# Wood et al Data Memo

*Daniel M. Swan*

*Monday, May 11, 2015*

This document contains a brief memo on the penalized likelihood estimator (PLE) for PIR and WIR data that we have been developing as applied to a subset of the data from Wood, Hojnoski, Laracy, and Olson (2015). All of the confidence intervals are obtained via a parametric bootstrap procedure.

Continuous duration recording (CDR) is the ideal method for direct observation of behavior, because it provides unbiased estimate of prevalence and incidence. Prevalence we define as the true proportion of time the behavior occurs, and incidence we define as the rate at which new behaviors occur. However, because of the amount of sustained effort it requires of the observer, CDR is not commonly used in field settings. More frequently, some form of interval recording procedure is used. Partial interval recording (PIR), whole interval recording (WIR), and momentary time sampling (MTS) are three common, well known methods used in direct observation of behavior. Partial interval recording slices up each observation session into a number of equal-length intervals – for instance a 15 minute observation session might be cut into 45 20-second intervals. Any interval containing the behavior, no matter how brief the behavior, is scored as 1 and any interval with no instance of the behavior is scored as a 0. Whole interval recording is similar, except that an interval must contain the behavior for its whole length to be scored a 1 and is otherwise scored a 0. Momemtary time sampling, instead of measuring the presence or absence of the behavior within the interval, measures the presence or absence of the behavior at the "moment"" at beginning and end of each interval. Once again, presence of the behavior gives that moment a score of 1 and the absence gives that moment a score of 0.

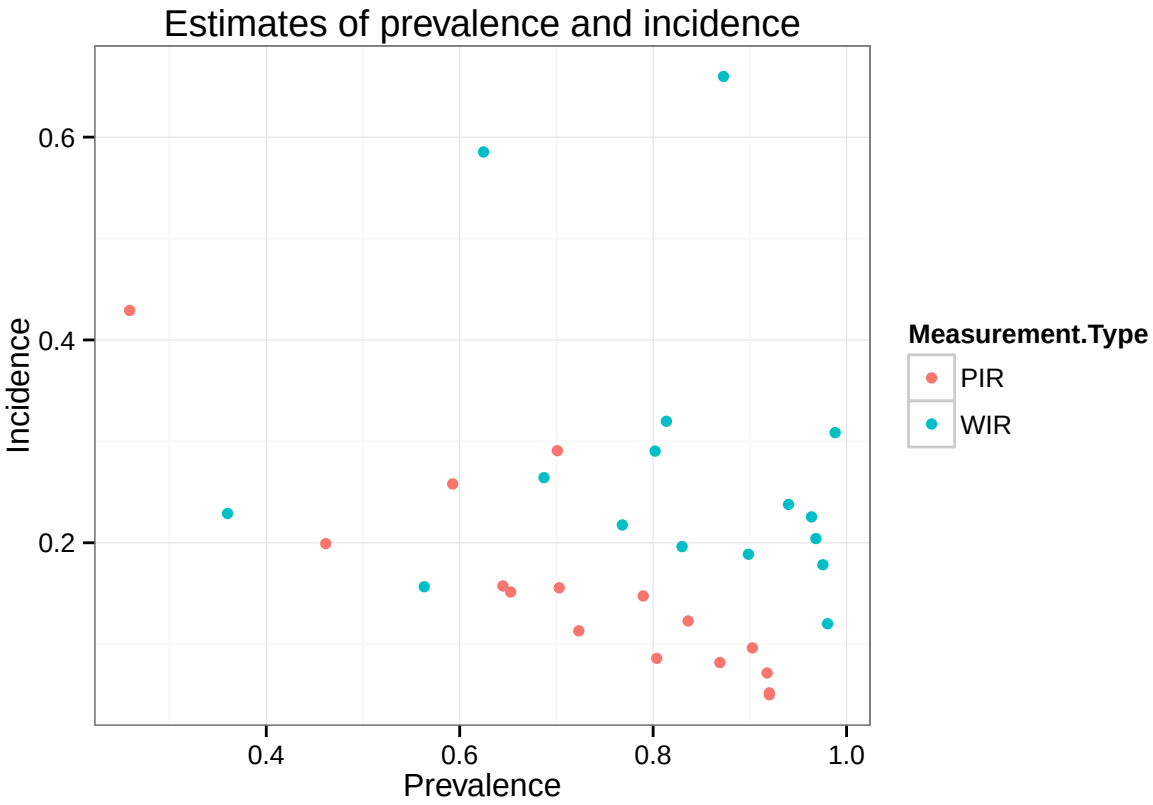
The problem with these three methods is that PIR, WIR, and MTS as they are traditionally employed only allow for a single parameter to be estimated. This is typically a "summary" estimate, the proportion of intervals scored as having the behavior of interest present. When trying to summarize any "state" behavior - that is any behavior with a discrete length rather than one that occurs instantaneously - we are interested in characterizing both prevalence and incidence. In the case of PIR and WIR, the summary measure is often treated as an estimate of prevalence. In reality the summary proportion is a combination of both prevalence and incidence, with PIR overestimating and WIR underestimating prevalence. In the case of MTS, the summary proportion is an unbiased estimate of prevalence under very weak assumptions, but we are still left without any estimate of incidence. Both parameters are necessarily of interest, as a behavior with high prevalence and low incidence is very different from one with high prevalence and high incidence. In the case of high prevalence and low incidence, you might have a child who is actively engaged in learning for a large proportion of the time with only a very small number of instances where they are off task. In the case of high prevalence and high incidence, you have a child who is "engaged" for short bouts, but also has many instances of off-task behavior. These very different scenarios likely lead to very different learning outcomes, yet with only an estimate of prevalence we might characterize these two scenarios as being very similar. Our PLE method is an attempt to give researchers who wish to continue utilizing interval recording procedures a method of estimating both prevalence and incidence. . #Data and Analysis

The original data for Wood et al (2015) came from 13 video-taped sessions of 24 target student participants, with between one and four children captured by each video. The first author coded the sessions using PIR, WIR, and MTS. A senior graduate student coded the videos using continuous duration recording, so that the first authors' coding would not be influenced by knowledge of the "true" value of prevalence. The authors employed a variant of PIR sometimes seen in the direct observation literature that we refer to as fractional interval recording (FIR). In FIR, rather than marking any interval containing the target behavior a 1, the behavior must take last some pre-determined proportion of the interval to be considered present in the interval. In the case of the Wood et al (2015) data, the behavior needed to last for at least 5 seconds, or 1/3 of the interval, to be considered present in the interval. The FIR data from the paper does not precisely conform to the model used in our PIR PLEs, but this allows us to examine the performance of the PLEs when the model is slightly mis-specified and compare it to the data from when the model is correctly specified by examining the WIR estimates.

We were provided with scans of the original hard copies of interval level-data for 16 of the 24 participants in the study. Two coders transcribed the data to a spreadsheet independently, and then the data was checked for agreement. For the purposes of this analysis, any missing intervals in the middle of the recording were simply discarded and we treated only those intervals containing data as the "complete" record without accounting for missingness in our model.

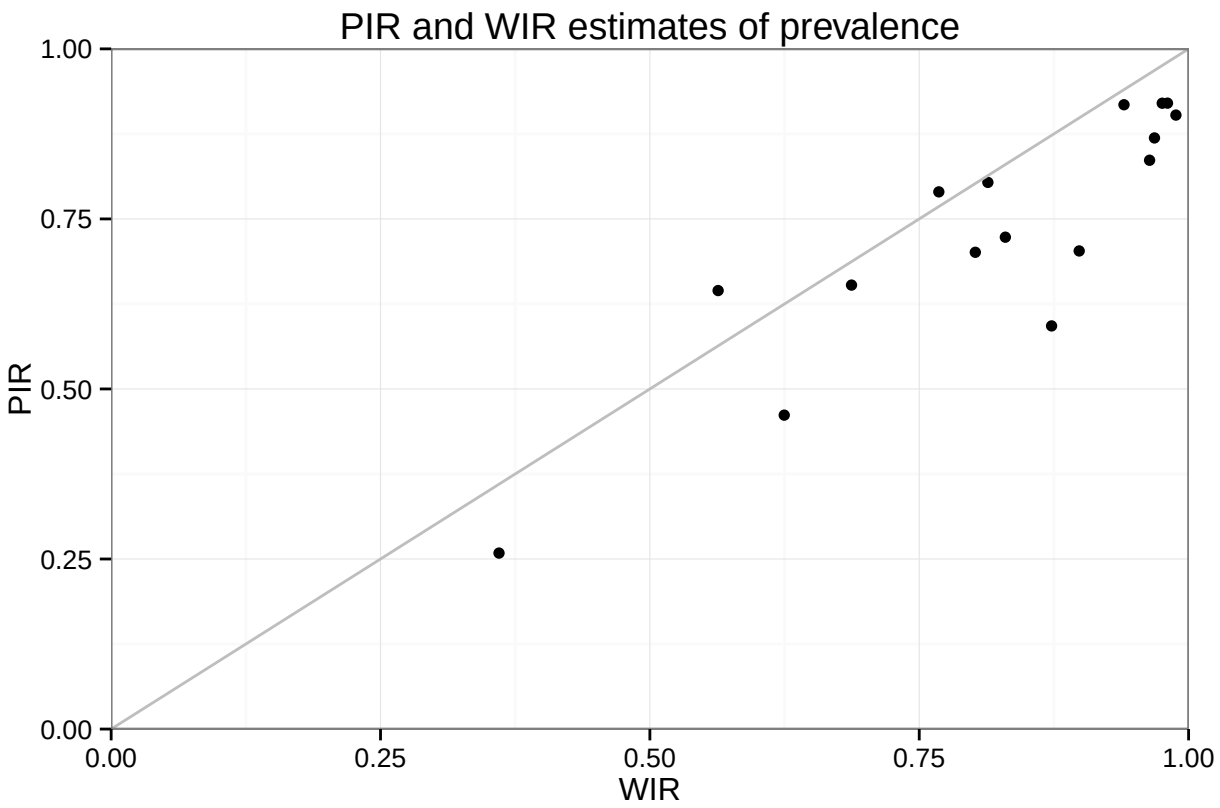
In order to compare the estimates to a "true" metric, we hand-transcribed the continuous duration recording values of prevalence from Table 1 of Wood et al (2015) – along with the MTS values, teacher Ratings, and expert ratings – by matching the summary values of both PIR and WIR to the other estimates in the table. Both the CDR and MTS values are displayed in this document as proportions rather than percentages to conform to the convention we typically use.

## Prevalence and Incidence

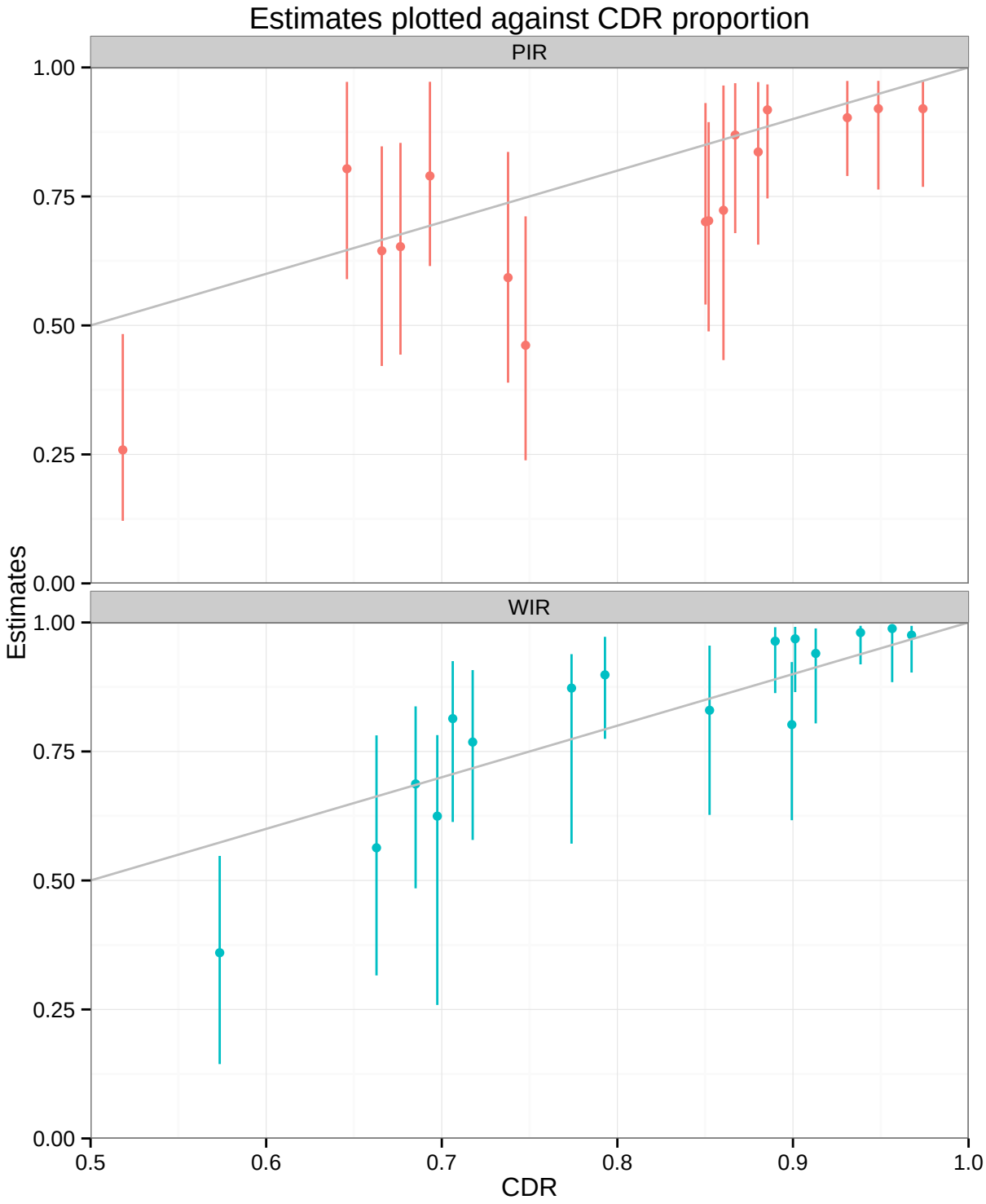



The first plot displays both PIR and WIR estimates of prevalence plotted against incidence. Here we show that generally WIR estimates have higher incidence and prevalence than the PIR estimates. This may simply be an artifact of the modified PIR used to gather the original data. However, without comparing these values to the true values we can't know if the difference is due to systematic bias, which of the observation methods is biased, and how we might characterize this bias.

## Prevalence

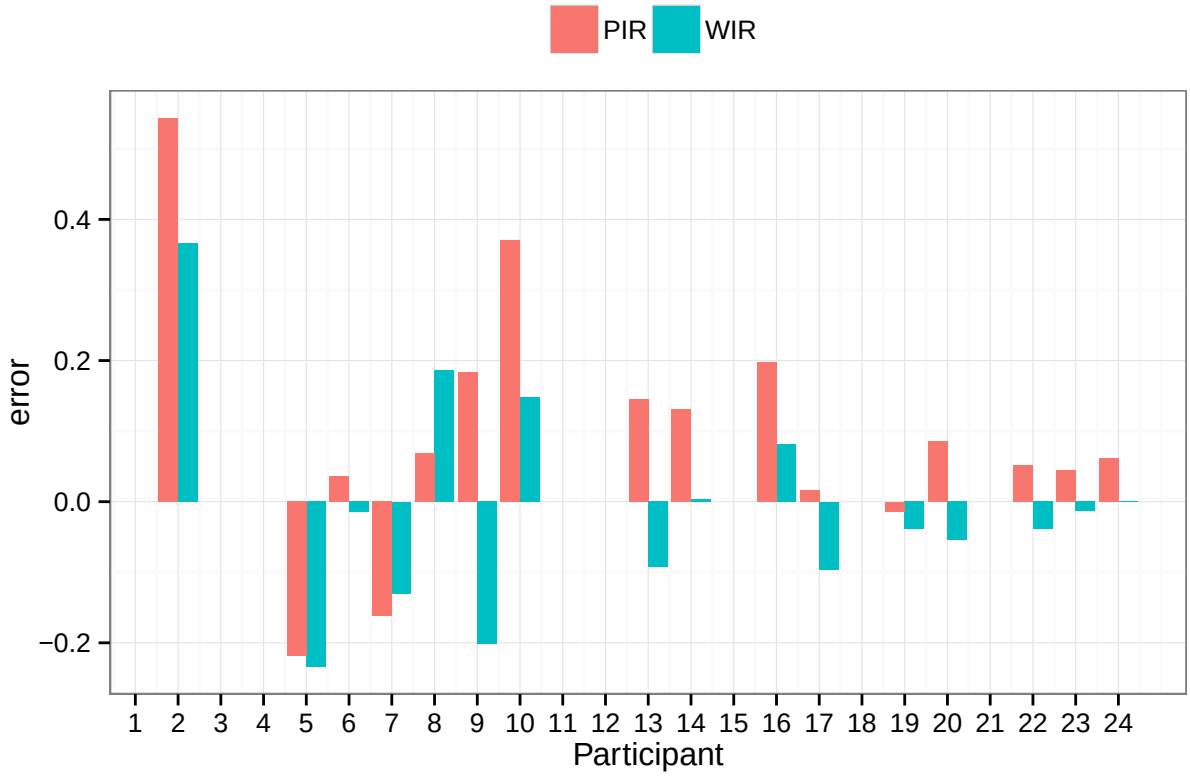



The second plot displays the PLE estimates of prevalence from PIR and WIR plotted against one another. The grey line is a line that passes through $x = y$, where the points would lay if the agreement between the PIR and WIR observations was exact. The plot suggests that the PIR estimates of prevalence are generally lower than the WIR estimates. It is possible that this is attributable to the fact that the PIR used in Wood et al (2015) was a modified version. Putting a minimum on the length of time required for a behavior to "count" as having occurred likely reduced the overall upward bias on prevalence, bias that we attempt to account for in our model.




The third plot displays the PLE estimates of prevalence from both PIR and WIR plotted against the continuous duration recording proportion. The vertical bars overlaid on the points represents the parametric bootstrapped confidence intervals. We can use this as a benchmark for the bias of our estimates. The grey line is a line through $x = y$, where all of the points would lie if there was perfect agreement between CDR and the other observation methods. This line also allows for easy assessment of whether or not the confidence

interval provides coverage of the "true" value. The PIR data appears to be an underestimate of prevalence, while the WIR data is an overestimate, although to a lesser degree. The WIR CIs also generally cover the CDR proportion (coverage = 0.94) whereas the PIR CIs coverage is slightly lower (coverage = 0.81).




The fourth plot displays measurement error relative to the CDR value of prevalence. Unlike the original manuscript, this plot doesn't show the same consistent downward bias for WIR data, although the upward bias for PIR data is still present. The magnitude of the bias for PIR appears slightly larger than in the original manuscript, but the general magnitude of the bias for the WIR estimates is considerably lower.

Table 1: Error

| Measurement Type | PLE | Summary |
|---|---|---|
| PIR | 0.0183 | 0.0093 |
| WIR | 0.0093 | 0.0618 |
| MTS | - | 0.0092 |

Table 1 contains the value of the "mean squared error" of the PLEs as well as the summary measurements for the three methods observational methods. In this case we define "mean squared error" as $\frac{\sum_{i=1}^{n}(\hat{\theta}-CDR)^2}{n}$, where $\hat{\theta}$ is a given estimate of prevalence. The error for PIR PLE is about twice that of the summary measurement, while the error for WIR PLE is very small compared to the summary measurement. The MTS error is roughly equivalant to the WIR error, which is excellent considering that MTS estimates are generally considered to be "unbiased" under very minimal assumptions. The discrepancey between PIR and WIR error is probably an issue of the model not accounting for the slightly different method of PIR used in this data, whereas our model is appropriately specified for the WIR data.

Table 2: Spearman's Rho - Teacher Ratings

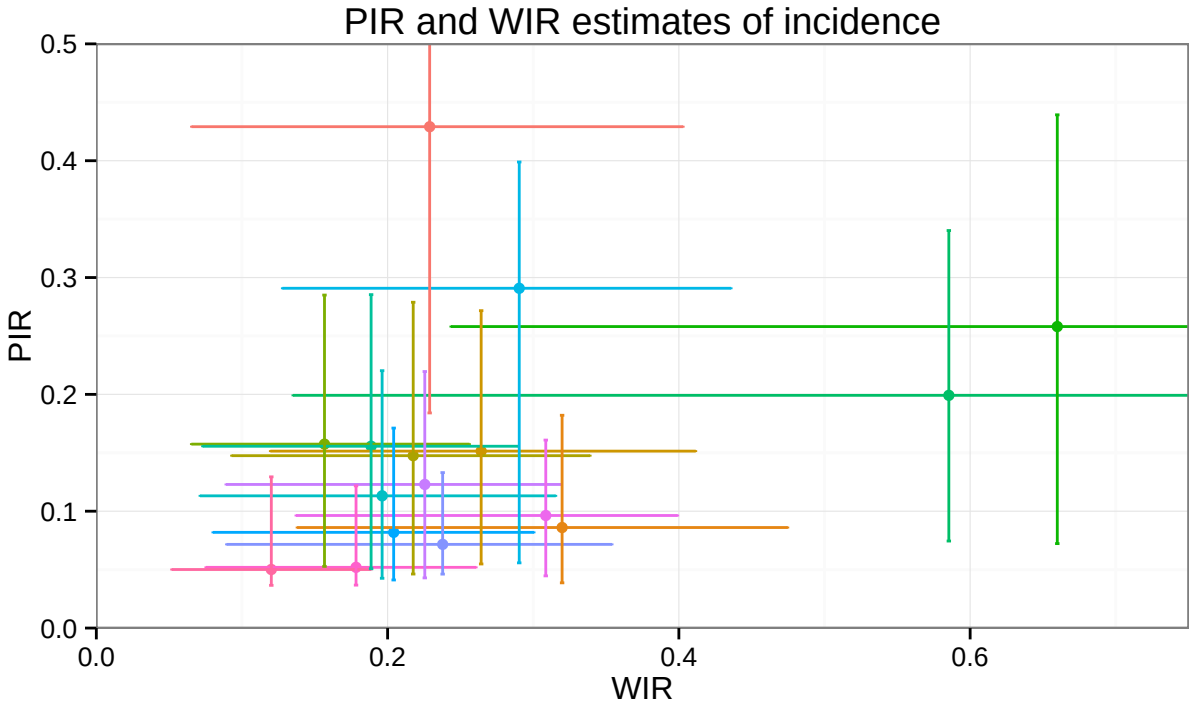
| | Observational Method | PLE | p-value | Summary | p-value |
|---|---|---|---|---|---|
| 3 | PIR | 0.33 | 0.21 | 0.21 | 0.43 |
| 4 | WIR | 0.29 | 0.27 | 0.44 | 0.08 |
| 2 | MTS | - | - | 0.35 | 0.19 |
| 1 | CDR | - | - | 0.16 | 0.55 |

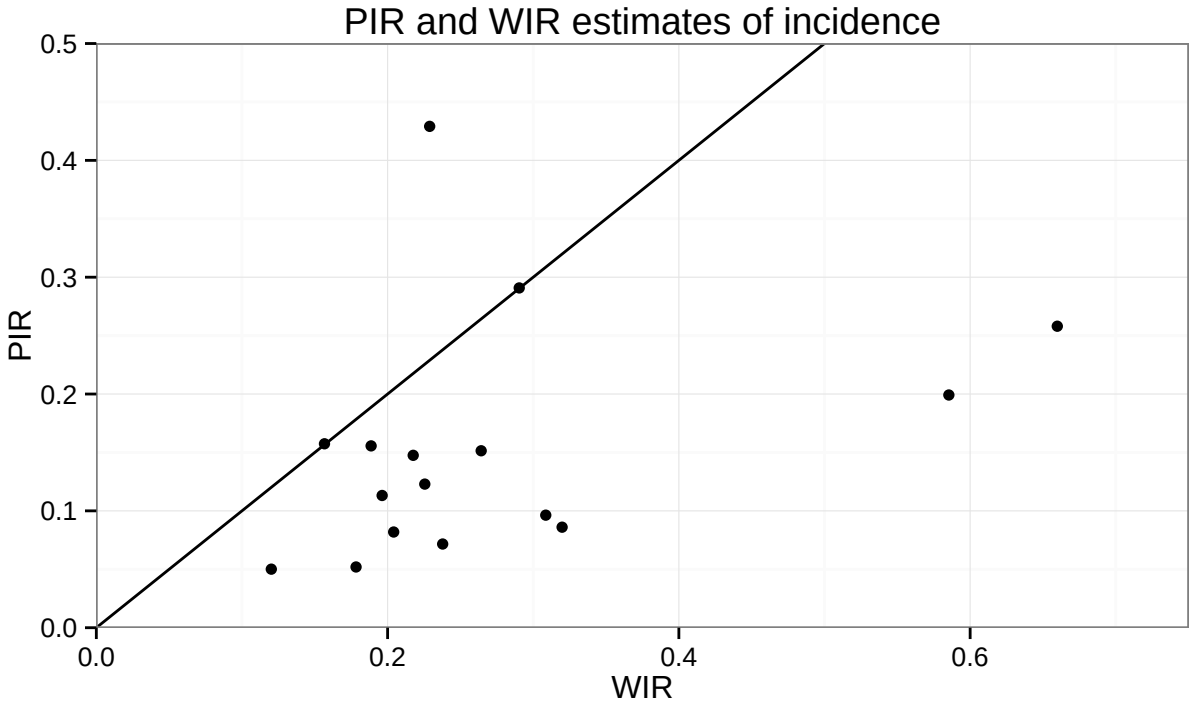
Table 3: Spearman's Rho - Expert Ratings

| | Observational Method | PLE | p-value | Summary | p-value |
|---|---|---|---|---|---|
| 3 | PIR | 0.62 | 0.01 | 0.61 | 0.01 |
| 4 | WIR | 0.55 | 0.03 | 0.57 | 0.02 |
| 2 | MTS | - | - | 0.67 | 0.00 |
| 1 | CDR | - | - | 0.71 | 0.00 |

Tables 2 and 3 display the correlation between PLE of prevalence or the summary proportions and the Teacher and Expert ratings as well as the estimated p-value for each correlation. This table excludes those 8 observations that were not included in the data we were provided. In the case of the Teachers, none of these correlations are significant. In the case of the Experts, all of them are. The correlations between the PLEs and the expert ratings are slighly, although only modestly, worse than the other methods (except for being comprable to the summary in the case of PIR). However, this is does not necessarily point to a disadvantage in the PLEs. When offering a global assessment of any state behavior, that assessment is likely to depend on both the prevalence and the incidence. A child who is engaged most of the time with few instances of being off task is likely to have a very different learning experience than a child who is engaged a large proportion of the time but also has many instances where they are off task or distracted. Ignoring incidence ignores an important component in state behaviors.

## Incidence




The fifth plot and sixth plots display the PIR and WIR estimates of incidence displayed against one another. The third plot also displays the CIS for incidence. Both plots have been provided because the tight clustering of incidence can make it difficult to interpret the fifth plot. The values of incidence have been scaled on a

per-interval basis. That is to say, if the estimate of incidence is 0.25 we have on average one quarter of a behavior per interval, or about one new behavior every four intervals on average. This second value is easily calculated from the first - if we denote incidence as $\zeta$ then the average number of intervals per new behavior is simply $1/\zeta$.

As with prevalence, the general pattern is that the PIR estimates are lower than the WIR estimates, as well as having some even more extreme deviations than prevalence. Unlike prevalence, we have no direct estimates of incidence to compare our estimates to, so it is difficult to characterize which of the two types of observation procedures best estimate the "true" value based on the data alone.

# Conclusions

Generally speaking, it appears that the WIR estimates for prevalence are better than the PIR estimates for prevalence. The plot of measurement error suggests that there these estimates are biased neither systematically upward nor downward. In addition, the magnitude of the bias is much lower, suggesting that our PLE reduces the bias in the WIR estimates considerably. The mean squared error estimate also suggests that the WIR estimates are the better of the two in this case. While the agreement between the expert raters and the WIR estimates of prevalence is not as high as we might like, that correlation ignores the important of incidence in characterizing a state behavior like academic engagement.

In addition, the results of our analysis suggests that the PLEs may be sensitive to model mispecification when the PIR model is used with FIR data. Further investigation of the impact of the fractional method on PLE estimates appears warranted.