The following methodology is developed to help guide the data collection methods, sample size determination, analysis and reporting.

Method

Data will be collected within three phases: (1) A small group (Early Birds) of data collectors will be utilized in order to fully assess the Inter-Rater processes and instruments including the protocols and ensuring filed practice will inform training. (2) A larger pilot (First Flight) will be used to establish quality assurance standards, and test knowledge implementation. (3) All staff (Second Flight) baseline rating and continued maintenance.

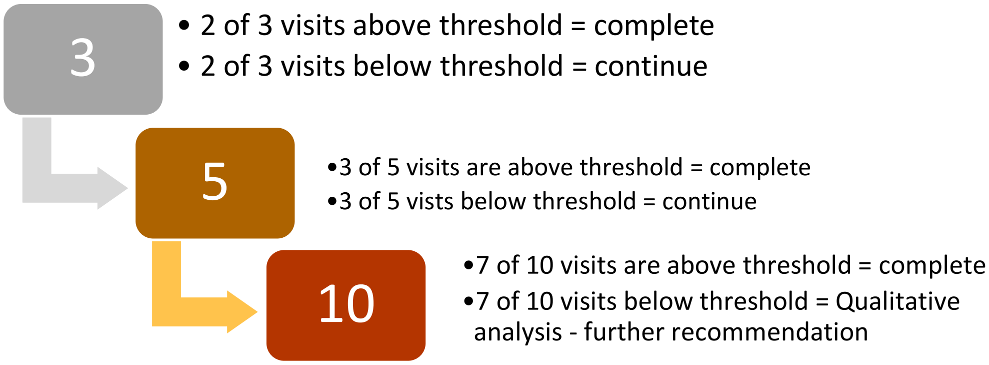
Sample

Sample for the purpose of this process means the amount of visits that will be required by each collector to determine individual reliability scores. Sample size is dependent on each collector's ability to reach a reliable threshold determined by the results of the early bird cohort in combination with empirically proven thresholds (80% using the Cohen's Kappa coefficient or 90% using simple agreement).

Phase One (Early Bird Flight): 8 data gathers

This phase will use a 3 to 5 strategy where Early Birds (EB) will be paired together to complete one to two visits to become familiar with the instrument and identify additional training needs. This could include virtual visits during the training. Once initial evaluations have been completed, Early Birds will then complete second and/or third visit for the purpose of achieving reliability as well as assessing reliability thresholds*. Once three visits are completed in the field and a data collector reaches a minimum score of 80% (Using Cohens Kappa) or higher on two or more visits they will be considered reliable. If two of the three scores are below 80% the collector will complete two more checklists. If the scores are still below 80% additional analysis will be used to identify cause, recommend changes and/or training. The data collector will then be required to start the process over until reaching the desired scores.

Note* If the inter-rater coefficient is higher than 80% for the first group then the expected levels for IRR could be higher for the following group of raters.



3

- 2 of 3 visits above threshold = complete
- 2 of 3 visits below threshold = continue

5

- 3 of 5 visits are above threshold = complete
- 3 of 5 vists below threshold = continue

10

- 7 of 10 visits are above threshold = complete
- 7 of 10 visits below threshold = Qualitative analysis - further recommendation

Phase Two (First Flight): 24 data collectors

This phase will use a 3 to 10 strategy where data collectors will be paired with an EB to complete one visit, then a second EB will be paired for the second visit and so on. Once three visits (inclusive of both family home and center) are complete and a data collector reaches a score of 80% (this amount could be higher depending on results of the Early Bird) or higher on two or more visits they will be considered reliable and complete. If two of the three scores are below 80% the collector will complete two more checklists. If the scores are still below 80% additional analysis will be used to identify cause, recommend changes and/or training. The data collector will then be required to complete five more visits before reliability can be determined.

Data Collection

Checklist Data: Data will be collected on tablet computers using the GoCanvas Application. Each team will be trained in the use of the tablet program by Technical Assistant Trainers. Data will be transmitted to DCYF secure server in Excel format, to the Licensing Analyst team. These data will document:

- Visit ID/Licensor
- Capacity; Type of Facility/Ages of Children Served
- Demographical Data: language of provider/interpreter used/language of licensor; was interpreter used
- Number of consistent/inconsistent WACs between pairs
- List WACs without agreement
- Areas on the checklist that are N/A
- Capturing all drop downs
- Time Measurements for the field visit (length of time on site, on the checklist, time of day/date stamp
- Document comments
- How long a licensor assigned to a provider (Second Flight)

Analysis

IRR Rating: Checklists results will be generated into an excel spreadsheet and sent to the licensing analysts where results can be compared.

Quantitative: IRR scores will be determined by comparing checklist outcomes using Cohen's kappa coefficient (κ). This is a statistic calculation which measures inter-rater agreement between all the checklist (categorical) items at any one visit. Using Kappa calculations is thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance. Alternatively, it is also possible to use a simple agreement calculation whereby the total of agreement is divided by the total amount of items inspected. An empirically proven acceptability rate for this agreement is 90%. Agreements will be translated into a value of one and disagreements will be recorded into a zero value. Once each visit is scored, the rating will be translated onto a worksheet tracking each licensor until acceptable thresholds are met.

Qualitative: Collector narratives will be collected through weekly feedback sessions and Question and Answer forms. These will be compiled and organized in a fashion to allow for the evaluation of interpretations, record decision points and guide training. These data points are TBD.

Reporting

Once analysis is complete a report and recommendations from the Licensing Analysts will be forwarded to the decision making team.

Phase Three (All staff – year one of implementation)

A first flight reliable rater will accompany each licensor on site visits at least twice in the first year of implementation using a full checklist to determine absolute reliability. Once a licensor has two scores at or above the acceptable threshold the licensor will not be required to continue visits for the monitoring year. This will need to be inclusive of a family home and center providers. Licensors will continue to have inter-rater checks until scores reach or exceed the preset reliability level for each provider type.

Provider Sample

We will use the 400-600 monitoring visits identified for measures and output validations beginning with the implementation of 110-300 WAC content. We will first identify 400 licensing monitoring visits that under currently policy would require a comprehensive checklist. It is possible we will need to pull an extra 100-200 to cover any discrepancies in reliability discovered during inter-rater reliability. We will also need to ensure there is an appropriate sample of each provider type as well as other factors such as rural/urban, language representation and so on. Because new monitoring policy allowing visits to take place anytime within a fiscal year we will need to inform the licensing units of the sites selected for long visits prior to July 1st, 2019 to ensure they will not be completed until the new regulations are in effect.

Ongoing Inter-Rater Assessment Needs

Beginning in August 2019 through August 2020 the focus will be to ensure “absolute” reliability assessment using comprehensive checklists from the identified provider sample used during the various validation studies. After absolute reliability is determined, the department will need to identify methodology inclusive of ongoing “relative” interrater checks in the subsequent years post FY 2019/2020 completed using differential monitoring until a full data set (four year rotations) can be determined per licensor inclusive of all licensing regulations. This methodology is yet to be developed and will include extensive data collection for each licensor. New licensing staff will need to complete training and the absolute inter-rater reliability assessment within the first 12 month of employment.