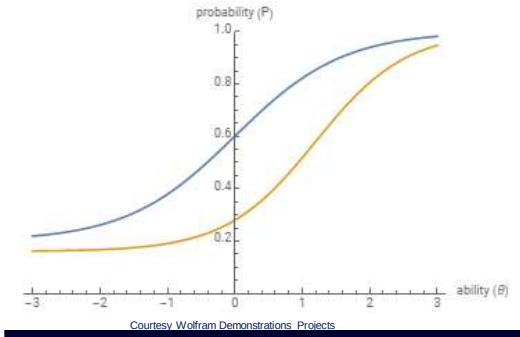
Tutorial: How to Develop and Implement Unidimensional Computer Adaptive Tests

A Tutorial



Courtesy Wolfram Demonstrations Projects

Scott K. Burtnick (U.S. Customs and Border Protection)
Jeffrey M. Cucina (U.S. Customs and Border Protection)
Kevin A. Byle (U.S. Customs and Border Protection)

The views expressed in this paper are those of the authors and do not necessarily reflect the views of U.S. Customs and Border Protection or the U.S. Federal Government.

Who we are: US Customs and Border Protection

- US Customs and Border Protection (CBP)
  - America's unified border agency
  - Twin goals of anti-terrorism and facilitating legitimate trade and travel
  - Secures 328 ports of entry into US and borders between ports
  - Prevents narcotics, agricultural pests, smuggled goods and inadmissible visitors (e.g., aliens with outstanding criminal warrants) from entering US
- Personnel Research and Assessment Division (PRAD)
  - Part of CBP's Office of Human Resources Management
  - Group of I/O psychologists who design, develop, validate, and implement wide range of competency-based assessments and conduct organizational development work (e.g., survey research)
    - Entry-level and promotional assessments

U.S. Customs and Border Protection

Who we are:
US Customs and Border Protection

Scott K. Burtnick

- Personnel Research Psychologist
- Expertise in psychometrics, test development, item response theory, differential item functioning, and data analysis.
- Leads survey research and conducts psychometrics analyses for CAT programs at U.S. Customs and Border Protection
- Previous work experience as psychometric lead for DoD security and intelligence certifications at Global Skills X-Change and conducting research for the U.S. Merit Systems Protection Board

Jeffrey M. Cucina

- Personnel Research Psychologist
- Expertise in psychometrics, test development, item response theory, data analysis, and individual differences research.
- Psychometric lead for CAT programs at U.S. Customs and Border Protection.
- Leads criterion-related validity studies, test development, and job analyses.

U.S. Customs and Border Protection

Who we are:
US Customs and Border Protection

Kevin A. Byle

- Supervisory Personnel Research Psychologist
- Expertise in psychometrics, test development, item response theory, and data analysis.
- Leads job analyses, test development and validation, and survey research at U.S. Customs and Border Protection.

U.S. Customs and Border Protection

Overview of Tutorial

This tutorial covers the development of unidimensional computer adaptive tests (CATs) for dichotomously scored items

1. Overview of CATs
2. Brief review of item response theory (IRT)
3. Item pool development
4. Monte Carlo simulations for selecting algorithms
5. Scaling/equating/metric issue
6. Experimental item collection strategies
7. Creating instructions for programmers
8. User testing
9. Implementing CATs
10. Lessons learned/things to consider

Overview of CAT

- Computer Adaptive Testing (CAT):
  - Built from a large item pool, which is different than static tests which have the same set of items that all test takers complete.
  - A form of computer-based testing that adjusts to a test taker's ability level item to item.
  - The next set of items to be drawn from is dependent on the response to the previous item.
- Typically used for large scale testing programs, as CAT requires a high amount of time and resources to develop and maintain.

Overview of CAT

- **Advantages vs. Static Testing**
  - **Testing Time:**
    - Has the potential to reduce testing times by estimating ability levels with fewer items than static tests.
  - **Test Security:**
    - Offers more test security because items are drawn from very large test banks, and each test taker receives a different set of items.
    - This is becoming increasingly important as organizations increasingly are shifting to more mobile or remote forms of employment testing.

Overview of CAT

- **Disadvantages vs. Static Testing**
  - Requires a high degree of time and resources to develop.
  - The number of items that need to be developed and pre-tested compared to static tests is typically about 5:1.
  - Large sample sizes are required to pre-test items, which can become complicated given the number of items needed for an item pool.

Overview of IRT

- Item Response Theory (IRT) uses item parameters to evaluate items on an Item Characteristic Curve (ICC).
  - *a* (item discrimination)
  - *b* (item difficulty)
  - *c* (pseudoguessing)
- IRT relies on several assumptions that must be tested before computing IRT item parameters and developing CATs.
  - Dimensionality (one vs. many)
  - Examinee independence
  - Test item independence

Overview of IRT

- Several considerations should be made when reviewing the ICCs including whether:
  - The set of items contains an adequate distribution of item difficulties.
  - Item discrimination meets key thresholds.
  - The pseudoguessing parameter is appropriate.
- Examining ICCs will help you decide whether to retain or discard items from the item bank, as well as whether a two or three parameter logistic model will best fit the data.

Item Pool Development

- In order to develop a CAT, an item pool of “hundreds, and possibly thousands” of items is needed (Hambleton, Swaminathan, & Rogers, 1991, p. 149).

- Data for each item would need to be collected on large samples:
  - 1-PLM (100-500 examinees; Wright, 1977; de Ayala, 2009, p. 42)
  - 2-PLM (500 examinees with 20 or more items; de Ayala, 2009, p. 105)
  - 3-PLM (BILOG works well with 1,000 examinees and 20 items [Mislevy, 1986]; 1,000+ examinees “strongly recommended” [de Ayala, 2009, p. 131])

Item Pool Development

- Of course, the rules-of-thumb on the previous page are guidelines.
  - What is needed for your CAT depends on your goals (e.g., unproctored CAT would likely need a large number of items).
  - Regardless, you need to collect data on a lot of items using a lot of examinees (many more than for a static test).
- Also need to consider that only a percentage of items that will be developed will survive content reviews and item analyses.
  - We suggest looking at past performance of your item writers for the particular content domain and assuming that an extra 10% of items might need to be dropped solely for IRT calibration reasons.

U.S. Customs and Border Protection

Item Pool Development

- Let's suppose you want to develop an unproctored version of a 30-item cognitive static test.
  - Using 3-PLM with 5 item pools of 150 items each.
  - Assuming 75% of items would be retained after analyses, a total of 1,000 (150 × 5 ÷ .75) items need to be administered to 1,000 examinees each.

Item Pool Development

- Source of examinees

1. Operational examinees: Collect data on experimental items using the individuals who are operationally taking the test (e.g., job applicants, certification test examinees).
   - Pros: This could be the best approach psychometrically, especially if the current pool of operational examinees matches those of future pools.
   - Cons:
     - A potential issue might occur if the number of operational examinees is small.
       - Suppose there are 1,000 operational examinees per year and 20 experimental items per examinee – it would take 10 years to collect data for one CAT.
     - Might require increasing testing time or holding scores (more on this later).

Item Pool Development

- Source of examinees

2. Convenience sample with incumbents: Collect data using people who are currently in the position (e.g., current onboard employees, established certified professionals).
   - Pros: Might be able to collect data from a large population, without being dependent on examinee volume.
   - Cons:
     - Could be problematic if range restriction is present. Need good representation at the low end of ability to estimate c parameter.
     - Incumbents might have lower motivation than operational examinees.
       - Response rate might be low.
       - Effort put into responding might be lower than examinees.
     - Might incur significant salary costs and consume staff time.

Item Pool Development

- Source of examinees

3. Convenience sample with non-incumbents: Collect data using people who are not currently in the position (e.g., use incumbents in similar occupation, undergraduates, Mechanical Turk).

    - Pros: Might be able to collect data from a large population very quickly with little cost.
    - Cons:
        - The convenience sample might not be reflective of the operational examinees
            - Differences in ability distributions, item responses, test-taking motivation, etc. could be problematic.
            - Manifestations of the target construct might also differ (e.g., if CAT will measure specialized knowledge, then couldn't collect data from the general population).

Item Pool Development

- Data collection strategy

1. Common item/non-equivalent groups: Administer an anchor test along with different batches of experimental items. Rotate through different batches until 1,000 examinees have taken each item.
   - The anchor items could be the operational test.
   - Could also shorten the operational test and use experimental items to get full length
     - This would require holding scores and conducting item analyses and equating for each batch of items.
   - Need to consider examinee test-taking fatigue (although it may be no different than typical testing situations involving common item/non-equivalent groups equating).

Item Pool Development

- Data collection strategy

1. Common item/non-equivalent groups: Administer an anchor test along with different batches of experimental items. Rotate through different batches until 1,000 examinees have taken each item.

    - From an IRT perspective, you would conduct separate calibrations on each combined anchor/experimental test. Next, you would apply a scale transformation to place the theta, a, and b parameters on the same scale (more on this later).

    - Hambleton, Swaminathan, and Rogers (1991, p. 129) mention this is typically the most feasible design for IRT purposes.

Item Pool Development

- Data collection strategy

2. Spiraling: Randomly assign examinees to take either operational test or a form containing only experimental items.

    - Would need to hold scores and conduct item analysis and equating (to place experimental items on scale of operational test) each time a new form of experimental items is introduced.

    - Wouldn't increase test-taking fatigue and could finish data collection quicker; however, it requires more item analysis and equating work to produce scores for operational purposes.

        - For each administration, would have to conduct item analysis and equating to produce operational scores for the experimental items.

Item Pool Development

- Data collection strategy

2. Spiraling: Randomly assign examinees to take either operational test or a form containing only experimental items.

   - From an IRT perspective, when spiraling is used, the item parameters and thetas for the two groups within one administration should be on the same scale (due to the random assignment; Kolen & Brennan, 2004, p. 166). However, this only works for the first administration when collecting experimental CAT items.

   - In subsequent administrations, would need to treat item parameters for operational (or already IRT-analyzed) test as fixed and conduct transformation of thetas and a and b parameters for new test to old test's scale.

Item Pool Development

- Data collection strategy

3. Single group: One single group of examinees takes operational items and all experimental items. Could counterbalance order of operational and experimental items if desired.

    - We mention this approach for completeness. Unless the item pool will be very small or the items can be completed very quickly with little examinee fatigue, this approach is not going to be practical.
    - From an IRT perspective, only one calibration would be conducted using both the operational and experimental items. No scale transformation would be needed.
    - Common persons equating is a variant of this method. A single group takes both sets of items and two separate groups take either experimental or operational items.

Item Pool Development

- Data collection strategy

4. Harvesting old data: Use data from established forms to create an item bank.

   - This approach assumes you have data for enough different forms of the test to create the CAT item pool. This might be possible for a large-scale testing program with a large number of equated forms and examinees.
   - Equating strategies 1-3 listed on the previous slides would have to have been implemented when the data were originally collected.
   - Would conduct IRT analyses on the existing data.
   - Need to consider scale drift and changes to the examinee population over time (e.g., item parameters for data collected 20 years ago might not be appropriate today).

Item Pool Development

- Data collection strategy

5. Combined static and CAT data collection: First collect data (using strategies 1-4) for experimental items for one CAT item pool using a static test. Next, develop the CAT and collect data for subsequent CATs by administering static experimental items alongside the CAT.

   - This approach assumes you are developing multiple CATs for a single test. Rather than collecting data on enough experimental items to develop multiple CATs, you first collect data for one CAT and implement it. After it is implemented, you collect data for the subsequent CATs.

   - From an IRT perspective, the pre-CAT data collection and calibration is the same as for strategies 1-4. After the CAT is implemented, a CAT pretest item calibration approach is used to place the experimental item parameters on the same scale as the operational CAT (more on this later).

Item Pool Development

- Example for Strategy 1 (Common item/non-equivalent groups): Placing parameters from different calibrations onto same scale.
- Suppose you collected data for 20 batches of 10 experimental items and had 30 common items across all administrations.
- Each batch of 10 experimental items and 30 common anchor items would be calibrated separately.
- This yields 20 different IRT calibrations. Note that the scales for each calibration are different. For example, the mean theta will be 0 in all calibrations, despite the fact that the examinees might have differed in ability levels across the different administrations.
  - This is because by default, BILOG and most other IRT programs assign a mean theta value of 0 for each calibration.

Item Pool Development

- You need to place the parameters from the 20 different IRT calibrations onto the same scale using the 30 common items as anchors.
  - You will want to choose 1 of the 20 calibrations to be the base form/calibration and transform the item parameters from each of the 19 remaining calibrations to those on the base.

- The below equations can be used to place the a and b parameters for two different tests (X and Y) on a common scale. The equations use the a and b parameters obtained from the separate calibrations of the common items (denoted c):

$$b_{Yc} = \alpha b_{Xc} + \beta$$

$$a_{Yc} = \frac{a_{Xc}}{\alpha}$$

(Hambleton, Swaminathan, & Rogers, 1991, p. 129)

Item Pool Development

- Before we can use the equation, we need to estimate α and β. There are a few different methods:

- Regression: Create regression equation to predict item parameters for test Y using those for test X. There is a symmetry issue since these will not be the same as for predicting X using Y; therefore, this method is not widely used.

- Mean/Sigma: Examines differences in means and SDs of anchor item parameters to compute α and β:

$$\alpha = \frac{SD_{Yc}}{SD_{Xc}}$$

$$\beta = \overline{b_{Yc}} - \alpha\overline{b_{Xc}}$$

Item Pool Development

- Mean/Mean: Used in 1-PLM Rasch Model:

$$\beta = \overline{b_{Yc}} - \overline{b_{Xc}}$$

- Robust Mean/Sigma: Similar to the Mean/Sigma method; however, it takes into account standard errors in estimating item parameters using a 5-step process (Linn et al., 1981). Stocking and Lord's (1991) adaptation takes into account outliers.

- Test Characteristic Curve: Iterative process that incorporates differences in discrimination (a) parameters (other methods focus on b) using test characteristic curve.
  - There are two variations of this: Haebara (1980) and Stocking and Lord (1983)
  - Overall, the Stocking and Lord (1983) method seems to be the preferred method in the literature (de Ayala, 2009; Kaskowitz & de Ayala, 2001; Baker & Al-Karni, 1991).

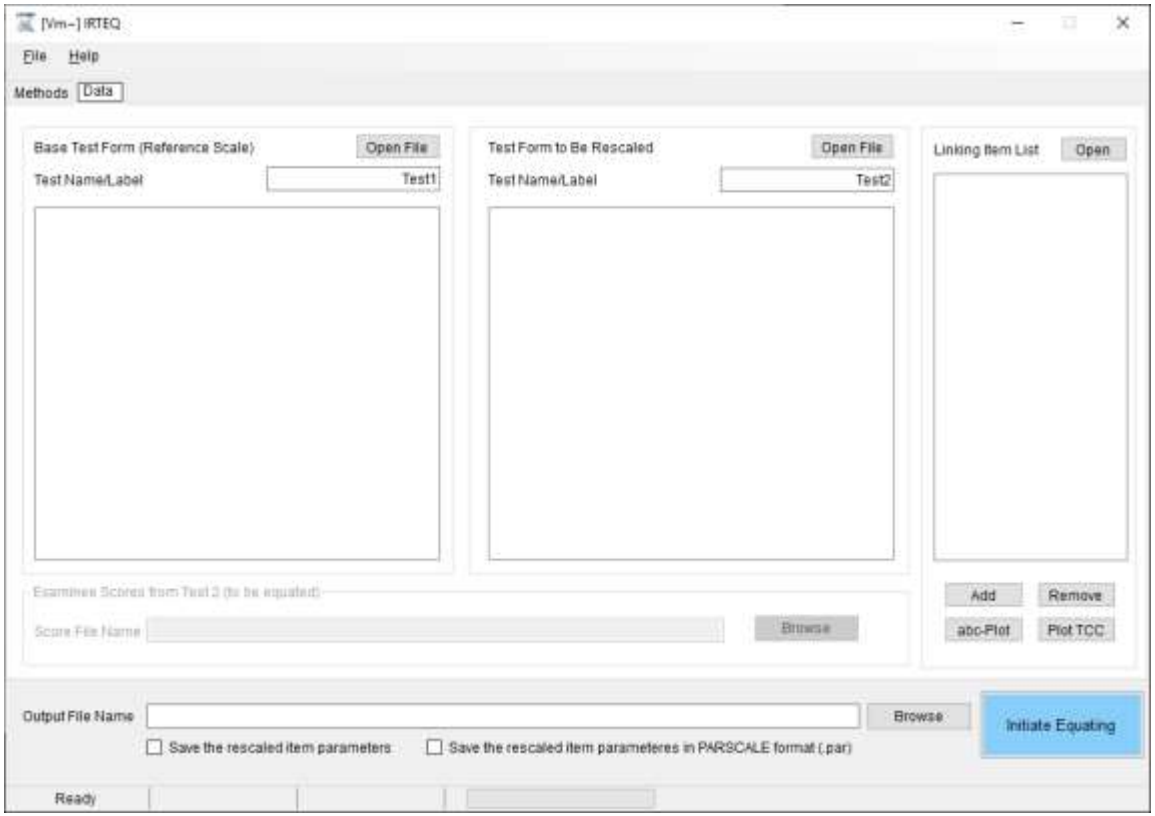
Item Pool Development

- There are a number of programs that can be used to implement these methods:
  - EQUATE (Baker, Al-Karni, & Al-Dosary, 1991) and EQUATE 2.0 (1993).
  - equateIRT R package (Battauz, 2018; Wiberg, 2018)
  - We prefer IRTeq (Han, 2009), a free windows-based program that has a GUI and syntax option:

https://www.umass.edu/remp/software/simcata/irteq/

Item Pool Development

- Using IRTeq
  - First open IRTeq and navigate to the Data tab:



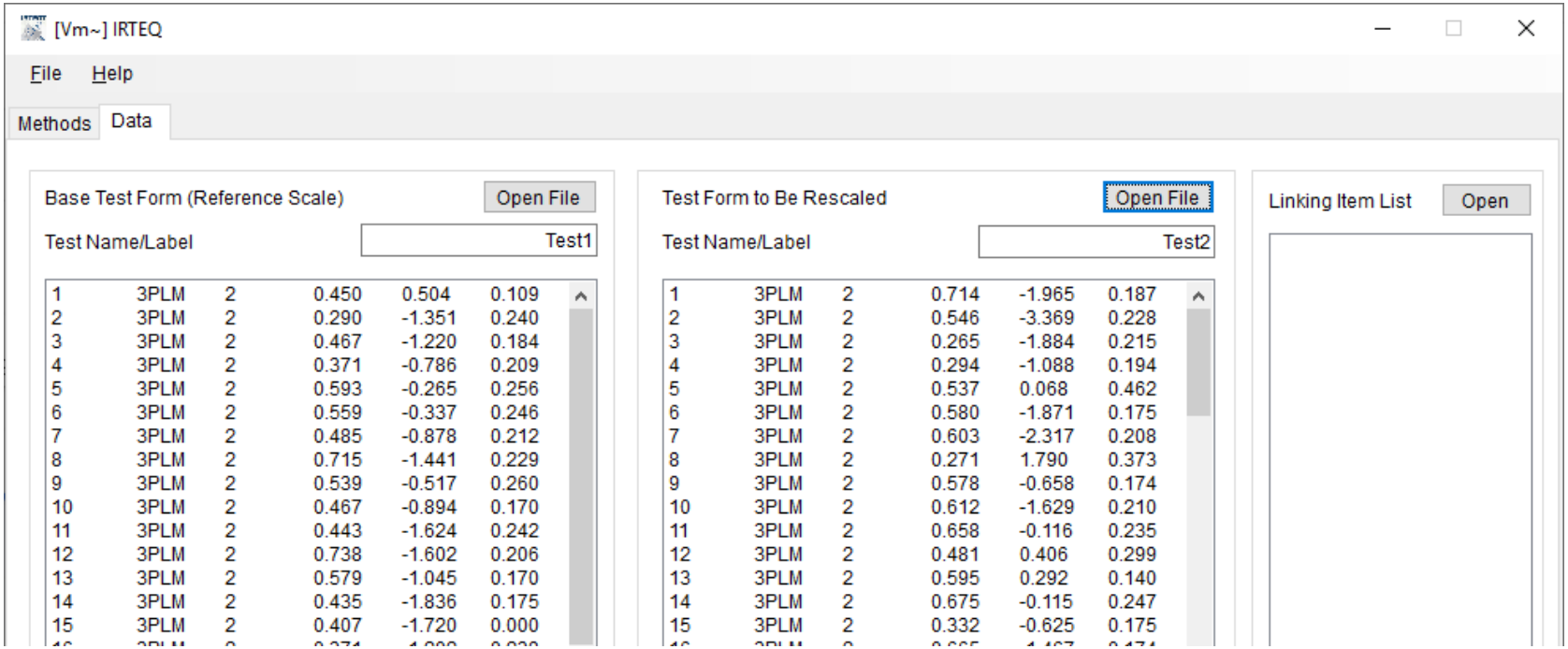
Item Pool Development

- IRTeq comes with several example files:
  - examinee_700.wge (thetas for 700 examinees using a WinGen format)
  - example1.LIL (Linking Item Link text file showing which items in the two test forms are anchor items)
  - example1.syn (syntax file for the example)
  - test1.PAR (test 1 item parameters and standard errors in PARSCALE format)
  - test2.PAR (test 2 item parameters and standard errors)

- Depending on which IRT software you are using, you may need to manually convert your theta and item parameter output to match the ones that IRTeq accepts (i.e., WinGen and PARSCALE).

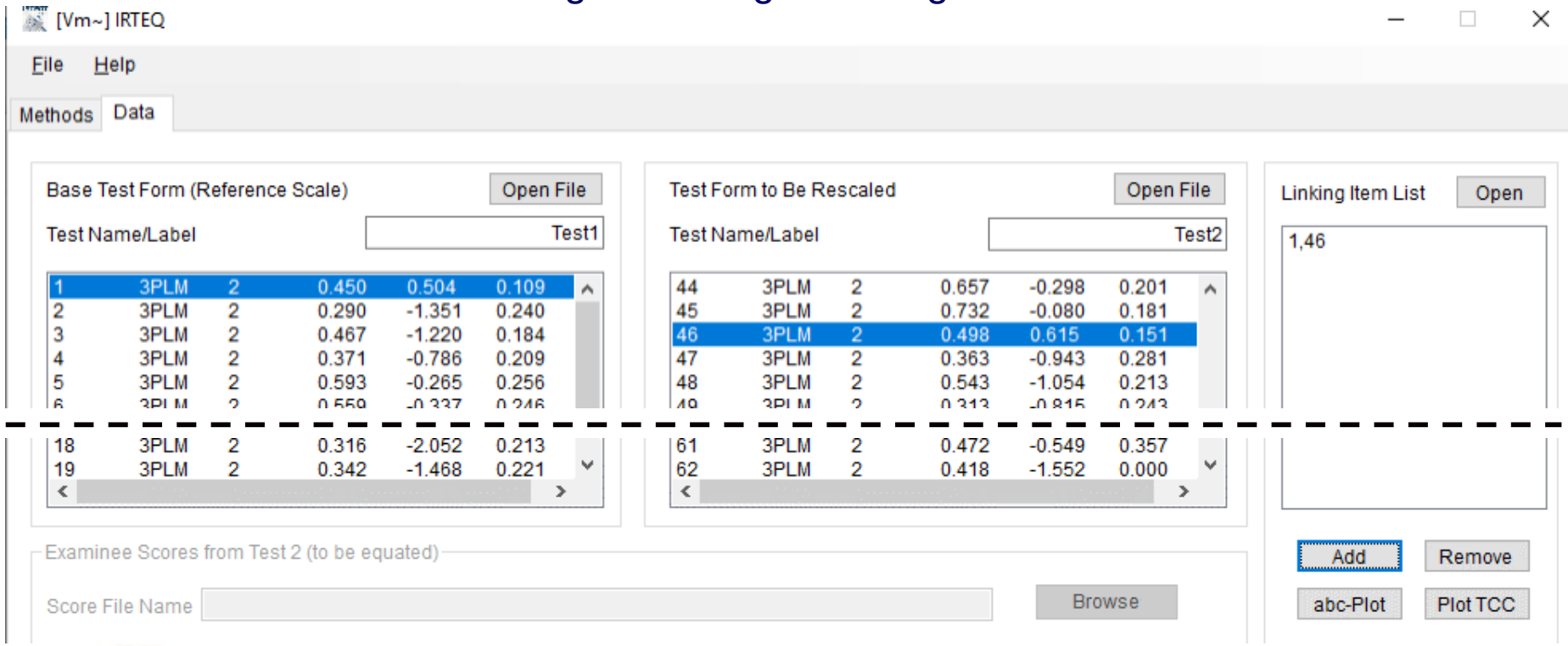
Item Pool Development

- Let's begin by opening the data files for the two test administrations. Click on Open File for the Base Test Form and select test1.PAR and then click on Open File for Test Form to be Rescaled and select test2.PAR.



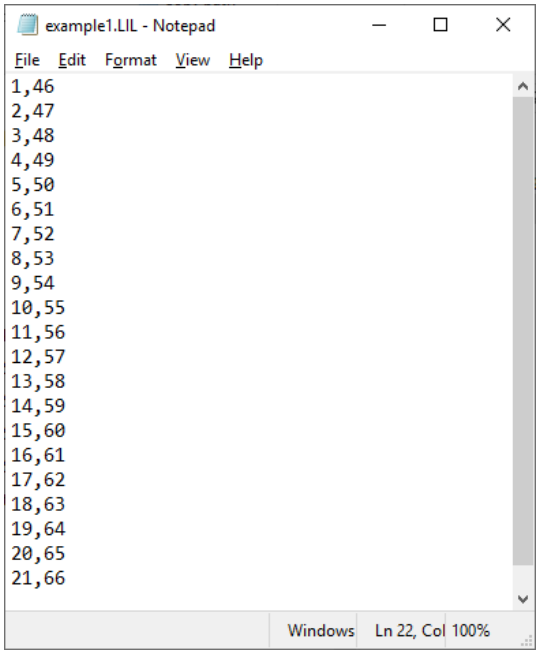
Item Pool Development

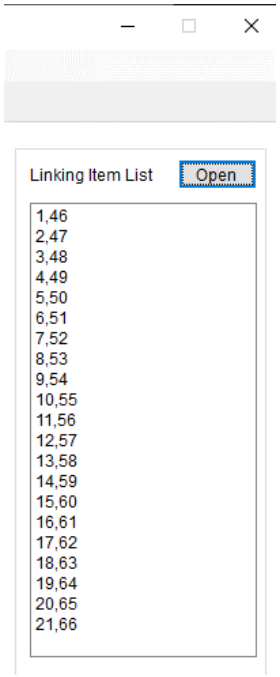
- Next, we need to tell IRTeq which pairs of items are the anchors in both forms. For example, suppose item 1 in test 1 is the same as item 46 in test 2. You can click on the parameters for both items and then click add to begin creating a Linking Item List.



Item Pool Development

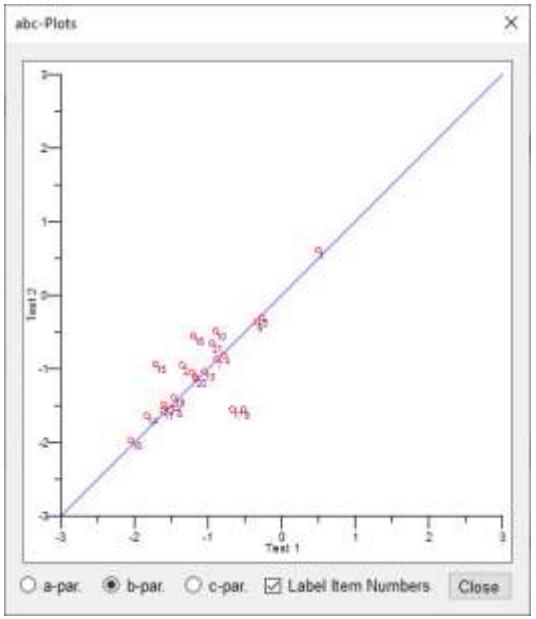
- Or, you can create a text file that shows the item pairings using commas and import it into IRTeq using the Open button in the Linking Item List. The provided example1.LIL is an example of how to set this up. Make sure you save the file with a .LIL extension.

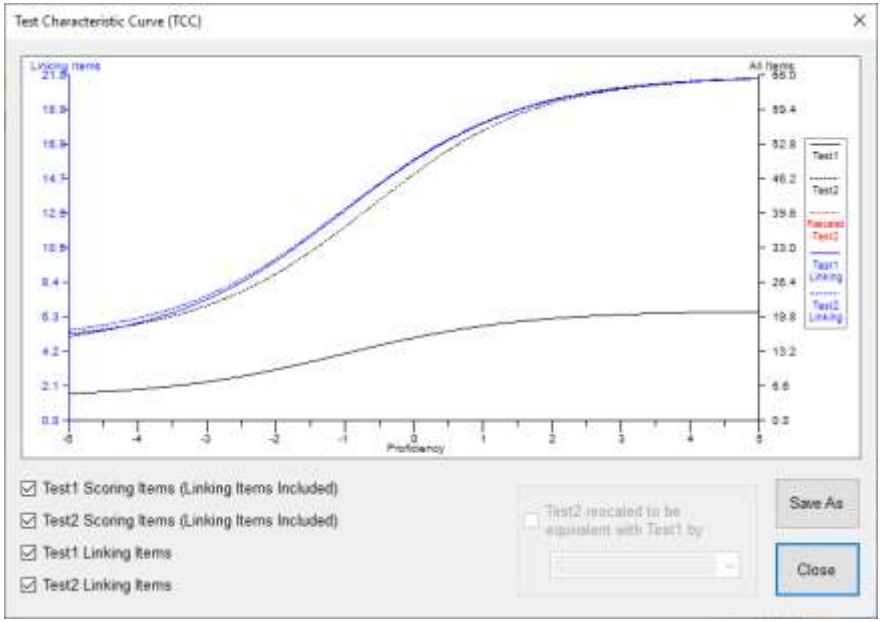




Item Pool Development

- (Side note: At this point, you can also click the abc-Plot and Plot TCC buttons in the Linking Item List area to explore plots of the item parameters across series for the anchor items).

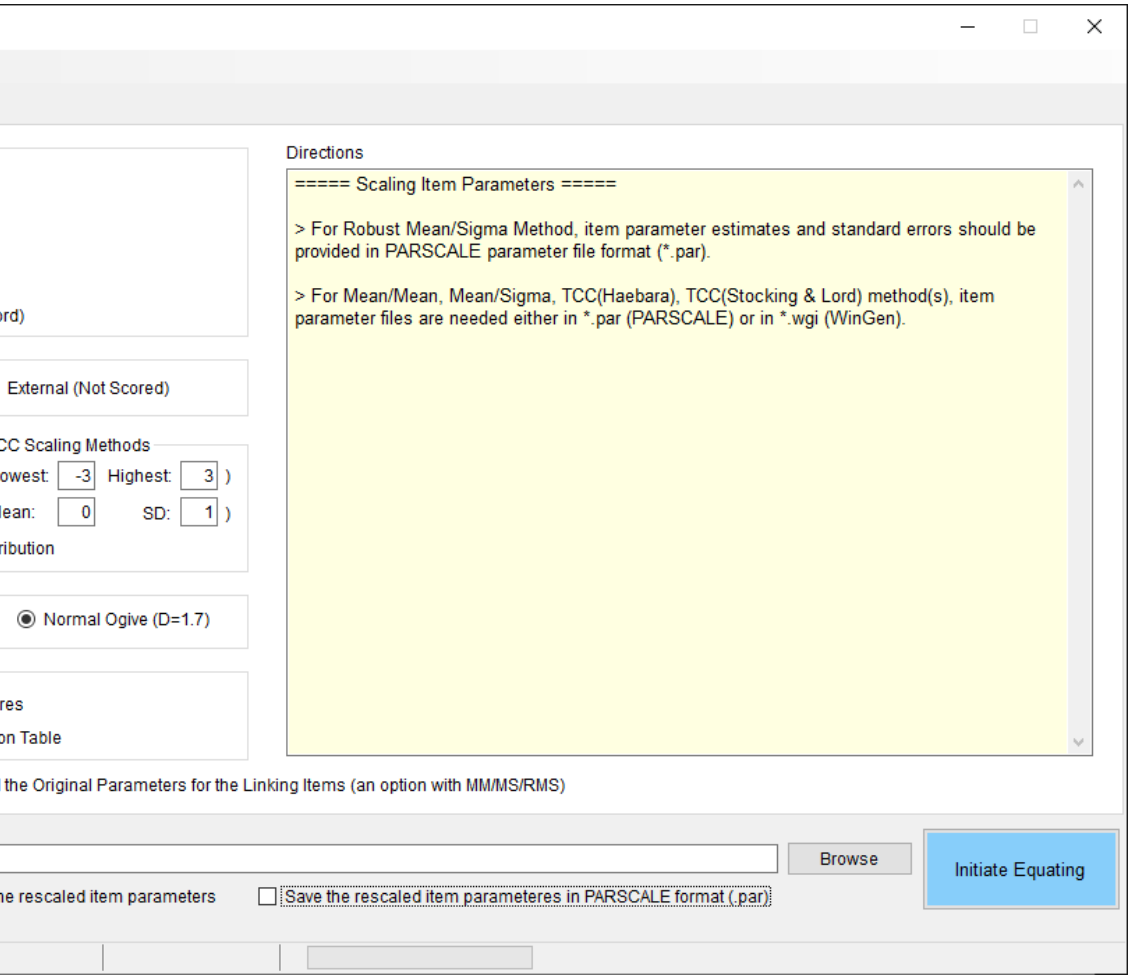




Item Pool Development

- Now, go back to the Methods tab and select the IRT equating methods you want to use under the Item Scaling Method section.
- For CAT development purposes, you can ignore the options for external linking items and true score equating.
- You can probably also ignore the option for averaging the parameters fro the linking items.
  - Sometimes psychometrician equating only two forms will average the item parameters for the anchor items across the two calibrations. Since our example has 20 different calibrations (and IRTeq only handles 2 at a time), this approach won't work for us.
- Make sure that you use the same scale (D = 1 or 1.7) that your IRT software used in your initial calibrations.

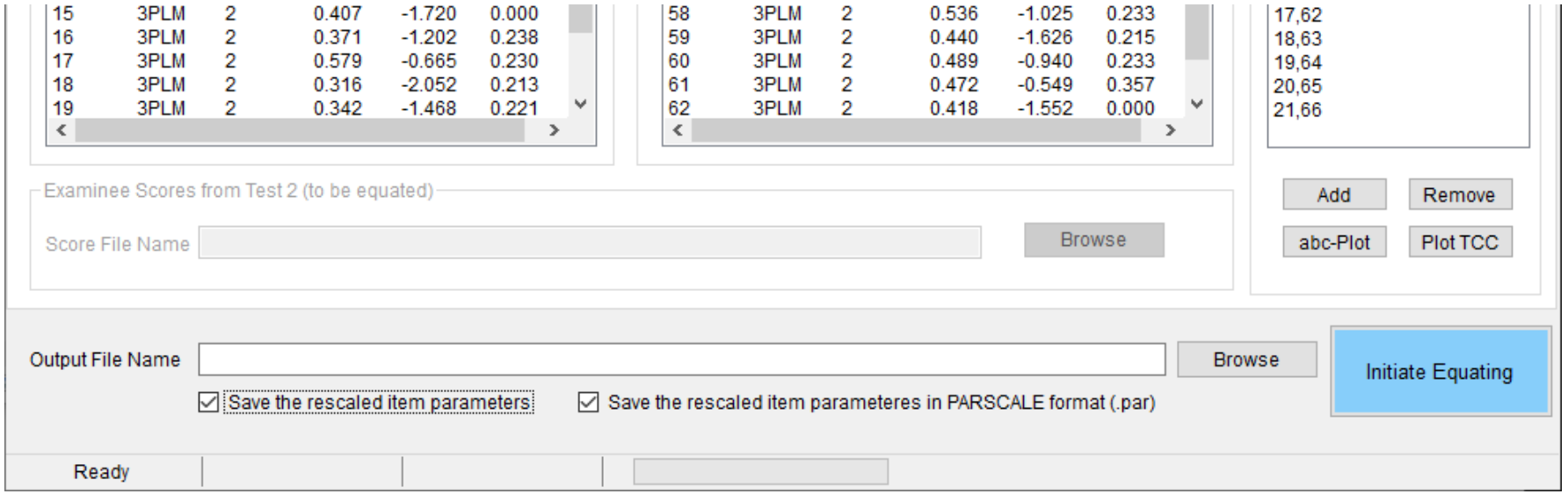
Item Pool Development



U.S. Customs and Border Protection

Item Pool Development

- Now we need to specify the name of an output file using the bottom left section of either the Methods or Data tab. You can also check the boxes for saving the rescaled item parameters.
- Then quick Initiate Equating on either tab to run the analysis.



Item Pool Development

- An output file will be generated and should open automatically.
- It provides some summary info for the number of items, which you should verify.
- The output also gives the mean and SDs for the item parameters across the two calibrations and the correlations between these.

example1.OUT - Notepad

File Edit Format View Help

```
IRTEQ (Han, 2007)
output file (5/12/2020 4:17:23 PM)

================================================================================

Test 1: C:\temp\IRTEQ\EXAMPLES\test1.PAR
   # of items: 21

Test 2: C:\temp\IRTEQ\EXAMPLES\test2.PAR
   # of items: 66

# of pairs of linking items: 21

================================================================================

mean discrimination of linking items in Test1:          0.477
mean difficulty of linking items in Test1:              -1.072
standard deviation of difficulty parameters of linking items in Test1: 0.603

mean discrimination of linking items in Test2:          0.490
mean difficulty of linking items in Test2:              -1.007
standard deviation of difficulty parameters of linking items in Test2: 0.596

Correlation coeff. for discrimination of linking items:  0.824
Correlation coeff. for difficulty of linking items:      0.767
Correlation coeff. for guessing of linking items:        -0.056

================================================================================
```

Windows (CRLF) Ln 1, Col 1 100%

Item Pool Development

- Once you scroll down, you'll see the α and β coefficients (labeled as A and B) for each of the methods.
- The remainder of the output tells you where the rescaled parameter values are saved.

example1.OUT - Notepad

File Edit Format View Help

```
========================================================================

Equating coefficient A with Mean-Mean method:              1.029
Equating coefficient B with Mean-Mean method:              -0.036

Equating coefficient A with Mean-Sigma method:             1.013
Equating coefficient B with Mean-Sigma method:             -0.052

Equating coefficient A with Robust Mean-Sigma method:      0.984
Equating coefficient B with Robust Mean-Sigma method:      -0.130

Equating coefficient A with TCC (Haebara) method:                1.00
Equating coefficient B with TCC (Haebara) method:                0.03
Minimized Loss Function Value with TCC (Haebara) method:         0.00765

Equating coefficient A with TCC (Stocking & Lord) method:        0.99
Equating coefficient B with TCC (Stocking & Lord) method:        0.03
Minimized Loss Function value with TCC (Stocking & Lord) method: 0.00127

========================================================================

Test2 items rescaled onto the scale of Test1 using Mean/Mean Method were saved in
          example1.MM.wgi (5/12/2020 4:17:37 PM)
Test2 items rescaled onto the scale of Test1 using Mean/Signma Method were saved
in          example1.MS.wgi (5/12/2020 4:17:37 PM)
Test2 items rescaled onto the scale of Test1 using Robust Mean/Sigma Method were
saved in    example1.RMS.wgi (5/12/2020 4:17:37 PM)
Test2 items rescaled onto the scale of Test1 using TCC(Haebara) Method were saved
```

Windows (CRLF) Ln 1, Col 1 100%

U.S. Customs and Border Protection

Item Pool Development

- In our example with 20 calibrations, you'll repeat this process 18 more times (1 for each of the remaining calibrations) and compile the α and β coefficients. Next, you'll plug these into the equations shown earlier to place all of the item parameters on the same scale as the base series.
- IRTeq does have a syntax file option whereby you can use code (rather than pointing and clicking) to set up the analyses. Here is the syntax for the example:
- You can also create a cue file that lists different syntax files and run a large number of syntax files automatically at once

example1.syn - Notepad

File Edit Format View Help

```
IRTEQ syntax file saved by a user
12/7/2007 12:55:03 PM
MM,MS,SL,SAVE,PARSCALE
-3,3,NORM,0,1,AVERAGE
LOGISTIC
INTERNAL
EQUATE,CONVERSION
C:\IRTEQ\EXAMPLES\example1.out
C:\IRTEQ\EXAMPLES\test1.PAR
C:\IRTEQ\EXAMPLES\test2.PAR
C:\IRTEQ\EXAMPLES\example1.lil
C:\IRTEQ\EXAMPLES\examinee_700.wge
```

Windows Ln 13, Co 100%

Monte Carlo Simulations

- At this point, you should have a listing of all of the item parameters for your item pool on the same metric.
- If you have collected enough data for multiple CATs, you'll want to divide the items into different pools, trying to ensure that the test information functions follow the distributions of test scores (or are focused on the cut score) as much as possible.
- We will assume for now that you only have enough data for one CAT at this point.

Monte Carlo Simulations

- We need to decide what type of algorithm to use. The choice is often guided by the characteristics of the item pool, your professional judgment, and the philosophy of the testing program.
  - At this point the literature doesn't explicitly recommend one algorithm over another.
    - The answer often depends on the characteristics of the item pool.
    - There are also too many possible algorithms to study all of them across different types of item pools.
    - So, you need to do a Monte Carlo simulation to select an algorithm (or confirm/compare ones you have in mind)

Monte Carlo Simulations

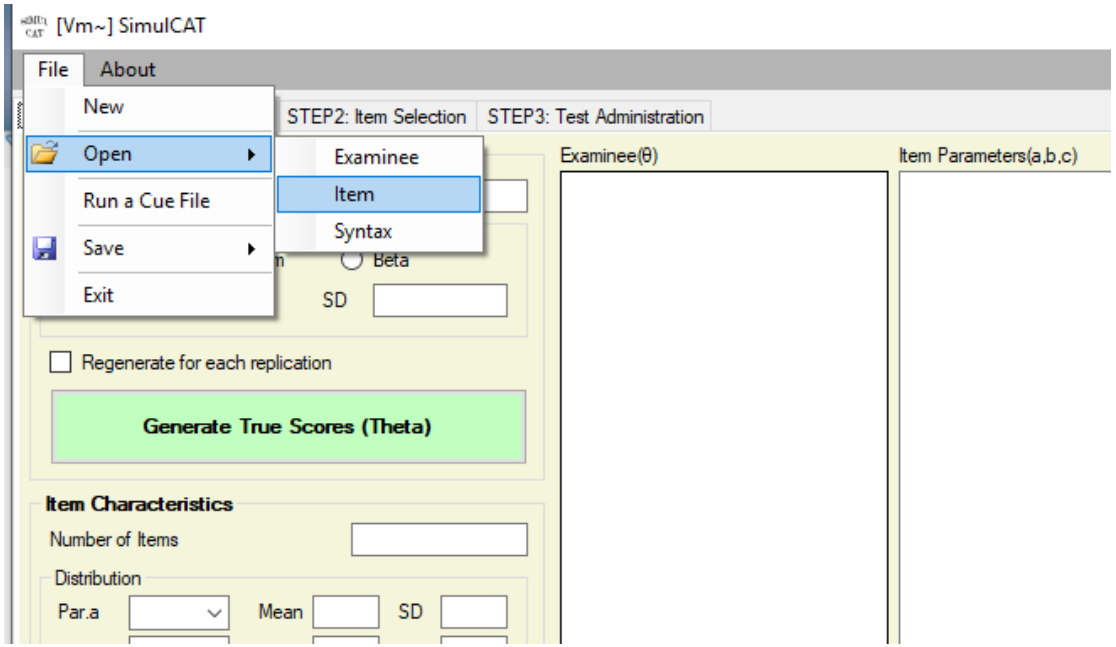
- A number of programs exist for conducting the Monte Carlo simulation:
  - CATSim (Weiss & Guyer, 2012): Available for purchase at https://assess.com/catsim/.
  - R packages: catIrt (Nydick, 2014); xxIRT (Luo, 2019)
  - We prefer SimulCAT (Han, 2012), a free windows-based program that has a GUI and syntax option:

https://www.umass.edu/remp/software/simcata/simulcat/

Monte Carlo Simulations

- This program also comes with example files. We'll only be using one today: Example_ItemPool500.wgix (this contains item parameters for 500 items using a 3-PLM).

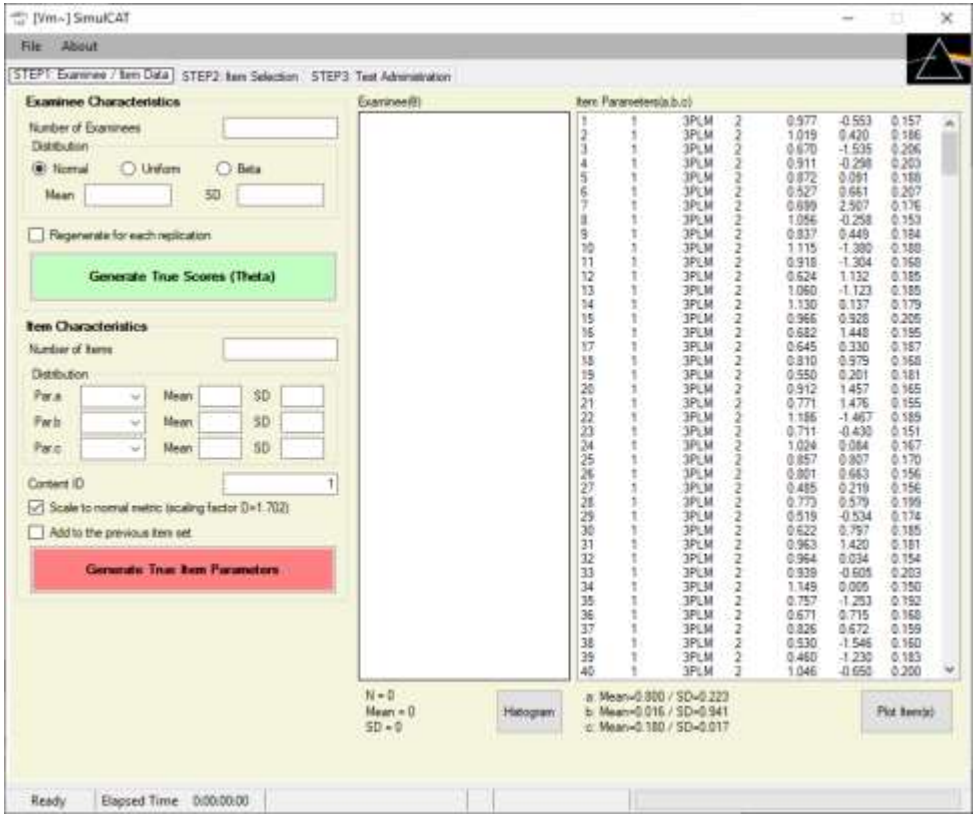
- Let's open SimulCAT and load this file:



Monte Carlo Simulations

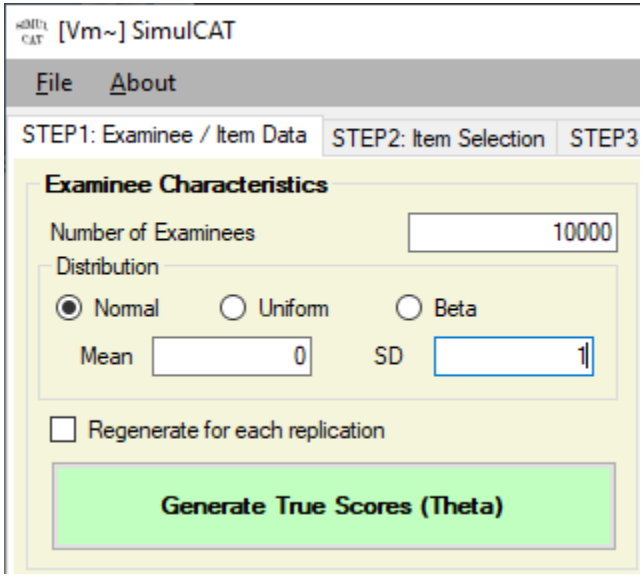
- All of the item parameters will appear on the right side of the window:

- You can click plot items and view the ICCs, item information, TCC, and test information plots



Monte Carlo Simulations

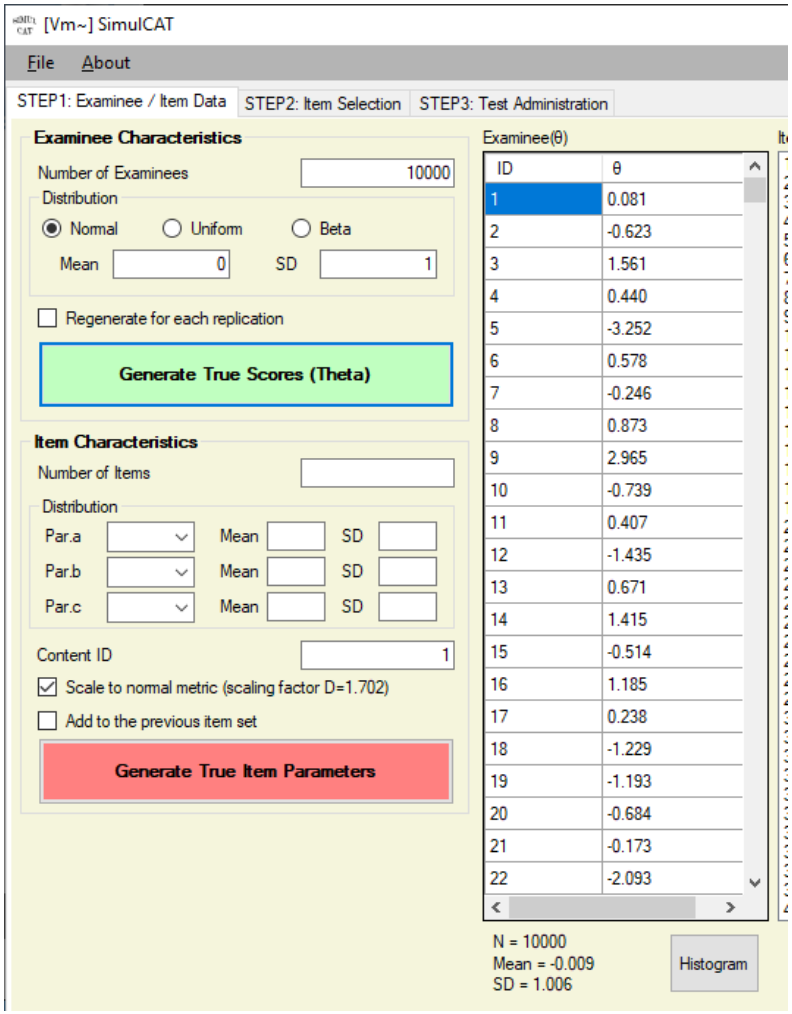
- Next, you should specify the number of simulees/examinees (10,000 is usually a good number for a Monte Carlo simulation) and the mean and SD of the theta distribution (0,1).



- If you have a special theta distribution, you could generate it using another program and import the thetas.

Monte Carlo Simulations

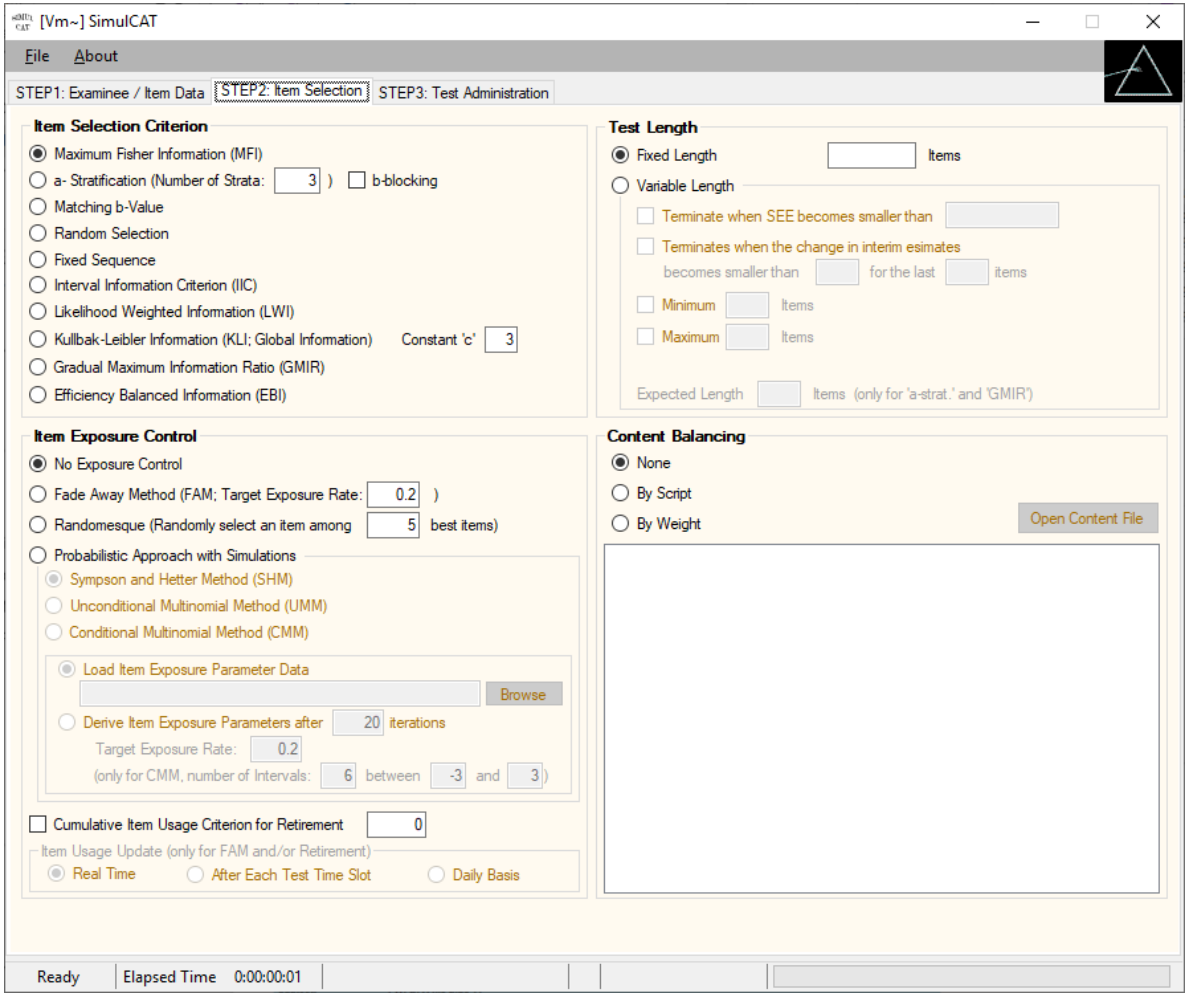
- Click generate true scores (Theta) and the theta values will be randomly generated and will appear in the Examinee(θ) box.
- You can click the histogram button to view the distribution and various descriptive statistics.
- You also want to confirm the D value against your IRT program



U.S. Customs and Border Protection

Monte Carlo Simulations

- Next, move to the STEP2: Item Selection tab.
- Here you will see options for many different types of algorithms.
- In the following slides, we will give a high-level overview of each.



U.S. Customs and Border Protection

Monte Carlo Simulations

- **Item Selection Methods**
- Maximum Fisher Information (MFI): Selects the item that would provide the maximum information value for the examinee's current theta estimate.
- a-stratification: The items are divided into different strata based on their a parameters. At the beginning of the CAT, the least discriminating stratum is used. The item with the closest b value to the examinee's current theta estimate is chosen. You can specify the number of strata. (Chang & Ying, 1999)
  - b-blocking: In addition to being formed based on a parameters, the strata are also designed so that they have balanced b parameter distributions. (Chang et al., 2001)

Monte Carlo Simulations

- Matching-b: Selects the item with an item difficulty b parameter that is closest to the current theta estimate.
- Randomization: This is a control condition in which item are randomly selected.
- Interval Information Criterion: A variation of the MFI method (Veerkamp & Berger, 1997). This approaches sets up a confidence interval around the theta estimate and takes an averages of the information function across the interval.
- Likelihood Weight Information Criterion: Another variation of the MFI method (Veerkamp & Berger, 1997). This approach takes the likelihood function based on the previously administered items and uses it to weight the information function across the theta scale. The values are then summed.

Monte Carlo Simulations

- Kullback-Leibler (1951) Information (Global Information): (Chang & Ying, 1996). Kullback-Leibler information is a type of information function. The global information method takes a moving average of that function.

- Gradual Maximum Information Ratio: Both MFI and the effective efficiency (i.e., how well the potential information for an item is realized for a given theta estimate) for items are computed. Both are summed together and the sum is used to select items. At the beginning of the CAT, more weight is given toward efficiency and at the end of the CAT more weight is given to MFI. (Han, 2009)

Monte Carlo Simulations

- Efficiency-Balanced Information: Similar to the Gradual Maximum Information Ratio, except that item efficiency and MFI are evaluated across the theta estimate interval (which depends on standard errors; Han, 2010).

- **Item Exposure Methods**

- Randomesque: rather than selecting the single best item, the best *k* items are identified and one of those is randomly selected to administer to the examinee (Kingsbury & Zara, 1989).

Monte Carlo Simulations

- Sympson-Hetter (1985): A target probability that an item is administered is established. A single random number between 0 and 1 is computed. The items are rank-ordered by the item selection method. Starting from the top, the target probability is compared to the random number. If the random number is smaller, the item is given, if not, the next item is considered.

- Unconditional Multimonial Method: A variation of Sympson-Hetter. A multinomial distribution is computed and then compared to the random number. (Stocking & Lewis, 1995)

Monte Carlo Simulations

- Conditional Multimonial Method: A variation of the Unconditional Multimonial Method that applies the item exposure controls to different regions of the theta distribution. This approach would control item exposure for groups of examinees with similar ability. (Stocking & Lewis, 1995)

- Fade-away method: This approach requires continually tracking and updating observed item exposure data. The variable that item selection is based on (e.g., MFI) is weighted by the target exposure rate divided by the actual exposure rate. Items that are frequently used tend to fade-away from being administered.

Monte Carlo Simulations

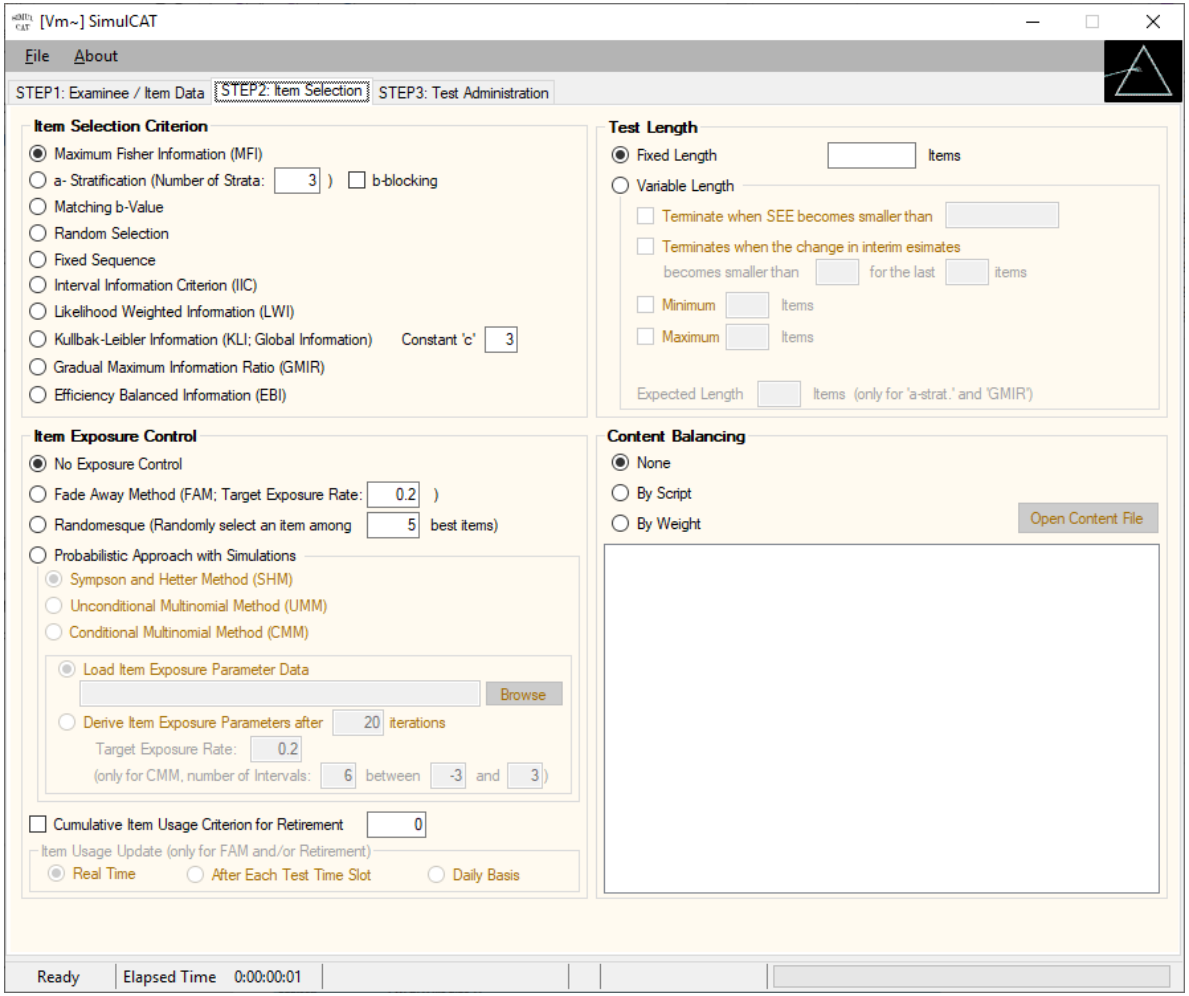
- **Test Length**
- Fixed length: Each examinee receives the same fixed number of items.
- Variable length: Instead of administering the same number of items to each examinee, the actual number administered depends on the stability of the theta estimate.
  - Can stop testing when the standard error of the theta estimate is lower than a user-specified value
  - Can stop when changes in the theta estimates are below a user-specified value.
  - Can also specify a minimum and maximum number of items.

Monte Carlo Simulations

- **Content-Balancing Methods**
- Script method: Several scripts are created representing different content areas.
- By weight/Constrained CAT method: An item is selected from a content area and each content area has a target percentage. The content area whose actual percentage of items administered and target percentage is the most different is selected. (Kingsbury & Zara, 1989)

Monte Carlo Simulations

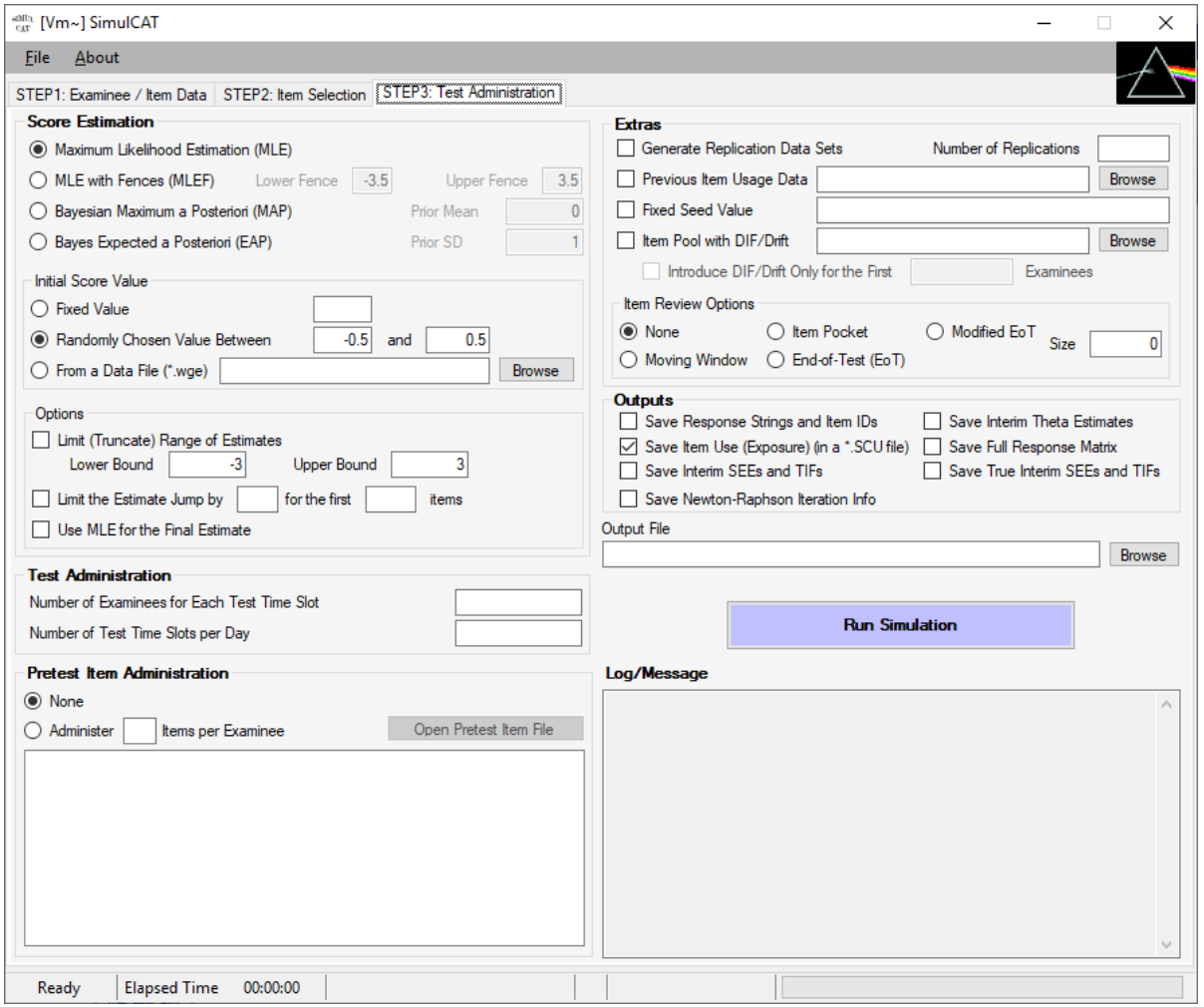
- At this point, determine which methods you want to explore.
- You can only test one method at a time.
- However, you can create syntax and cue files to run multiple methods.



U.S. Customs and Border Protection

Monte Carlo Simulations

- Next move to Step3.
- Here you can indicate test administration and scoring conditions.
- We discuss each on the following slides.



U.S. Customs and Border Protection

Monte Carlo Simulations

- **Score Estimation**
- These are essentially the classic theta estimation approaches: Maximum likelihood estimation (MLE), Bayesian maximum a posteriori (MAP), and Bayes expected a posteriori (EAP).
  - MLE with Fences is MLE but artificial items with fixed responses are added to allow estimating theta for abnormal response patterns and those that consist of all 0s or 1s. The fence values are user-specified b parameters (see Han, 2016, for more information).
- We recommend using whatever approach you used to estimate the item parameters

Monte Carlo Simulations

- **Initial Score Value**
- Fixed value: All examinees start the CAT with the same initial theta (often set to 0).
- Randomly chosen value between a user-specified interval
- From a data file: In this option, you can supply starting values for each examinee. This might be used if you had a previous test, pre-test, or other information for choosing a starting value.

Monte Carlo Simulations

- **Options**
- You can also truncate the range of estimates to a specific interval (e.g., if a theta estimate was -4, you could truncate it to -3), limit how much the theta estimates change and apply that limit to the first k items, or use MLE as the final theta estimate.
- **Test Administration**
  - Some of the item exposure controls depend on item exposure rates that are computed daily; here you can customize how many examinees are in a time slot and how many time slots are in a day.

Monte Carlo Simulations

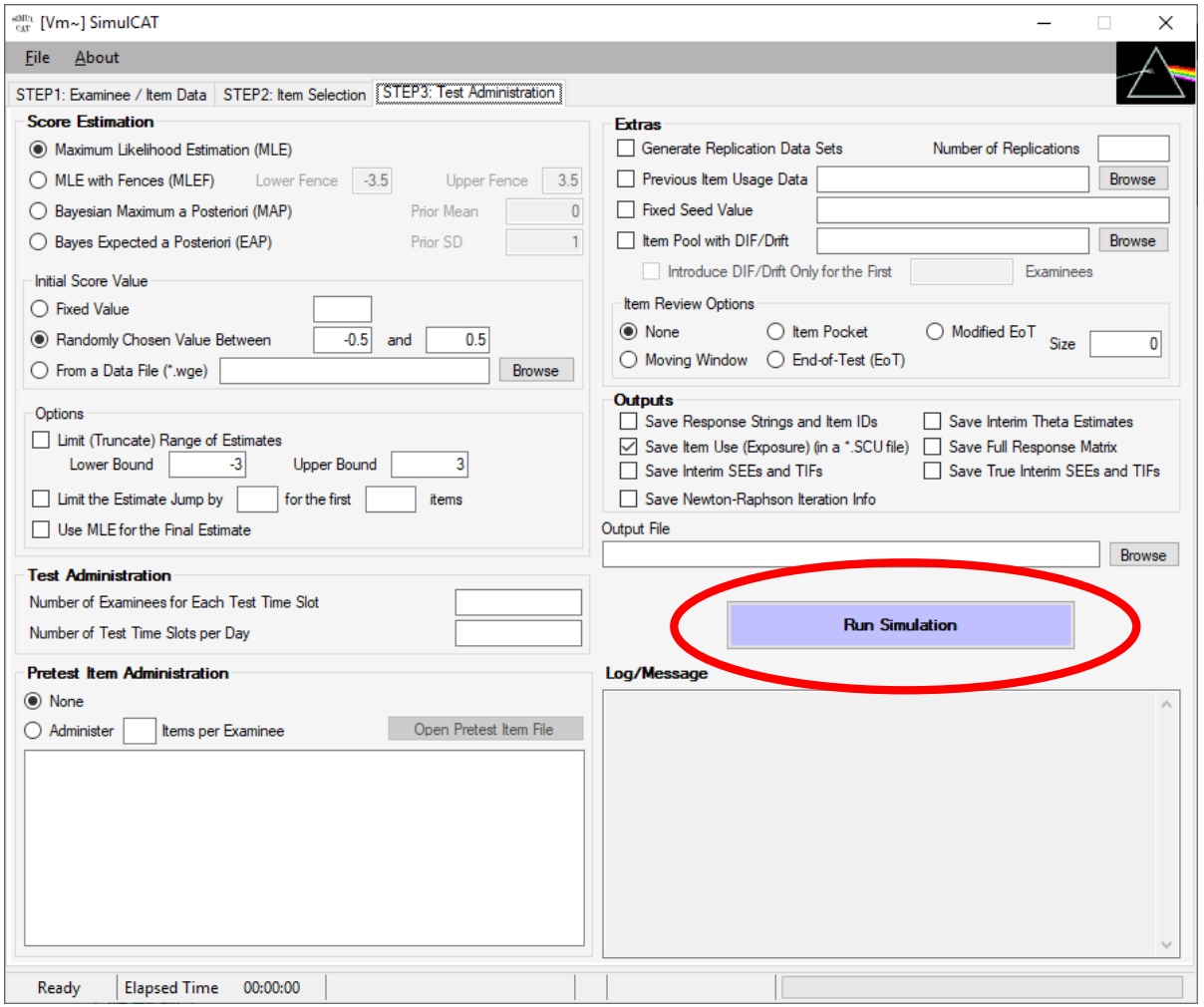
- **Pretest item administration**
- You can simulate administering a pretest to examinees.
- **Extras**
  - It's also possible to conduct multiple replications, study differential item functioning (DIF) and item drift, consider previous item usage, and fix a random number seed value.
    - We recommend fixing the random number seed value for each run so that you could replicate the exact results if you later needed to.

Monte Carlo Simulations

- **Outputs**
- You can save all types of output for each run. We found it useful to save the item use exposure (.SCU) file (so that you can track item exposure without having to do too much data manipulation).
- At a bare minimum you also need the true and final theta estimates.
- You could, of course, check all the boxes and save everything. However, if you are testing many conditions and have a large sample size, file sizes and disk space can quickly become an issue.

Monte Carlo Simulations

- Finally, click run simulation.



Monte Carlo Simulations

- **Designing your Monte Carlo Simulation**
- You will have to think strategically about which conditions to study. The design can get out of hand pretty quickly.
  - For one CAT, we thought about testing almost all of the options but then realized it would have taken 10 years to run the simulation.
  - We recommend thinking carefully about each option and choosing those you are comfortable with considering psychometrics, programming feasibility, and face validity.
    - For example, starting with a randomly-chosen initial theta value might help with item exposure, but some examinees might perceive this as unfair and arbitrary.

Monte Carlo Simulations

- **Analyzing the Results from your Monte Carlo Simulation**
- We recommend computing a number of summary statistics for each condition and then comparing the results.
- Mean Bias: This is the final theta estimate minus the true theta estimate averaged across all examinees. It will tell you if the algorithm, on average, tends to over or underestimate theta.
- Mean |Bias|: It's also helpful to convert the mean bias values to positive values to allow you to sort the results for each condition and find those that are closest to zero.

Monte Carlo Simulations

- **Analyzing the Results from your Monte Carlo Simulation**
- Mean Square Error: (theta final – true theta)^2.
- Root Mean Square Error: Square root of MSE
- Correlation: correlation between final theta estimates and true thetas.
- Reliability: The correlation squared.

Monte Carlo Simulations

- **Analyzing the Results from your Monte Carlo Simulation**
- Chen et al. (2003) average test overlap rate: $\bar{T} = \frac{S^2+\mu^2}{\mu}$
  - Where S^2 is the variance of the item exposure rates (which are the proportion of examinees that received an item), μ is the mean of the item exposure rates.
  - A value of 40% means that, on average, two examinees have 40% of their items in common.
- Number of underused items (i.e., number of items administered to less than .05 [5%] of examinees)
- Percentage of items that are underused

Monte Carlo Simulations

- **Analyzing the Results from your Monte Carlo Simulation**
- Scaled χ^2 (Chang & Ying, 1999): This compares the observed item exposure to an ideal/expected uniform item exposure rate. The ideal/expected rate is the length of the CAT divided by the total number of items in the item pool.
- Essentially, you will compare these statistics across different conditions and decide on a final algorithm.
  - You might also make refinements and run additional conditions.

Scaling/Equating/Metric Issues

- The CAT will generate a final θ estimate for each examinee.
- You could use that the final θ estimate as the operational score.
  - It is on a *z*-scale and has positive, negative, and zero values. This makes it more difficult to explain to non-psychometricians.
  - It can be useful to place the θ estimates onto the same scale as a static test.
  - Oftentimes, you are replacing a static test with a CAT.
  - The static test's metric/scale might have been used in validation and standard-setting studies.

Scaling/Equating/Metric Issues

- Hambleton, Swaminathan, and Rogers (1991, p. 84-7) explain how to place θ on a number correct scale for a static test.
  - You first need to obtain the item parameters for the static test.
  - Next, determine the desired level of precision in the equating (e.g., do you want to equate θ in increments of 0.1, 0.01, etc.).

Scaling/Equating/Metric Issues

- Create a spreadsheet with different θ values in each row and P (probability of correct response) for each item in columns. Use the 1-, 2-, or 3-PLM equation to compute the Ps. Next sum the Ps across the items to give the number of correct items (T, using Hambleton et al.'s notation). You can divide this by the total number of items to give the proportion correct (π), and convert to a percentage if desired.
- On the next slide, we give an example for a fictitious 10-item test with θ in unit increments.

Scaling/Equating/Metric Issues

Theta transformation for CAT Tutorial.xlsx - Excel

File Home Insert Page Layout Formulas Data Review View ACROBAT Tell me what you want to do... CUCINA, JEFFREY Share

B2 =B$12+((1-B$12)*((2.718^(1.7*B$10*($A2-B$11)))/(1+(2.718^(1.7*B$10*($A2-B$11))))))

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Θ	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	T	π
2	-3	0.214	0.231	0.102	0.259	0.209	0.164	0.325	0.173	0.246	0.113	2.036	**20%**
3	-2	0.369	0.235	0.118	0.269	0.288	0.211	0.335	0.234	0.253	0.131	2.445	**24%**
4	-1	0.677	0.263	0.193	0.311	0.528	0.403	0.389	0.392	0.315	0.211	3.683	**37%**
5	0	0.903	0.431	0.454	0.450	0.826	0.750	0.577	0.646	0.614	0.463	6.113	**61%**
6	1	0.978	0.800	0.805	0.710	0.958	0.942	0.842	0.856	0.923	0.795	8.610	**86%**
7	2	0.995	0.968	0.957	0.906	0.991	0.989	0.964	0.953	0.991	0.951	9.666	**97%**
8	3	0.999	0.996	0.992	0.976	0.998	0.998	0.993	0.986	0.999	0.990	9.928	**99%**
9	Item Parameters												
10	a	0.923	1.234	1.01	0.875	0.955	1.022	0.998	0.735	1.31	0.958		
11	b	-1.3	0.5	0.25	0.7	-0.8	-0.5	0.3	-0.279	0.021	0.257		
12	c	0.159	0.231	0.099	0.256	0.187	0.153	0.322	0.145	0.245	0.109		
13													

Scaling/Equating/Metric Issues

Theta transformation for CAT Tutorial.xlsx - Excel

File Home Insert Page Layout Formulas Data Review View ACROBAT Tell me what you want to do... CUCINA, JEFFREY Share

B2 =B$12+((1-B$12)*((2.718^(1.7*B$10*($A2-B$11)))/(1+(2.718^(1.7*B$10*($A2-B$11))))))

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Θ	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	T	π
2	-3	0.214	0.231	0.102				0.325	0.173	0.246	0.113	2.036	20%
3	-2	0.369						0.335	0.234	0.253	0.131	2.445	24%
4	-1	0.677	0.263					0.389	0.392	0.315		3.683	37%
5	0	0.903	0.431					0.577	0.646	0.614	463	6.113	61%
6	1	0.978	0.800	0.80								8.610	86%
7	2	0.995	0.968	0.957								9.66	97%
8	3	0.999	0.996	0.992								9.9	99%
9	Item Parameters												
10	a	0.923	1.234	1.01									
11	b	-1.3	0.5	0.25									
12	c	0.159	0.231	0.099									
13													

The values in cells B2:K8 are the Probability of correct responses using the 3-PLM equation, which is typed in the bar above.

The values in cells L2:L8 are the Sum of the values in columns B through K

This is the number correct converted to a percentage.

Scaling/Equating/Metric Issues

- A few notes:
  - The number correct scores and θs are monotonically related (not necessarily linear).
  - You can apply this procedure to compute number correct scores for item that the examine didn't receive, provided that you have the θs and item parameters (and both are on the same metric).
  - This is actually a transformation of true θ scores to true number correct scores.
  - This procedure is a derived from the summation of item characteristic curves to yield a test characteristic curve. However, the curves are usually for the static test items, not the entire CAT item pool; therefore, the curves we discussed earlier and that are produced in SimulCAT (for example) are not relevant here.

Experimental Item Collection

- After CAT is implemented, you may desire to continue collecting data on experimental items
  - This will help you to develop additional CATs
  - Might need to add new items if content domain changes over time (e.g., some aspects of job knowledge might change over time).
  - You might also need to recalibrate existing items to control for item drift.
    - Suppose you developed 5 CATs each with their own pool. At some point in the future, you might consider stopping use of one of those CATs and readministering the items as experimental items in order to recalibrate the item parameters.

Experimental Item Collection

- There are some strategies for collecting experimental items
  - Administered a static test in addition to the CAT
    - Need to consider order effects and fatigue
    - Examinees likely need to be told that one of the tests contains unscored experimental items.
      - They might be able to identify which test is the non-CAT static experimental one and then reduce their motivation when completing items.

Experimental Item Collection

- There are some strategies for collecting experimental items
  - Embed experimental items within the CAT
    - Experimental items could be in fixed or random locations
    - We suggest avoiding placing experimental items at the very beginning or end of the CAT
      - Examinees often know that the first item should be of average ability and the last should be near their own ability levels. They might panic if the first item is very hard or if the last item is very easy.
      - Examinees also might be gauging their performance throughout the CAT and could possibly identify or be thrown off by experimental items that are far from interim theta.

Experimental Item Collection

- There are some strategies for collecting experimental items
  - Collect experimental items outside of CAT environment (e.g., using a static test that is part of the testing program).
- In most cases, you will have a set of unscored static test items, CAT item responses, and thetas.
- In the literature, the preferred method seems to be to have enough usable (i.e., those that will survive item analysis) static test items to conduct a calibration and obtain item parameters and thetas. (This can range from 10 to 60 items depending on which “rules of thumb” you choose.)

Experimental Item Collection

- The static items are calibrated separately from the CAT responses and could be on a different scale.
- Pommerich & Segall (2003) developed a transformation which works well according to simulation studies for placing static test (Test 2) parameters on same scale as CAT (Test 1); note that only a and b, but not c, need transformation:

$a_{Transformed,\ Test\ 2} = a_{Test\ 2}/A$

$b_{Transformed,\ Test\ 2} = A(b_{Test\ 2})+B$

Where:

$A = \sigma_{Test1}/\sigma_{Test2}$

$B=\mu_{Test1}-A(\mu_{Test2})$

Experimental Item Collection

- Other approaches include
  - Attempting to calibrate CAT responses with static items.
    - This yields a sparse data matrix with a lot of missing data in the CAT responses.
  - Using BILOG-MGs external θ command (which allows you to place new item parameters on same scale as an existing set of θs)
    - However, we could not locate any technical details on how this is done in BILOG-MG or how well it works

Experimental Item Collection

- Other approaches include
  - Treating CAT θ estimates as known and estimating item parameters when θ is known.
    - However, θs are actually estimates with error and not known and most IRT software packages don't implement this option.
  - Including a set of static anchor items with know item parameters and equating the experimental items to the anchor items using IRTEQ (for example).

Creating Instructions for Developers

- Typically, online testing systems are developed and maintained by computer programmers who are not experts in psychometrics.
- Because of this it is recommended to create a set of instructions detailing how to program the algorithm and test equating.
- Recommended information to include in the instructions:
  - Item selection method (e.g., randomesque)
  - Test stopping rules (e.g., # of items)
  - Equations for estimating theta

Creating Instructions for Developers

- Recommended information to include in the instructions (continued):
  - Number of quadrature points (e.g., 20, 40)
  - Instructions for conducting random selections of items
  - Equating instructions to other test forms
  - Location of experimental items (for pretesting other items)
  - Output file formats

Creating Instructions for Developers

- A critical step in developing the instructions is to think strategically about which variables should be computed and recorded when the CAT is administered.
- This is valuable in case the CAT is being used in a high-stakes situation where scores may be challenged.
- It is recommended to audit the programmed CAT algorithm to ensure no mistakes were made in the programming and the ability to do this depends largely on the variables that are recorded and outputted.

CAT User Testing

- After a CAT algorithm is chosen and implemented, it is recommended that careful user testing is conducted to ensure that the CAT algorithm is functioning properly.
- User testing ensures no errors were made during the programming process. Ensuring the CAT algorithm is error-free provides protection for organizations and appropriate legal defensibility.
- We recommend creating a very large (e.g., 500,000) number of test cases and running them through the CAT algorithm using a Monte Carlo simulation to ensure that all possible item iterations are verified.

CAT User Testing

- Several aspects of CAT output that we have identified as crucial to user testing include:
  - Interim theta estimates (i.e., the ability estimates between item administrations).
  - Item selection correctness (i.e., whether the correct items are being selected based on the interim thetas).
  - Final thetas (i.e., the final ability estimates of the test).
- Interim theta estimates and item selection correctness are important because theoretically a test taker could have a correct final ability estimate, but an incorrect item iteration or interim theta estimates.

CAT User Testing

- Final thetas should be examined for correctness to the last decimal place and rounding errors, as these issues may be problematic around where a test cut score is set.
- Additionally, there may be legal or other challenges that require verifying or proving that the test was administered correctly, and these aspects of the test may need to be verified or produced.

Implementing CATs

- When implementing a CAT, it is important to update any applicable examinee communications
  - Provide a description of the new CAT
  - Give insight on what examinees can expect and how they should prepare for taking a CAT
  - Update relevant website(s)
    - Provide updated information on CAT
    - Indicate when the CAT will be implemented
- The CAT developer may also need to work with stakeholders (e.g., HR staff, recruiters) so that they are aware of the new test and implementation guidelines

Implementing CATs

- CATs can be either developed from scratch or from an existing static test
  - If a psychometrician is converting a static test to a CAT, then he or she will need to decide if it is necessary to discontinue use of existing static test scores upon implementation of CAT
    - A decision will need to be made whether applicants or candidates with passing scores on the static test need to take the adaptive version or if they will be allowed to keep their passing score.
  - This will not be a concern if the CAT is being developed as a brand new test

Implementing CATs

- After the CAT is implemented, it is important to confirm the algorithm was implemented correctly
  - Checking live data is the only way to know if the CAT is functioning as intended
    - Is the CAT selecting the correct items for test takers?
    - Are incorrect and correct response options being scored correctly?
    - Are the stopping rules working correctly?
    - Are calculations for interim and final thetas correct?
    - Do final thetas match any equated scores?
    - Is the scoring for the overall test correct?

Implementing CATs

- To minimize any potential issues with the algorithm that could arise after the implementation of the CAT, it is recommended that a plan be created for checking the data prior to implementation
  - Indicate what part(s) of the algorithm and scoring you need to check
  - Provide the steps for how those checks will be carried out

Lessons Learned/Things to Consider

- CAT provides a number of important advantages over traditional testing
- CAT can provide more accuracy because each examinee is given a unique test that is tailored to his or her ability level
  - Questions that do not provide sufficient information about the examinee's proficiency are avoided
    - For example, really easy questions are typically not asked to individuals with a very high ability level and vice versa
  - This can provide a higher level of precision across a wider range of ability levels

Lessons Learned/Things to Consider

- In comparison to static tests, CATs can reduce testing length by more than 50%, while maintaining a comparable level of reliability
  - Fewer items are needed to achieve acceptable accuracy
- CAT testing has also been shown to improve test security
  - Since each test is unique to the examinee, this makes it more difficult for an individual to capture the entire pool of items
    - This drastically reduces the likelihood of widespread cheating and the subsequent need to redo an entire test, which would cost a significant amount of time and money

Lessons Learned/Things to Consider

- There are some notable limitations to using the CAT method
- Unlike traditional testing, CAT is difficult to develop in that it requires the expertise of psychometricians to calibrate items using an algorithm based on an IRT model that accurately measures the ability level of examinees
  - Item calibration requires that extensive data be collected on a large item pool
    - Developing a sufficiently large item pool can take a lot of time and effort and can end up using a lot more resources than traditional testing

Lessons Learned/Things to Consider

- Some subjects and skills cannot be accurately measured with CAT because IRT cannot be readily applied to those areas
- Using IRT models can also cause item constraints that result in an overly narrow selection of questions being presented to examinees
  - This is likely to occur if the content isn't balanced across different item difficulty levels.
  - These constraints can result in examinees completing sets of items that are broadly the same, thus losing the advantage over traditional tests
    - This presents a potential issue when CAT is used for knowledge and content-focused tests there can be blatant inequities when comparing the scores of examinees
    - Content balancing methods are a possible solution

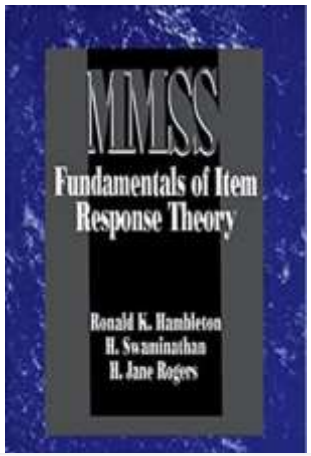
Lessons Learned/Things to Consider

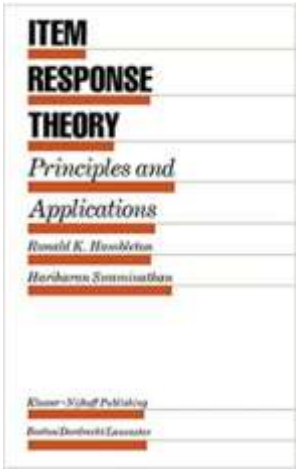
- Some of the limitations of CAT can be addressed using another method of testing called linear-on-the-fly testing (LOFT)
- LOFT takes items from a very large item pool and constructs a unique exam for each examinee
  - This is done through a program that pseudo-randomly selects items so that examinees receive tests that are equivalent with respect to content and statistical characteristics, but with different items
    - Like CAT, this selection method typically uses IRT, but results in longer testing times and less precision than CAT

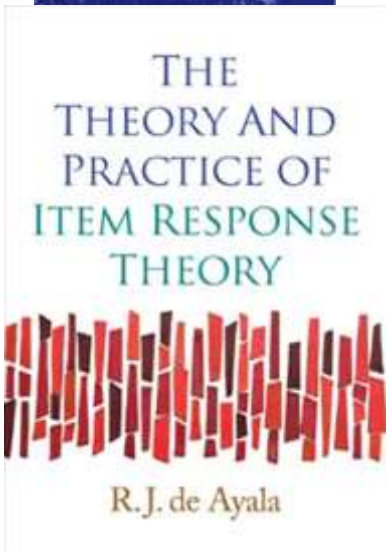
Lessons Learned/Things to Consider

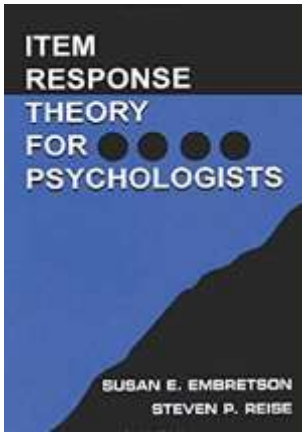
- Ultimately, when it comes to different modes of testing (e.g., static vs. CAT vs. LOFT), there will always be some trade-offs between the different methods
  - No single method will address every limitation
- When deciding whether to use methods such as CAT, it is important for test developers to weigh the advantages and disadvantages of each method and decide which approach works better for their situation

Resources/Recommended Reading









- Web-based item characteristic curve plotting
  - https://demonstrations.wolfram.com/ItemCharacteristicCurves/
- IRTEQ, SimulCAT, WinGen, and other free programs: http://www.hantest.net/
- Hambleton, R.K., Swaminathan, H, & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE Publications.
- de Ayala, R.J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.
- Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. New York: Springer Science & Business Media.
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.*

U.S. Customs and Border Protection

Resources/Recommended Reading

- Weiss, D.J. (1983). *New Horizons in Testing: Latent trait test theory and computer adaptive testing.* New York, NY: Academic Press.
- Van der Linden, W.J., & Glas, C.A.W. (2000). *Computer adaptive testing: Theory and practice.* Dordrecht, The Netherlands: Kluwer Academic Publishers, Inc.
- Wainer, H. (2014). *Computer adaptive testing: A primer.* (2nd ed.). New York: Routledge.

U.S. Customs and Border Protection