Sample Size Justification Submitted to IRBs Prior to Study Launch

In terms of test accuracy, sensitivity is a measure of a test's ability to identify true positive cases of a condition, and specificity, on the other hand, is a measure of true negatives. There is often a tradeoff between sensitivity and specificity. A test with high specificity would result in few false positives, but could misclassify too many patients as negative for the condition when they are really positive (false negative). In the case of life threatening conditions, there is a large cost to false negatives. In the case of perinatal depression, however, we see more harm stemming from false positives than from false negatives. Aside from placing unnecessary burden on overburdened healthcare systems, false positives can also be stigmatizing, especially in a setting like rural Kenya. Thus while aiming for an optimal balance, we have a slight preference for high specificity over high sensitivity.

As reported in the Sweetland et al. (2014) systematic review of 65 depression screening validity studies from 16 countries in sub-Saharan Africa, there is substantial variability in assessment measures, gold standards, and diagnostic accuracy used to assess the criterion validity of depression screening instruments. In Table 1, we summarize reported sensitivity and specificity collapsed across scales, gold standards, and samples.

Summary statistics for reported sensitivity and specificity of depression screening instruments reviewed in Sweetland et al. (2014)

			Average (SD)		Number of Studies	
Scale	Gold Standard	Sample	Sensitivity	Specificity	All	Clinical Interview
Any	Clinical interview	Any	76.4 (17.1)	80.6 (19.3)	16	16
Any	Any	Perinatal	78.0 (10.1)	76.9 (15.6)	17	4
EPDS	Any	Perinatal	82.8 (8.4)	79.1 (19.8)	8	2
PHQ-9	Any	Any	88.3 (4.8)	91.1 (13.9)	3	0

Based on these results, a reasonable expectation for specificity of our new test would be 0.80. Using the method of Chu and Cole (2007), we estimate that we would need to include 130 non-cases to ensure with 0.95 probability that the lower 95% confidence limit does not fall below 0.70, assuming power of 0.80. If the prevalence of depression in this setting is 31.3%, this would mean recruiting 60 cases for a total of 190 women.[1]

```
cases = noncases/[(1-prevalance)/prevalence]
```

Without published estimates of perinatal depression in this part of Kenya to guide us, we have to look elsewhere in the literature. Villegas et al. (2011) found a combined prevalence rate of postpartum depression of 31.3% (95% CI 21.3-43.5%) across 10

[1] If our estimate of specificity is higher as hoped, say 0.95, then we would only need to recruit 68 non-cases (31 cases for a representative sample) to ensure that the lower 95% confidence limit does not fall below 0.85 with 0.95 probability.

studies conducted in developing countries.[2]

If the sensitivity of our new test is also 0.80—a reasonable estimate from Table 1—and if we successfully recruit 60 cases, then we can ensure with 0.95 probability that the lower 95% confidence limit will not fall below 0.64. Raising this lower limit to 0.70 would require us to recruit 130 cases, and that that is not feasible given our current resources; however, if sensitivity is higher as hoped—for instance, 0.85, 0.90, or 0.95—then 60 cases would be sufficient to ensure that the lower 95% confidence limit would be 0.70, 0.75, or 0.80, respectively.

Therefore, we conclude that an optimal validation sample size is 190 women, including 60 cases and 130 non-cases. If there is attrition between the survey and the clinical assessment, we will have to recruit additional women. Attrition is unlikely to exceed 10%, so we anticipate that the maximum required validation sample would be 209.[3]

[2] If the "true" prevalence is closer to the lower confidence limit estimated by Villegas et al. (2011), then we would only need to recruit 36 cases—not 60—for a total sample of 166 women.

[3] It is possible that we could exceed this total if we struggle to find cases and inadvertently oversample non-cases in an effort to recruit 60 cases.