1. To choose a metric confidently I would need to be better acquainted with the details of Ultimate's business model and accounting practices. However, the general idea would be to ascertain whether the reimbursements are likely to generate enough revenue for Ultimate to justify the cost. The specifically relevant measure would depend, for example, on what procedures determine how much revenue from a ride goes to the driver and how much goes to the firm, as well as perhaps on how the timing of rides affects the price. Ideally I would want to measure several quantities to get a better idea of what is going on, and to allow for the possibility that, even if the program isn't cost-effective under current procedures, there might be adjustments that could be made to make it beneficial. As an initial approach, if there is a measure of revenues to the firm, that could be compared against the costs of reimbursing tolls, that would be an appropriate measure.

2. (a) During alternate weeks (just Monday through Friday), offer toll reimbursements, alternating over, say, an initial 16-week period. (After the first 16 weeks, we could perform a power analysis on the data collected, to determine if it would be useful to continue the experiment.) Measure the amount of tolls reimbursed during reimbursement weeks, as well as the amount of revenue generated for the firm in both reimbursement and non-reimbursement weeks. Match each reimbursement week with the preceding non-reimbursement week and calculate the difference in revenues. Subtract the amount of tolls reimbursed from that to obtain a measure of success. Once the pilot experiment is complete, there will be 8 observations. Possibly we will need to adjust for the presence of holidays in certain weeks (maybe by just using Tuesday through Friday, if all the holidays are on Monday, or maybe by skipping holiday weeks and just alternating non-holiday weeks).

(b) Do a t-test of the null hypothesis that the population mean for revenue-difference-minus-toll-reimbursements is zero. Also do a binomial test of the null hypothesis that the probability of a positive difference is 0.5. (I am assuming that the observed differences turn out to be positive. If not, the tests could be done in the other direction, in which case a rejection would be considered a failure and a non-rejection would prompt a power analysis and a decision as to whether continuing the experiment would be worthwhile.)

(c) If both tests reject their respective null hypotheses (assuming positive results), report that the experiment has been a success, and recommend that the toll reimbursement program be made permanent, subject to a review of the data to see if there are any relevant issues that we may have missed. If the t-test rejects but the binomial test doesn't, take a closer look at the data distribution to see if the assumptions of the t-test appear to hold.

If neither test rejects, do a power analysis to determine how much more data would be required to come to a definite conclusion. The operations team can come to a decision about whether it will be worthwhile to extend the experiment, and perhaps, after taking a closer look at the data, I will be able to suggest an alternative experiment that could be performed.