Statistical Analysis Plan

Project Title: Cyclical Neuroactive Steroid Changes, Arousal, and Proximal Suicide Risk: An Experimental Approach

Principal Investigator: Tory Eisenlohr-Moul, PhD

Approved as of September 24, 2019

Statistical Analysis Plan

- There are twelve total outcomes, which are specified and numbered (#) below. In order to better set up the power analysis details that follow, I have noted the anticipated intra-class correlations for each outcome (in red) and note from which preliminary data these estimates arise.

  OUTCOME SPECIFICATION

  - **Suicide Outcome Variables.** In our primary analyses, proximal suicide risk will be operationalized as (1) **daily severity of suicidal ideation** using the total score on the Adult Suicidal Ideation Questionnaire in the daily survey (ICC=.64 in K99 trial), (2) **daily suicidal planning thoughts** (dichotomous) on the daily C-SSRS phone risk screen (~10% of days were positive for planning in the K99 trial; ICC=.45 in K99 trial) and (3) **daily suicide attempt** (dichotomous; aborted, interrupted, or full) on the daily C-SSRS phone risk screen (~1.2% of days were positive for suicide attempt in the K99 trial; ICC=.12 in K99 trial).

  - **Aim 1 Mediation Outcomes (for Mediation by Hopelessness, Arousal)**. Aim 1 focuses on mediation of our experimental effects by hopelessness, hyperarousal, and their interaction. The primary hopelessness outcome will be (4) daily hopelessness on items from the Beck Hopelessness Scale (validated against full BHS in the laboratory; ICC=.32 in K99 trial). The primary arousal outcomes will be (5) daily hyperarousal as measured with the Brief Agitation Measure (validated against POMS Arousal subscale in the laboratory; ICC=.26 in K99 trial), (6) daily total smartphone accelerometer displacement reflecting overall locomotor activity (measured continuously using our BiAffect app, per reviewer recommendation; ICC=.54 in Co-I Leow's prior study in patients with bipolar disorder), and (7) daily average typing speed instability reflecting greater reactivity to external stimuli (measured using our Biaffect App; ICC=.38 in Co-I Leow's prior study in patients with bipolar disorder), (8) degree of RSA withdrawal during the Face Matching Task (no ICC estimate available for repeated administration; we assume a moderate ICC between .4-.6), and (9) degree of PEP shortening during the Face Matching Task (no ICC estimate available for repeated administration; we assume a moderate ICC between .4-.6).

- **Aim 2 Mediation Variables (Mediation by Neurosteroids and Neurosteroidogenic Enzymes).** Aim 2 focuses on mediation of our experimental effects by neuroactive steroid levels and mRNA expression for neurosteroidogenic enzymes. The primary outcomes will be (10) levels of allopregnanolone as measured using GC-MS; average ICC of .29 in former K99 mentor's prior trials with 4 measures across the cycle in naturally cycling women), (11) gene expression for 5a-RI (no ICC estimate available for repeated administration; we assume a moderate ICC between .4-.6)., and (12) gene expression for 3a-HSDII.

ANALYTIC STEPS

- **Specification of Steps for Analytic Plan:**
  For each outcome, we will use model fit indices to select the best-fitting hierarchical model before proceeding with analyses. Based on our prior study (K99MH109667) with an similar experimental design and similar data collection structure for daily surveys and lab visits, we expect that the data structure for primary analyses will be observations nested within participants (two-level regression model). In our prior models, inclusion of a third level of nesting (i.e., observations nested within condition nested within individuals) did not improve model fit and did not substantively alter hypothesis tests. Based on the prior study, we also expect that a combination of a random intercept for participant and an autoregressive (day-1) within-person error structure will be included based on model fit.

  Testing Experimental Effects: Both of our two aims require that we first evaluate the impact of the experiment on each of the 3 primary suicide outcomes and 9 (RDoC, neurosteroid) mediators described above. Therefore, we will first examine effects of the within-person hormone stabilization experiment on repeated measures outcomes across the menstrual cycle.

  a. **For daily outcomes,** we will utilize multilevel models (with 52 days—26 per condition--nested within each individual; two-level models) to predict each repeated measures outcome from covariates (daily pain, illness, PRN medication use) and the interaction between (1) **experimental condition** (a dichotomous variable coded as 0 for natural hormone withdrawal condition and 1 for hormone stabilization condition) and (2) **condition phase,** a categorical (contrast) variable coded as follows: *midluteal baseline phase*, coded as days +1 to +7 after the positive ovulation test, *perimenstrual experimental phase*, coded as days +11 to +17 after the positive ovulation test, and the *medication withdrawal phase*, coded as days +22 to +26 after positive ovulation test (see Figure 6 in grant for study design and timeline). **Number of daily observations per person**. For the primary contrast between the midluteal and perimenstrual experimental phases, there are 28 observations per person—14 per condition (7 midluteal, 7 perimenstrual). Based on our preliminary data from the K99 trial (identical experimental manipulation), we anticipate receiving an average of 24 observations per participant.
  b. **For laboratory visit outcomes,** we will again utilize multilevel models (with 6 labs per person—3 per condition—nested within individuals; two-level models). A two-level model will predict each repeated measures lab outcome from covariates (daily pain, illness, PRN medication use) and the interaction between (1) **experimental condition** (a dichotomous variable coded as 0 for natural hormone withdrawal condition and 1 for hormone stabilization condition) and (2) **condition phase,** a categorical (contrast) variable coded as follows: *midluteal baseline visit*, *perimenstrual experimental visit*, and the *medication withdrawal visit*. **Number of lab observations per person**. For the primary contrast between the midluteal and perimenstrual experimental phases, there are 4 observations per person—2 per condition. Based on our preliminary data from the K99 trial (identical experimental manipulation), we anticipate very minimal missing laboratory data (see grant for details) but have estimated that on average each completer participant will miss .25 visits.

  c. **Hypothesized Pattern of Effects for Daily and Lab outcomes**: A significant interaction between experimental condition and condition phase is hypothesized for all twelve specified outcomes such that, in the natural withdrawal condition we expect deleterious changes from the midluteal phase to the premenstrual and menstrual phases (due to corresponding hormone withdrawal), but in the hormone stabilization condition we expect no deleterious

changes (or less severe changes) from the midluteal to the premenstrual and menstrual phases (since hormone withdrawal is prevented).

2. Testing Associations Between Mediators and Suicide Outcomes: Each aim next requires that we evaluate the day-to-day or week-to-week associations between single mediators (listed above) and suicide outcome variables on the same day or visit. This will be carried out in similar multilevel models as above, predicting each of the three primary suicide outcomes from person-centered values for each of the 9 mediators (listed above).

3. Testing Interactive Effects of Hopelessness and Arousal on Suicide Outcomes: Interactions among hopelessness and arousal will be examined as predictors of suicide outcomes and as mediators of the experimental effect. In order to test this, we examine day-to-day interactive effects of daily person-centered hopelessness (noted above) and daily person-centered arousal indices (noted above) predicting same-daily suicide outcomes (noted above). We predict that higher-than-usual hopelessness on a given day will predict greater same-day risk on suicide outcomes, but only when arousal is *also* higher than usual for that individual.

4. Evaluating Mediation Where Appropriate. Finally, each aim indicates that **mediation** will be explored as appropriate. Only when analyses indicate significant effects of the experiment on both the suicide outcome and any mediator, as well as significant associations between the person-centered variable (or interaction variable) and the suicide outcome, we will utilize methods provided by **Bauer, Preacher, and Gil (2006)** to construct a confidence interval for the level 1 indirect effect of experimental condition (condition X phase) on the suicide outcome via the mediator.

1 R01 MH120843-01 (PI: Eisenlohr-Moul)
POWER ANALYSIS.

With the funds allowed in this four-year grant, we are able to increase our proposed sample size slightly, from 80 to 85 completers, which would allow us to achieve 80% power to detect conventionally small-to-medium sized effects that, in our estimation, are clinically meaningful (i.e., $f^2 = .10$, which would mean that the interaction of condition and phase had uniquely accounted for 10% of the variance in the outcome in the perimenstrual window; Selya et al., 2012, "*A Practical Guide to Calculating Cohen's* f^2*, a Measure of Local Effect Size"*). We posit that if hormone withdrawal (vs. experimental stabilization) does not account for at least 10% of the variance in the outcome during this critical risky window of the menstrual cycle, then this line of research, which involves significant participant burden, is not worth pursuing further.

- Power in multi-level models has a unique consideration beyond sample size and alpha; it is dependent also on the intraclass correlation (ICC) and its implications for the independence of level-1 observations (i.e., the design effect). High ICCs indicate little variation of the outcome within a given person (i.e., stability of repeated observations within a given person), and suggest that the study's power depends more on the number of higher-level units. Low ICCs indicate high within-person variation within a given person and suggest that the study's power depends more on the number of lower-level units. As the ICC ranges from 0 to 1, the effective n will range from the number of lower-level units to the number of higher-level units. In general, the lower the ICC, the higher the power.
- Snijders and Bosker (1999; *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*) provide equations to determine the "Design Effect", which we have used to calculate the "effective lower-level N" based on our expected ICCs for each outcome and our expected number of raw observations with 85 completer participants (including 15% missing at daily level and an average of .25 missing visits per person). Once the design effect and effective N is calculated for a given outcome, one can calculate power to detect a given effect size (in our case, $f^2 = .10$) for a given number of these *effective* lower-level units (observations or visits)—see Table below for estimates for the current design. We considered a single fixed regression coefficient in a model with 3 predictors, one-tailed test, alpha=.004.

Variable	Measurement interval	ICC	Upper-Level Sample Size	Effective Lower-Level N based on ICC and Design Effect Equation	Power Achieved to detect $f^2 = .10$
Suicidal Ideation	Daily	.64	85	130	.81
Suicidal Planning	Daily	.45	85	180	.93
Suicide Attempt	Daily	.12	85	542	.99
Beck Hopelessness Scale	Daily	.32	85	244	.98
Brief Agitation Measure	Daily	.26	85	292	.99
Accelerometer activity	Daily	.54	85	152	.88
Typing Speed Instability	Daily	.38	85	209	.96
HRV withdrawal in the lab	Labs	.50	85	163	.90
PEP shortening in the lab	Labs	.50	85	134	.83
Allopregnanolone in serum	Labs	.29	85	177	.93
PBMC gene expression for 3a-HSD	Labs	.50	85	134	.83
PBMC gene expression for 5a-reductase	Labs	.50	85	134	.83

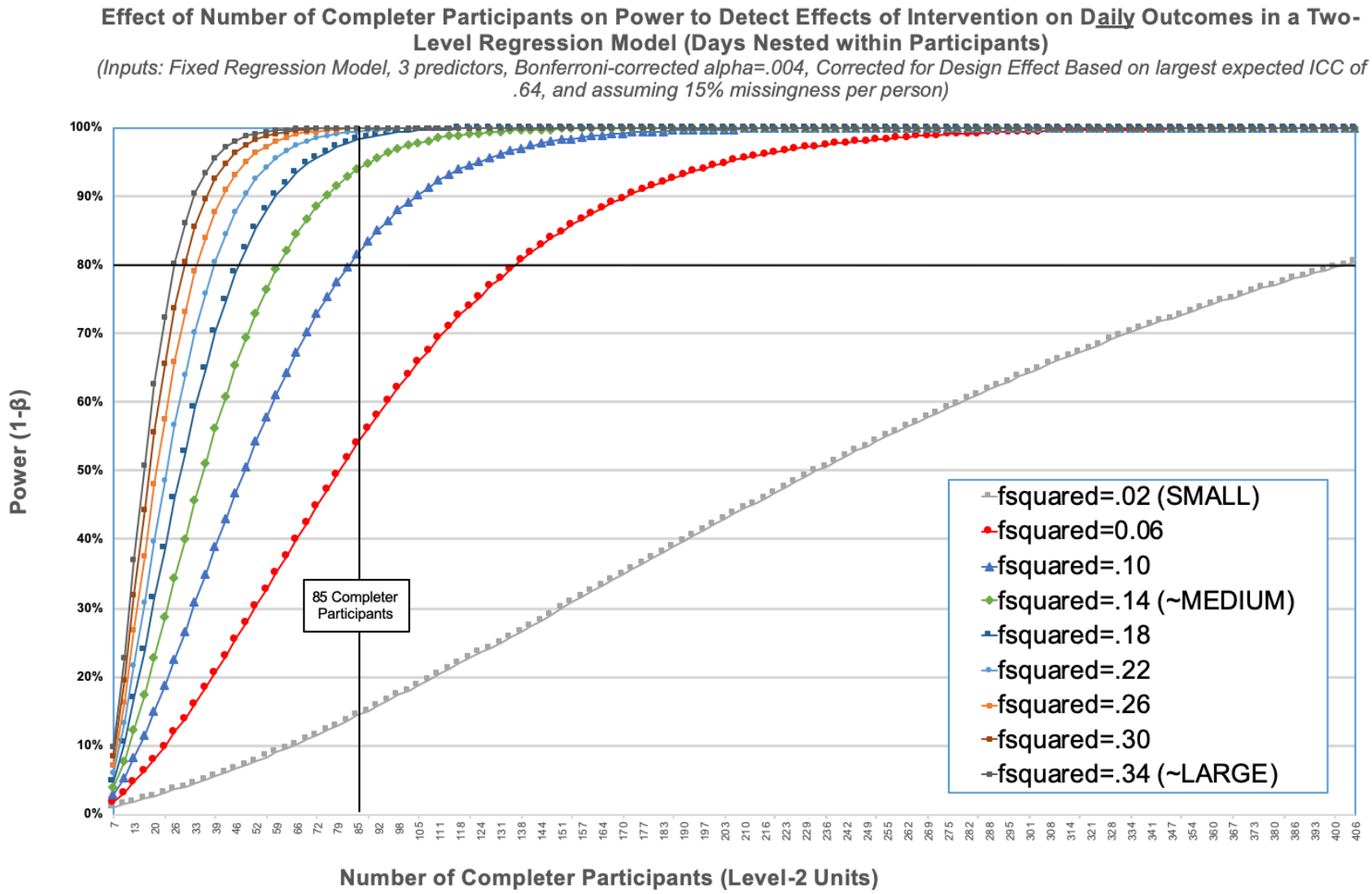
Additional Requested Supplemental Table (2019-08-16)

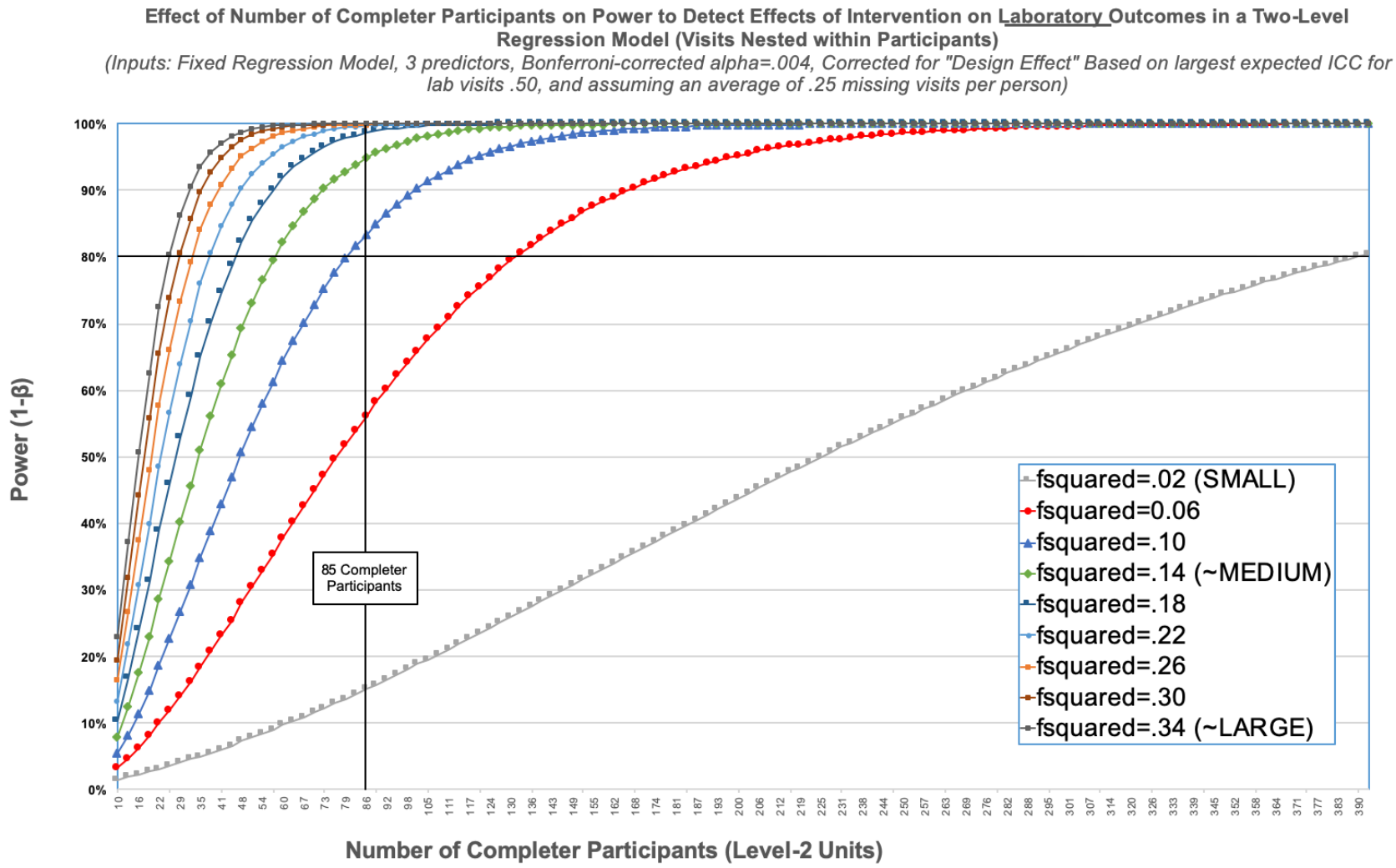
Variable	Measurement interval	ICC in pilot data	P value	f-squared	Participants Needed to Achieve 80% Power in Proposed Model*
Suicidal Ideation	Daily	**.64**	.004	.10	**83**
Suicidal Planning	Daily	**.45**	.004	.10	**57**
Suicide Attempt	Daily	**.12**	.004	.10	**21**
Beck Hopelessness Scale	Daily	**.32**	.004	.10	**45**
Brief Agitation Measure	Daily	**.26**	.004	.10	**37**
Accelerometer activity	Daily	**.54**	.004	.10	**71**
Typing Speed Instability	Daily	**.38**	.004	.10	**52**
HRV withdrawal in the lab	Labs	**.50**	.004	.10	**74**
PEP shortening in the lab	Labs	**.50**	.004	.10	**74**
Allopregnanolone in serum	Labs	**.29**	.004	.10	**53**
PBMC gene expression for 3a-HSD	Labs	**.50**	.004	.10	**74**
PBMC gene expression for 5a-reductase	Labs	**.50**	.004	.10	**74**

*First, we calculated the number of observations required using a normal GLM framework (non-multilevel). Then, we calculated how many participants would be required to achieve that number of adjusted observations after penalizing for the anticipated design effect, which adjusts for the degree of non-independence in values (Snijders & Bosker, 1999) using the anticipated intraclass correlation (ICC) for each outcome. Anticipated ICCs are based on pilot data from prior studies.

1 R01 MH120843-01 (PI: Eisenlohr-Moul)

- **Plotting Power Across a Range of <u>Participant-Level Sample Sizes</u>**. In addition, one can use the design effect equation to calculate expected power for a range of upper-level sample sizes (i.e., number of participants) adjusting for the most stringent design effect at that repeated measures interval (based on ICC=.64 for daily variables); see the figures below for power estimates across a range of possible participant-level sample sizes.


Effect of Number of Completer Participants on Power to Detect Effects of Intervention on Daily Outcomes in a Two-Level Regression Model (Days Nested within Participants)
(Inputs: Fixed Regression Model, 3 predictors, Bonferroni-corrected alpha=.004, Corrected for Design Effect Based on largest expected ICC of .64, and assuming 15% missingness per person)
Power (1-β)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
85 Completer Participants
fsquared=.02 (SMALL)
fsquared=0.06
fsquared=.10
fsquared=.14 (~MEDIUM)
fsquared=.18
fsquared=.22
fsquared=.26
fsquared=.30
fsquared=.34 (~LARGE)
Number of Completer Participants (Level-2 Units)


Effect of Number of Completer Participants on Power to Detect Effects of Intervention on Laboratory Outcomes in a Two-Level Regression Model (Visits Nested within Participants)
(Inputs: Fixed Regression Model, 3 predictors, Bonferroni-corrected alpha=.004, Corrected for "Design Effect" Based on largest expected ICC for lab visits .50, and assuming an average of .25 missing visits per person)
Power (1-β)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
85 Completer Participants
fsquared=.02 (SMALL)
fsquared=0.06
fsquared=.10
fsquared=.14 (~MEDIUM)
fsquared=.18
fsquared=.22
fsquared=.26
fsquared=.30
fsquared=.34 (~LARGE)
10 16 22 29 35 41 48 54 60 67 73 79 86 92 98 105 111 117 124 130 136 143 149 155 162 168 174 181 187 193 200 206 212 219 225 231 238 244 250 257 263 269 276 282 288 295 301 307 314 320 326 333 339 345 352 358 364 371 377 383 390
Number of Completer Participants (Level-2 Units)