Wood et al Data Memo

Daniel M. Swan and James E. Pustejovsky

June 2, 2015

```
##    user  system elapsed
##    0.07    0.00 1161.00
```

In this memo, we describe an application of penalized likelihood estimators (PLEs) for partial interval recording and whole interval recording to a subset of the data from Wood, Hojnoski, Laracy, and Olson (2015). These data contain direct observations of childrens' academic engagement behavior, recorded using several different methods including both PIR and WIR, as well as momentary time sampling and continuous duration recording.

Academic engagement is a *state behavior* (sometimes called a *duration-based behavior*), in that each bout of the behavior has a positive duration. State behavior has two primary characteristics: *prevalence* (also called *percentage duration*), which is the true proportion of time the behavior occurs, and *incidence* (also called *rate*), which is the frequency with which new bouts of behavior begin, per unit time.

Continuous duration recording (CDR) is theoretically the ideal method for direct observation of behavior because it provides simple, unbiased estimates of *both* prevalence and incidence. However, because it requires the observer to maintain sustained attention for the entire observation session, CDR is not always 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. PIR 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. WIR 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. In contrast, MTS records the presence of the behavior at the "moment"" at beginning or end of each interval, giving each moment a score of 1 if the behavior is present and a score of 0 if it is absent.

Usually, the interval scores from a session are summarized as the proportion of intervals with behavior. In the case of PIR and WIR, this summary measure is usually interpretted as an estimate of prevalence. However,the summary proportion is actually affected by both prevalence and incidence, with PIR systematically overestimating and WIR systematically 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.

In addition to the fact that PIR and WIR systematically mis-estimate prevalence, a further problem with all three methods is that they only estimate prevalence, when incidence may be of interest as well. Consider that 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.

The PLE method that we are developing is an attempt to give researchers who use interval recording procedures a method of estimating both prevalence and incidence, while avoiding the systematic biases of the PIR and WIR summary scores. The methods build on earlier work by Brown, Solomon, and Stephens (1977) that developed maximum likelihood estimators for MTS data.

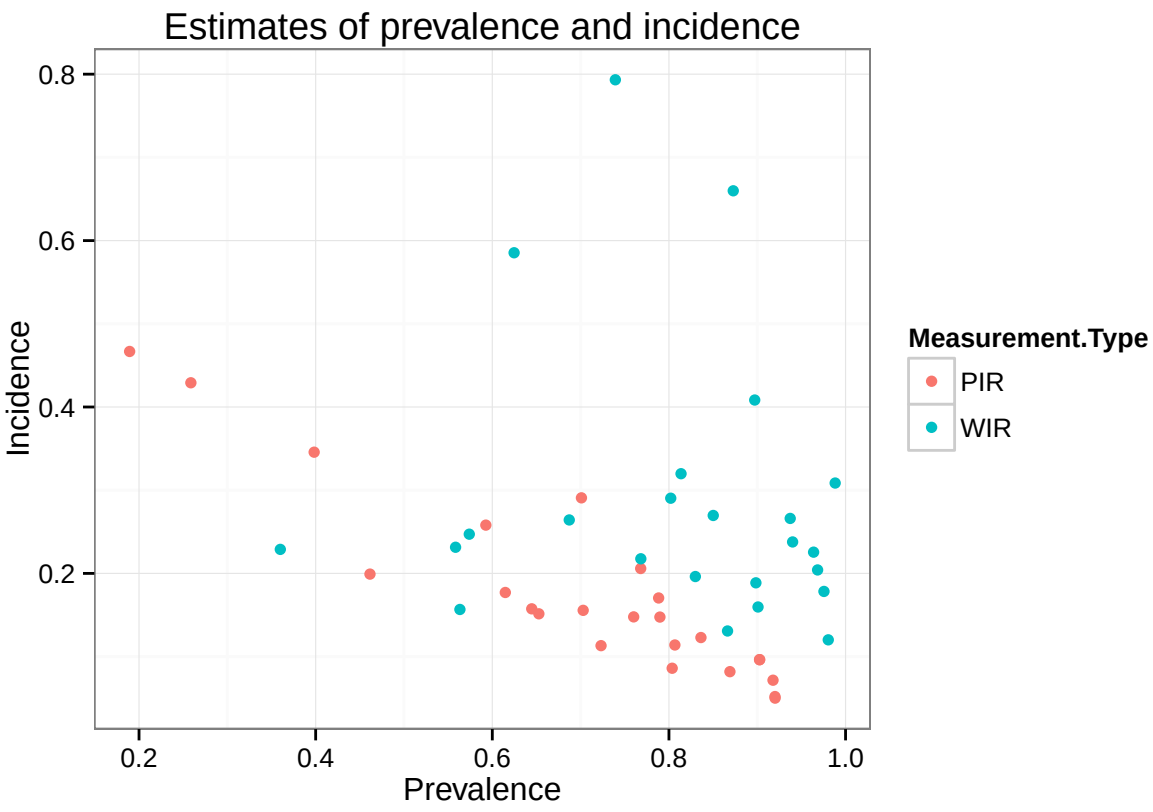
Data and Analysis

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, which we refer to as fractional interval recording (FIR). In FIR, rather than marking any interval containing the target behavior a 1, the behavior must last some pre-determined proportion of the interval to be considered present in the interval. In the present study, the behavior needed to last for at least 5 seconds (1/3 of the interval) to be considered present in the interval. The FIR data from the paper does not precisely conform to the modeling assumptions behind our PLE method. This discrepancy therefore allows us to examine the performance of the PLEs when the model is mis-specified. The WIR estimates were recorded using the conventional rule, and so the WIR data llows us to examine performance when this aspect of the model is correct.

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 observation session were discarded and we treated the intervals containing data as the "complete" record, without accounting for missingness in our model.

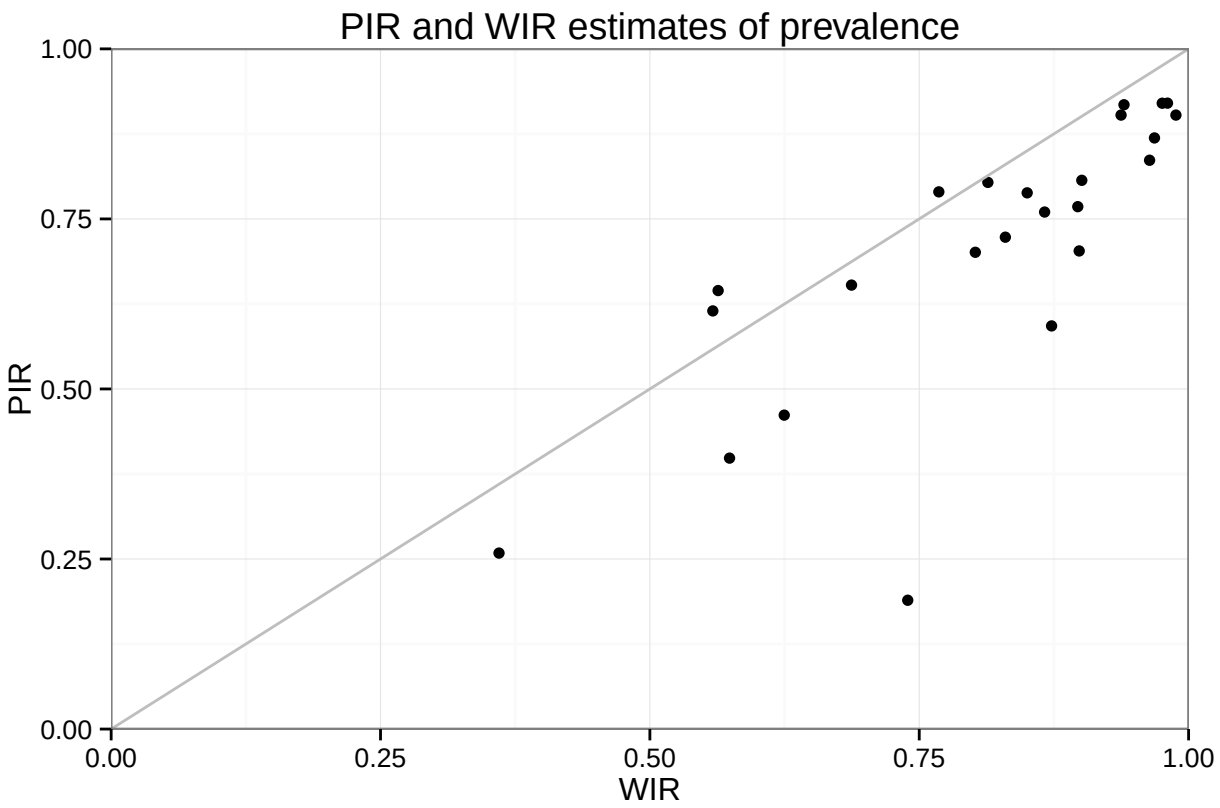
In order to benchmark the PLEs, we hand-transcribed the continuous duration recording values of prevalence from Table 1 of Wood et al (2015), along with the MTS values, teacher ratings, and expert ratings, by matching the summary values of both PIR and WIR to the other estimates in the table. Both the CDR and MTS values are displayed in this document as proportions rather than percentages to conform to the convention we typically use. Confidence intervals for the PLEs were obtained via a parametric bootstrap procedure.

Prevalence and Incidence

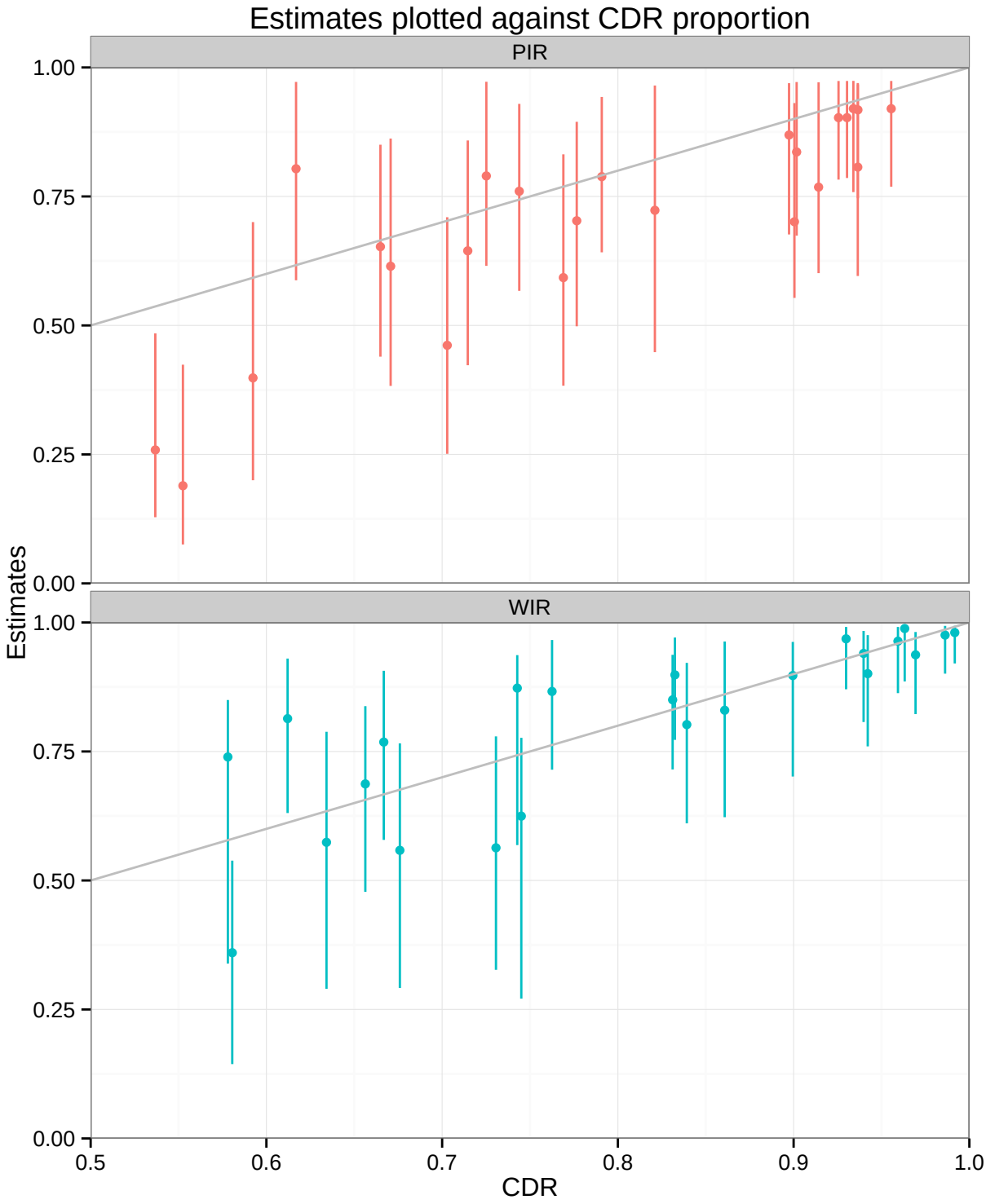


The first plot displays both PIR and WIR estimates of prevalence plotted against incidence (per interval). It appears that the WIR estimates have higher incidence and higher prevalence than the PIR estimates. This is probably an artifact of the fact that FIR (modified PIR) was used to score the observations. 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

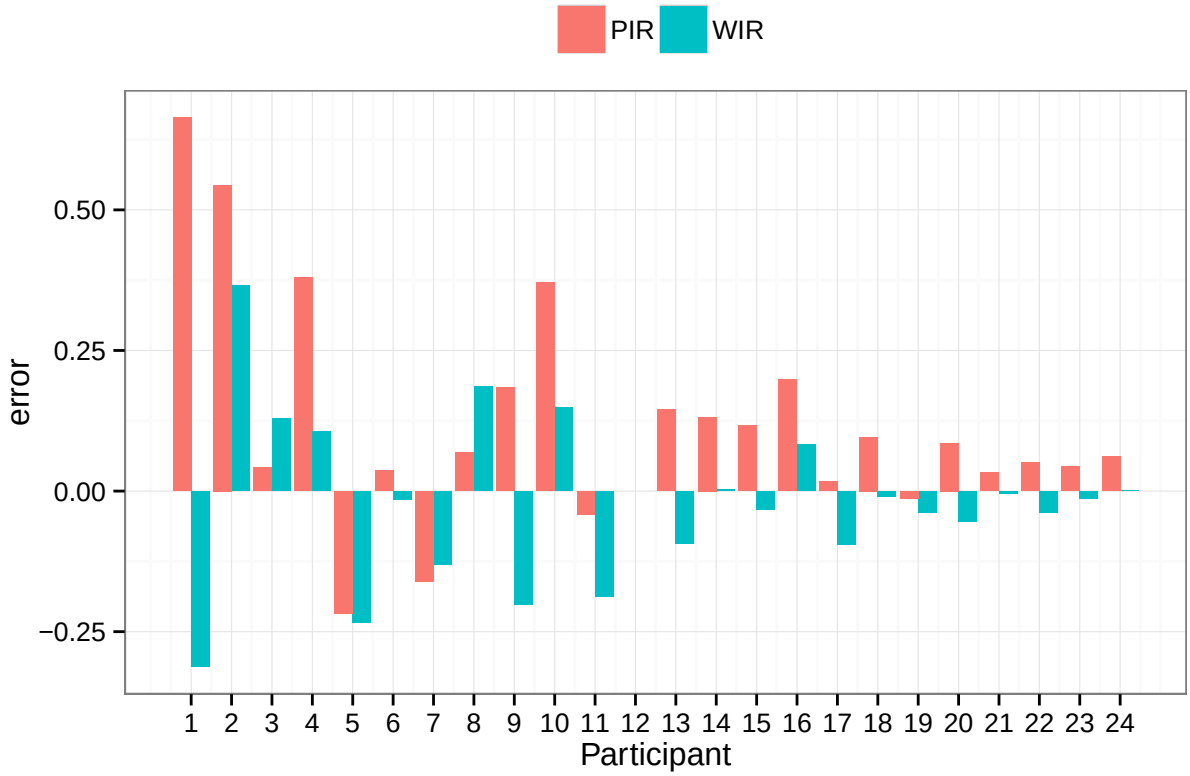


The second plot displays the PLE estimates of prevalence from PIR and WIR plotted against one another. The grey line is a line that passes through $x = y$, where the points would lay if the agreement between the PIR and WIR observations was exact. The plot suggests that the PIR estimates of prevalence 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 (FIR). 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 CDR proportion, which we treat as a benchmark for the PLEs. The vertical bars overlaid on the points represents the parametric bootstrap confidence intervals. 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. Confidence intervals that intersect with the grey line indicate that they cover the "true" score. The PIR estimates appear

to under-estimate of prevalence, while the WIR estimates over-estimate, but to a much lesser degree. The WIR CIs mostly cover the CDR proportion (coverage = 0.96), while the PIR CIs coverage is slightly lower (coverage = 0.83).



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; the magnitude of the bias for the WIR estimates is considerably lower.

Table 1: Error

Measurement Type	PLE	Summary
PIR	0.0214	0.0084
WIR	0.0089	0.0622
MTS	-	0.0367

Table 1 reports the *mean-squared error* (MSE) of the PLEs as well as that of the summary measurements, for each of the three recording procedures, where the MSE is defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - CDR_i)^2,$$

for a prevalence estimate $\hat{\theta}_i$ and CDR score CDR_i. The error for the PIR PLE is not quite three times that of the summary measurement, while the error for WIR PLE is very small compared to the summary measurement. The WIR error is smaller than the MTS error, which is excellent considering that MTS estimates are generally considered to be "unbiased" under very minimal assumptions. The discrepancy

between PIR and WIR error is probably an issue of the model not accounting for the modified PIR method (FIR) used to collect these data, whereas our model is appropriately specified for the WIR data.

Table 2: Spearman's Rho - Teacher Ratings

Observational Method	Prevalence PLE	p-value	Incidence PLE	p-value	Summary	p-value
PIR	0.25	0.23	-0.3	0.16	0.17	0.42
WIR	0.19	0.36	-0.37	0.08	0.37	0.08
MTS	-	-	-	-	0.11	0.60
CDR	-	-	-	-	0.21	0.33

Table 3: Spearman's Rho - Expert Ratings

Observational Method	Prevalence PLE	p-value	Incidence PLE	p-value	Summary	p-value
PIR	0.66	0	-0.48	0.02	0.66	0.00
WIR	0.57	0	-0.31	0.14	0.60	0.00
MTS	-	-	-	-	0.53	0.01
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. While none of the correlations for teachers are significant, the correlations between the prevalence PLEs and the expert ratings are comparable to or better than the other methods, except in the case of CDR where they are worse. 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. As discussed previously, 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, as is suggested by the negative relationship between incidence and expert ratings. Ignoring incidence ignores an important component in state behaviors.

Table 4: Squared Correlations - Teacher Ratings

Observational Method	Squared Rho	R-squared	p-value
PIR	0.0292	0.147	0.5515
WIR	0.1344	0.2413	0.038
MTS	0.0130	-	-
CDR	0.0439	-	-

Table 5: Squared Correlations - Expert Ratings

Observational Method	Squared Rho	R-squared	p-value
PIR	0.4321	0.503	0.167
WIR	0.3576	0.436	0.0184
MTS	0.2782	-	-
CDR	0.5037	-	-

Tables 4 and 5 contain squared correlations for both Teacher and Expert ratings. The first column contains

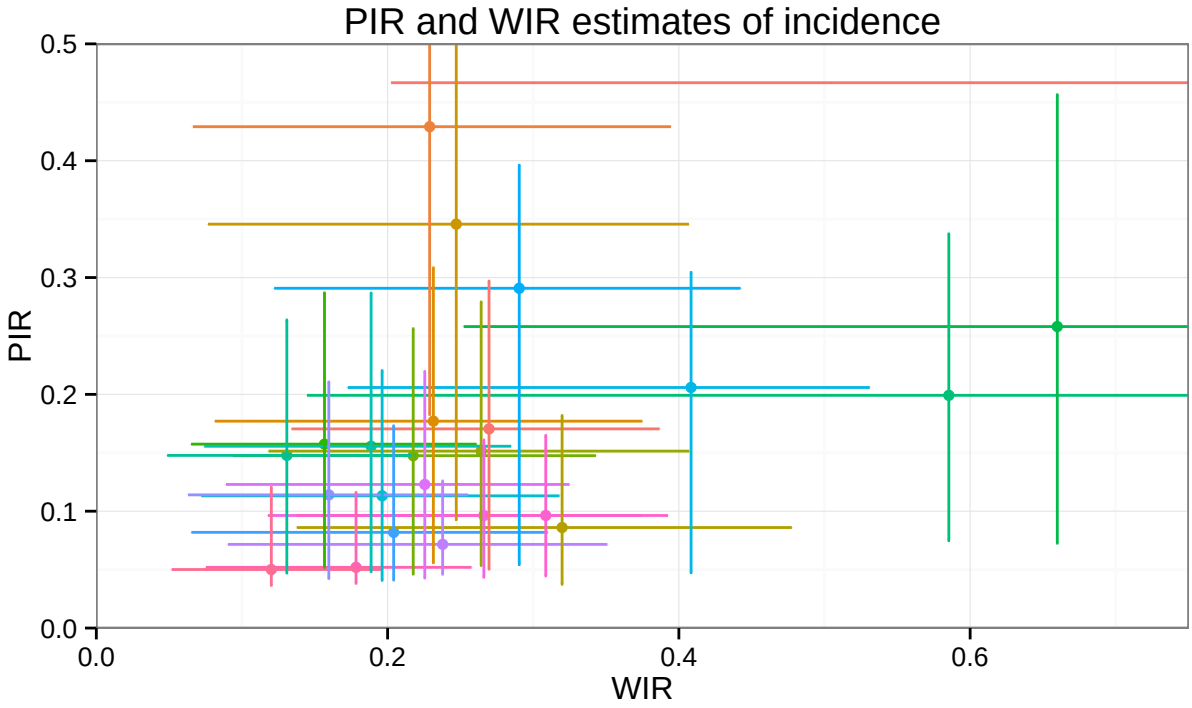
the squared values of the correlations contained in tables 2 and 3. The second column contains the R-squared value of a linear model regressing the ratings on the PLE estimates of both prevalence and incidence. The third column contains the p-value when comparing a linear model regressing *only* the estimate of prevalence against the rating to a model with both prevalence and incidence. P-values of less than 0.05 indicate that incidence explains a significant proportion of the varability in the ratings, above and beyond prevalence.

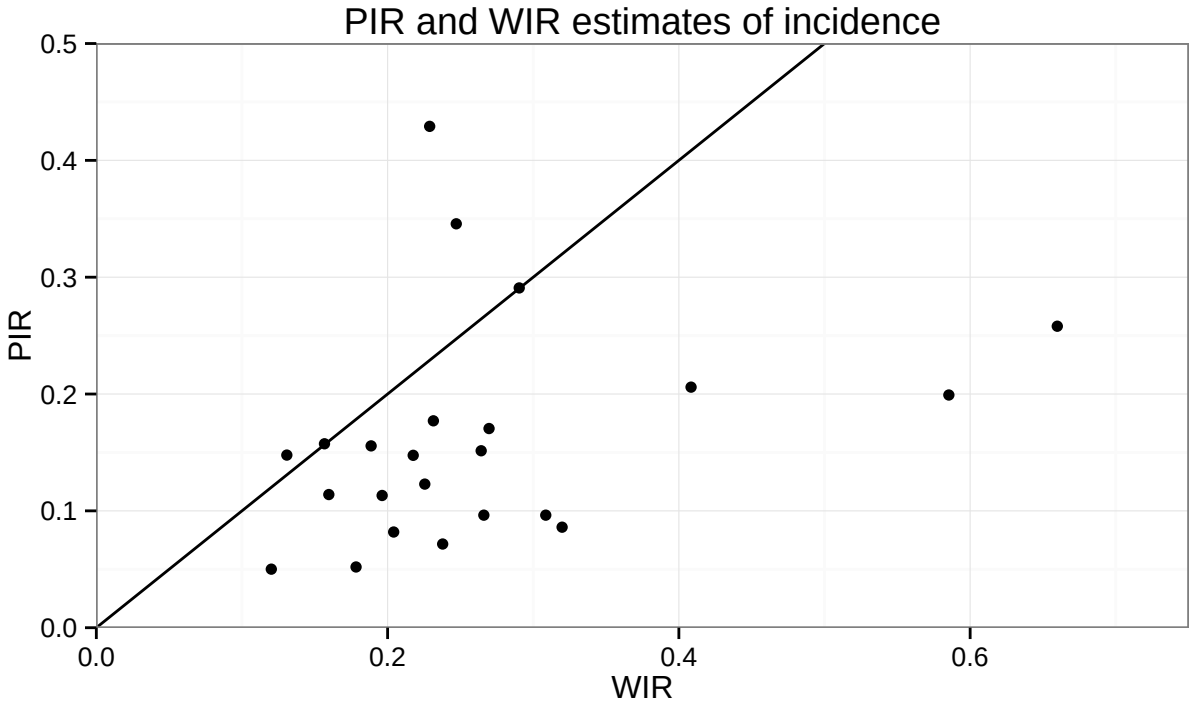
Table 6:

	Dependent variable:	
	Teacher Ratings	Expert Ratings
	(1)	(2)
phiest	1.047	5.274**
	(0.984)	(1.889)
zetaest	−2.177**	−4.822**
	(0.983)	(1.886)
Constant	1.822*	3.629**
	(0.888)	(1.704)
Observations	24	24
R^2	0.241	0.436
Adjusted R^2	0.169	0.382
Residual Std. Error (df = 21)	0.783	1.502
F Statistic (df = 2; 21)	3.340*	8.116***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 6 compares the Teacher and Expert rating models using the WIR data. Unsurprisingly, incidence is negatively related to teacher rating of academic engagement. Of particular interest is that prevalence is only significant in the case of expert raters. It may be that teachers are simply more interested in the *frequency* of off-task behavior rather than its overall length. This points to the importance of characterizing both prevalence and incidence.

Incidence





The fifth plot and sixth plots display the PIR and WIR estimates of incidence displayed against one another. The values of incidence have been scaled on a per-interval basis; for instance, 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. 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.

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

Overall, it appears that the WIR estimates for prevalence are better than the PIR estimates for prevalence. The plot of measurement error suggests that the WIR 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 might be desired, this 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 is warranted.