A Web Application For Psychometric Tests

Julian Applebaum, Steven Stemler, Ph.D
QAC Summer 2011, Wesleyan University

Introduction

In a psychometric lab, effective data collection depends on the ability to build tests and reliably administer them to participants. Historically, this has been accomplished with a with a combination of mail, email, and basic productivity software. While more specialized tools do exist, no available software fully addressed the needs of Wesleyan's Stemler Lab. The Stemler Lab requires software that can manage a large item pool, build a variety of assessments, interrupt tests with survey items, display multimedia, and quickly distribute tests worldwide.

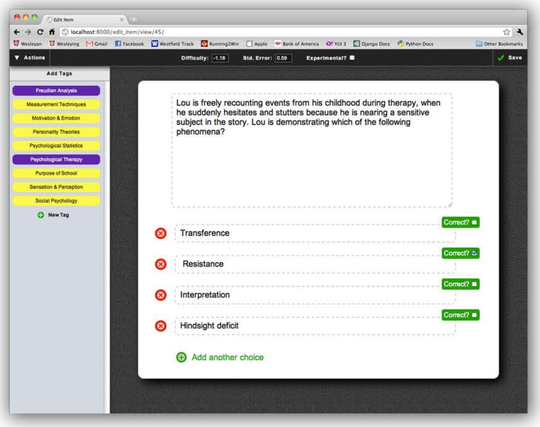

Figure 1: The interface for editing individual items

Goals

The primary goal of this ongoing project is to build a web application that meets all of the Stemler Lab's needs, functioning as a specialized research tool for psychometric assessments of any scale. The software also seeks to provide a reliable and valid implementation of Computer Adaptive Testing (CAT) for use in future studies.

At the present time, the software can build a limited number of assessments, manage an item pool, interrupt with survey items, and distribute tests to any location. It also has an experimental CAT implementation whose behavior is currently being evaluated and adjusted.

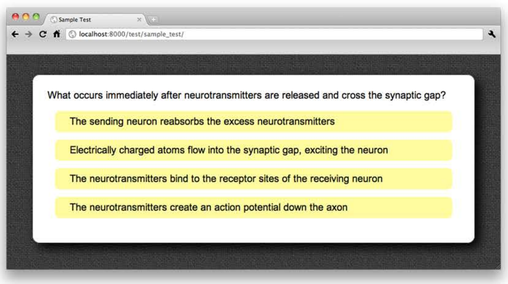

Figure 2: An item being administered during a test

Implementation

There are three main interfaces that a researcher interacts with when developing a test with this software: the item editor, the item pool view, and the test builder. The item editor (figure 1) allows a researcher to create a multiple choice item with any number of keys and distractors. In addition, the researcher can add metadata - such as keyword tags, difficulty rating, and standard error - that is needed to organize items and construct individual test sections.

The item pool view (not pictured) allows the researcher to rapidly search for items by tag, keyword, experimental status, difficulty rating, standard error, and number of participants. Having a powerful search makes maintaining multiple item pools a much more feasible task - an essential feature for a lab running multiple studies simultaneously.

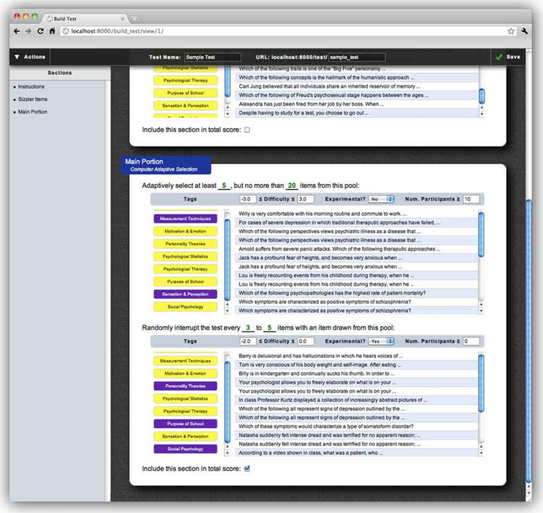

Figure 3: The interface for building a test

Currently, the test building interface (figure 3) allows the researcher to construct a test consisting of three sections: one that displays instructions, one that randomly draws from a subset of the item pool, and one that implements Computer Adaptive Testing. To aid data collection, the adaptive section also allows itself to be periodically interrupted with survey and/or experimental items that will not be included in the test's final score.

Both the random and adaptive sections' item choices can be limited using the same search criteria available in the item pool view. In future versions, researchers will be able to insert any number of sections in an arbitrary order, making the software a powerful tool for experimenting with new measurement instruments.

CAT Methodology

One of the major goals of this software is to explore Computer Adaptive Testing (CAT), a recently developed method capable of faster, more accurate measurement than traditional testing methods. Instead of administering a fixed set of items, a CAT continually attempts to calculate a participant's ability stimate $\theta' = \ln(\frac{\%\ \text{correct}}{\%\ \text{incorrect}})$[1], which it then uses to choose an appropriately difficult item for that person to respond to. These two steps, referred to together as the Item Selection Algorithm, repeat until a Termination Criterion is met, at which point the test is complete and a final score can be calculated.

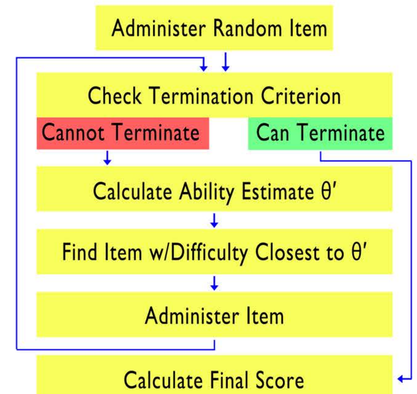

Figure 4: The general algorithm for administering a CAT

To accurately estimate a participant's true ability θ, a CAT begins by taking coarse guesses at the appropriate θ′ - without more information, it cannot yet decide very "intelligently". As the test continues and more information is collected, the calculation of θ′ becomes progressively more accurate, and the difference between θ and θ′ begins to dwindle. A well behaved CAT will terminate when θ′ has converged to within a 95% confidence interval of θ.[2] At this point, the participant's probability of a correct response will be roughly .5, which in the Rasch Model indicates a perfect match between item and participant.

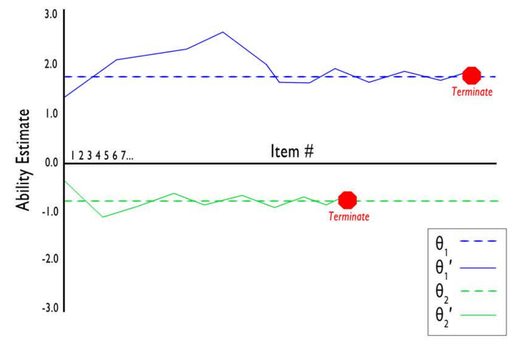

Figure 5: Illustration of CAT converging on true ability θ of two participants 1 and 2. Note that the number of items required is not constant - it can vary between participants.

1) Bond, Trevor G., and Christine M. Fox. "Appendix A." Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahwah, NJ: Lawrence Erlbaum Associates, 2007. 277-78. Print.
2) "Computerized Adaptive Testing (CAT) Overview." National Council of State Boards of Nursing. 2011. Web. 24 July 2011