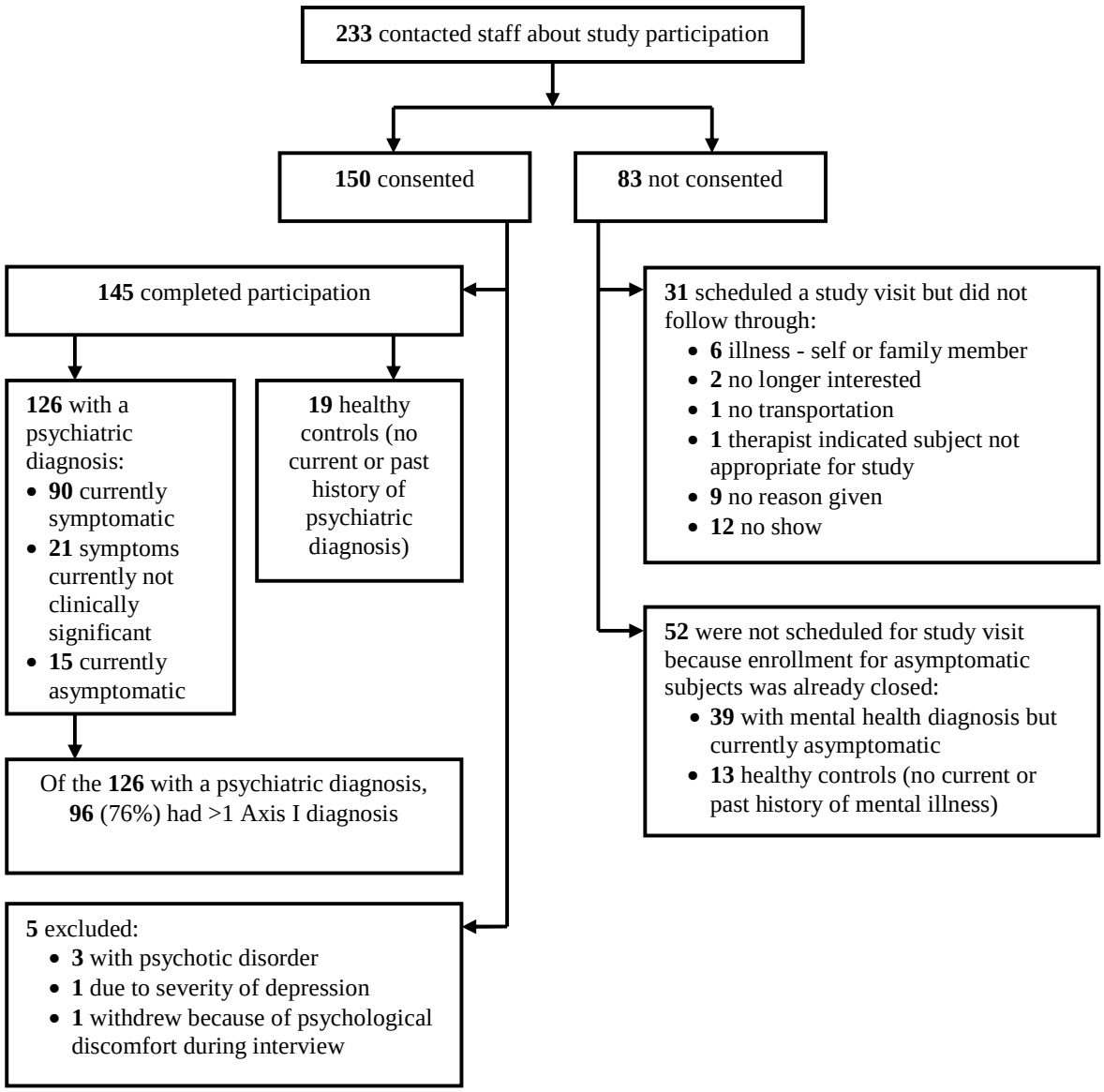
VALIDATION OF COMPUTERIZED ADAPTIVE TESTING IN A NON-ACADEMIC SETTING

eFigure 1. Enrollment



Online Supplemental Appendix

The Logic of Item Response Theory (IRT): Classical and IRT methods of measurement differ dramatically in the ways in which items are administered and scored. The difference is clarified by the following analogy. Imagine a track and field meet in which ten athletes participate in the 110-meter hurdles race and also in the high jump. Suppose that the hurdles race is not quite conventional in that the hurdles are not all the same height and the score is determined, not only by the runner's time, but also by the number of hurdles successfully cleared, *i.e.*, not tipped over. On the other hand the high jump is conducted in the conventional way: the cross bar is raised by, say, 2 cm increments on the uprights, and the athletes try to jump over the bar without dislodging it. The first of these two events is like a traditionally scored objective test: runners attempt to clear hurdles of varying heights analogous to questions of varying difficulty that examinees try to answer correctly in the time allowed. In either case, a specific counting operation measures ability to clear the hurdles or answer the questions. On the high jump, ability is measured by a scale in millimeters and centimeters at the highest scale position of the cross bar the athlete can clear. IRT measurement uses the same logic as the high jump. Test items are arranged on a continuum at certain fixed points of increasing difficulty. The examinee attempts to answer items until she can no longer do so correctly. Ability is measured by the location on the continuum of the last item answered correctly. Hence, in IRT, ability is measured by a scale point, rather than a numerical count.

These two approaches to measurement contrast sharply: if hurdles are arbitrarily added or removed, number of hurdles cleared cannot be compared. The same is true of traditional number-right scores of objective tests: scores lose their comparability if item composition is changed. The same is not true, however, of the high jump or of IRT scoring. If positions on the

bar are omitted, height cleared is unchanged and only the precision of the measurement at that point on the scale is affected. Similarly, in IRT scoring of tests, a certain number of items can be arbitrarily added, deleted or replaced without losing comparability of scores on the scale. This property of scaled measurement, as opposed to counts of events, is the most salient advantage of IRT over classical methods of psychological measurement.

Computerized Adaptive Testing (CAT): Consider a 1000-item mathematics test with items ranging in difficulty from basic arithmetic through advanced calculus. We now test two examinees, a fourth grader and a graduate student in mathematics. Most questions will be uninformative for both examinees (too difficult for the first and too easy for the second). To decrease examinee burden, we could create a short test of 10 items, equally spaced along the mathematics difficulty continuum. While this test would be quick to administer, it would provide very imprecise estimates of these two examinees' abilities because only an item or two would be appropriate for either examinee. A better approach would be to begin by administering an item of intermediate difficulty, and based on the response scored as "correct" or "incorrect" select the next item at a level of difficulty either lower or higher. This process would continue until the uncertainty in the estimated ability is smaller than a predefined threshold. When this process is automated and administered by a computer, it is called computerized adaptive testing (CAT). To use CAT, we must first calibrate a "bank" of test items using an IRT model that relates properties of the test items (e.g., their difficulty and discrimination) to the ability (or other trait, such as severity of depression) of the examinee. As such, the 'difficulty' or 'severity level' of an item refers to how indicative the item is of severity of illness. Thus an item such as 'I sometimes feel sad' would have a low level of severity since almost every human being would endorse it, while

an item such as ‘ I am so distraught that I plan to take my life’ would have a high level of severity because it is only endorsed by the most severely depressed patients. The paradigm shift is that rather than administering a fixed number of items that provide limited information for any given participant, we adaptively administer a small but varying number of items (from a much larger “item bank”) which are optimal for the participant’s specific level of symptom severity. CAT for the bifactor model has been described by Gibbons and colleagues.[1,2]

Computerized Adaptive Diagnosis: Diagnosis and measurement represent very different processes. While IRT is ideal for measurement, it is not ideal for diagnostic screening where an external criterion is available (e.g. a SCID-based DSM diagnosis of MDD). Decision trees[3] represent an attractive framework for designing adaptive predictive tests since their corresponding models can be represented as a sequence of binary decisions. Despite this intuitive appeal, decision trees have suffered from poor performance, largely as a result of variance associated with the specific algorithms used to estimate them and the limited modeling flexibility of small trees. Instead, models constructed of averages of hundreds of decision trees, called random forests, have received considerable attention in statistics and machine learning.[4] These models provide significant improvements in predictive performance, but lack the adaptive test structure inherent in individual trees. The important distinction between CAT and CAD is that in CAD we a have an external criterion or “gold standard” which in our case is a DSM diagnosis. Random forests allow us to create an adaptive sequence of items which provide a binary diagnostic decision with the smallest number of items and highest level of confidence for predicting the gold standard. By contrast CAT allows us to adaptively administer a small set of optimally selected items which maximally preserves the information contained in the entire item

bank and therefore the CAT-based score will be highly correlated with the entire item bank score. No external criterion is *needed* for CAT, however, an external criterion is *useful* for demonstrating the validity of a CAT-based severity score.

1. Gibbons RD, Hedeker D: Full-information item bifactor analysis. Psychometrika 57:423-436, 1992.

2. Gibbon RD, Bock RD, Hedeker D, et al: Full-information item bifactor analysis of graded response data. Applied Psychological Measurement 31:4-19, 2007.

3. Brieman L, Friedman JH, Olshen R, et al: Classification and Regression Trees. Wadsworth, 1984.

4. Brieman L. Random Forests. Machine Learning 45:5-32, 2001.