Reflections on the Development of "Calibrate Your Judgment"

Luke Muehlhauser
Open Philanthropy Project
December 2018

Starting in mid-2016, the Open Philanthropy Project contracted Spark Wave (led by Spencer Greenberg) to develop a probability calibration training application as part of a variety of efforts to improve the accuracy of our judgments and forecasts. The resulting app, *Calibrate Your Judgment*, was released publicly in December 2018.

We expect many users will find this program to be the most useful free online calibration training currently available.

That said, we think there are several ways in which a calibration app could be more engaging and useful than ours is, if someone were to invest substantially more development effort than we did. This document explains some of our ideas for how one might develop a better calibration app, and also explains some lessons we learned during the development process.

Lessons learned

Our first lesson was that creating a fully engaging probability calibration training app likely requires far more effort than we initially guessed. For example:

- We initially hoped that a question set capable of being automatically generated (such as the scatterplot module and the math problems module) would be sufficiently engaging, but in testing we concluded that real-world questions would be more engaging (and probably better for training). But real-world questions must be manually sourced and curated.
- Even after licensing thousands of pre-curated trivia questions from The Question Company, we found that most of the questions are likely not in the "sweet spot" between "too easy" and "too hard / no information" for most users.
- Once we learned that the "confidence interval" module seemed most engaging and perhaps also best for learning (e.g. because it allows one to train on one confidence level at a time), we hoped that a relatively straightforward adaptation of proper scoring rules to a prediction interval context would be possible. However, we soon learned that this problem is to some degree unsolved, and straightforward adaptations have counterintuitive properties that are problematic from a user experience perspective (see Greenberg 2018).
- The technical challenge of managing the status of thousands of questions for (potentially) thousands of users required more engineering work than we'd initially expected.
- More generally, we initially thought the app we wanted was simple enough that we could specify it almost entirely in advance. However, we now think the design of a fully engaging and useful probability calibration app is complicated challenge, one that should be approached via many rounds of idea development and user testing.

We also learned some things about how to manage a software development process better. In particular:

- This project was never a priority either for Open Philanthropy or for Spark Wave, and both parties had limited time available to devote to it each month. As a result, the project frequently languished for weeks at a time. If we want to make faster progress on a future software development project, we'll need to make sure that both the client (Open Philanthropy) and the developer can make the project a priority and devote substantial time to it.
- We should have scheduled specific milestone deadlines with Spark Wave from the beginning. Not doing so slowed progress.

Ideas for how to build a better calibration app

We now realize that designing a fully engaging and useful calibration app is a complex challenge, one that should be approached via many rounds of idea development and user testing. Below are some ideas I (Luke) have for how to build a better calibration app, though of course they are all subject to user testing results:

1. Extant and novel trivia questions could be tested (e.g. via Mechanical Turk) to learn which ones are in a "sweet spot" between "too easy" and "too hard" for the intended user base.
2. Large-scale user tests could reveal which scoring rules are best for user experience and rapid learning (without being gameable).
3. Various interface improvement could make the app more intuitive and engaging to use, and the app could be made natively compatible with more platforms.
4. Various "gamification" methods (badges, leaderboards, competitive play, etc.) might further incentivize app use and learning.
5. A continuously updating set of questions about current events could be more engaging, and relevant to real-world calibration, than classic trivia questions are.
6. One could also incorporate prospective forecasting into a calibration app — i.e. questions to which the answers aren't yet known, but they will be judged as having happened or not in the future. Perhaps this could be achieved in collaboration with an existing prediction tournament platform such as PredictIt or Good Judgment Open.
7. An interactive tutorial could help introduce users to the basic concepts of prediction and calibration, and give tips for how to improve, perhaps in response to the user's performance on the app so far.