GEN3PL_RawDATA: 3PL Item Model-Based Data Generation For Raw Responses

Richard M .Luecht, PhD
University of North Carolina at Greensboro

Version 2.0 (April, 2011)

GEN3PL_RawDATA_V2 generates selected- response (SR) or multiple-choice (MC) raw response data with the following features: (a) complete control over every item characteristic function, based on the IRT 3PL model; (b) user-defined single or multiple answer keys for every item; (c) proportional omit coding at the item level; (d) proportional not-reached coding at the item level; and (e) user-specified moments of the underlying proficiency distribution.

The program uses an underlying item response theory (IRT) model for dichotomous (0,1) response data to generate the raw responses. The IRT 3PL normal ogive model is used:

$$P_i(\theta) = c_i + \frac{1-c_i}{1+\exp\left[-Da_i(\theta-b_i)\right]} \qquad (1)$$

where θ is the latent ability, a_i is the item discrimination parameter (slope), b_i is the item difficulty parameter (location or threshold), and c_i is the item pseudo-guessing parameter (lower asymptote). An item control file is required. This file, described further on, specifies the item parameter values, a_i, b_i, and c_i for i=1,…,n items. The maximum test length is n=1,000 items.

The response generating algorithm is straight-forward. The a_i, b_i, and c_i parameters are read from a user-supplied file for i=1,…,n items. For j=1,…,N examinee abilities, θ_j is drawn from a normal distribution with user defined population moments, μ (mean) and σ (standard deviation). Equation 1 is used to computed the true response function, $P_\mathrm{i}(\theta_\mathrm{j})$ for each item. A uniform random number $0 \le \pi_\mathrm{ij} \le 1$ is generated and the simulated score item response is computed as $u_\mathrm{ij} = 1$ if $P_\mathrm{i}(\theta_\mathrm{j}) \ge \pi_\mathrm{ij}$; otherwise $u_\mathrm{ij} = 0$. This process is repeated for n items “administered” to N examinees. There is no practical limit on the number of examinees simulated. Formatting of some of the outputs, however, may become erratic if more than 999,999 examinees are specified.

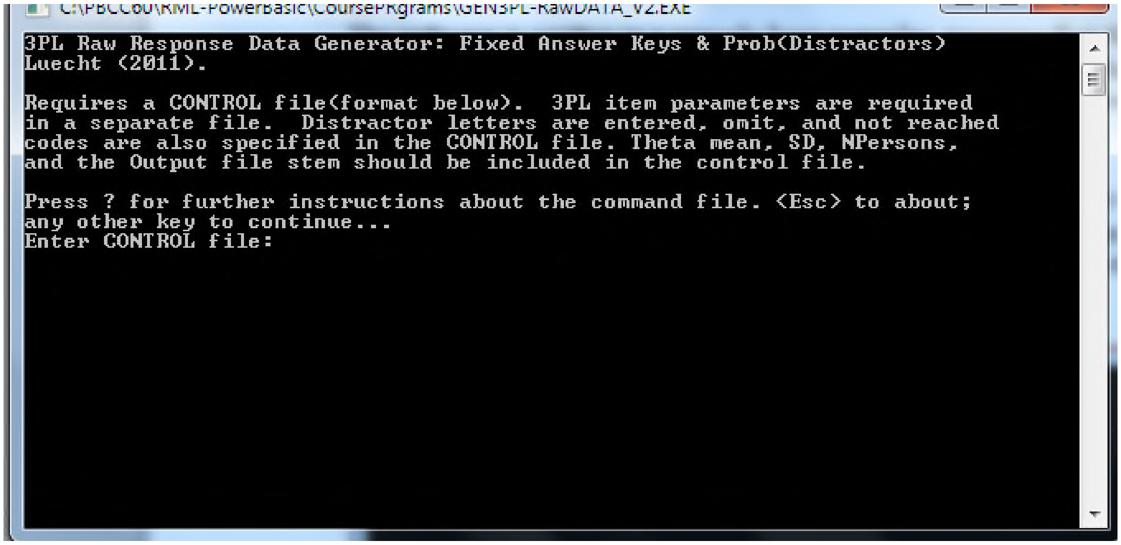
The raw responses are the generated as follows, using the dichotomous response data. All correct responses (u_{ij}=1) are set equal to the user-supplied item answer key. If more than one key is specified for the item, a uniform random selection is made from among all correct answer keys for that item. Incorrect responses are selected proportional to user-supplied weights for each item distractor option. This feature allows some incorrect distractor options to appear more popular than other options. Setting the weight for a particular option to zero even makes it possible to mimic poor MC or SR distractor options that every examinee can exclude as obviously incorrect. If the user-supplied item distractor option weights are not specified, one of the available incorrect distractor choices is selected with uniform probability across the options.

Omitted and/or not-reached item responses can also be simulated at the item level by specifying proportions in the item control file. This makes it possible to somewhat mimic speededness effects for items near the end of the test (not-reached) or items omitted under "penalty-for-guessing" instructions. By tying omissions to the more difficult items, some degree of correlation between omission and item difficulty can be induced.

Running GEN3PL_RawDATA_V2

The software is written in a console-base compiler, meaning that there is no windows interface. The actual program file is a very small (less than 60kB) executable file named **GEN3PL_RawDATA_V2.EXE**. The program can be run in one of two ways.

Manual Control File Entry. Clicking on the GEN3PL_RawDATA_V2.EXE file will start the program. A command window will open and will display some basic instructions, with the comment to press any key. The software will then prompt for the name of the CONTROL File (explained in the next section) as shown in Figure 1. The exact name of your control file, including any drive and folder path specifications (e.g., c:\mywork\mycontrol.con) should be entered.


C:\PBCC60\RML-PowerBasic\CoursePRgrams\GEN3PL-RawDATA_V2.EXE

```
3PL Raw Response Data Generator: Fixed Answer Keys & Prob(Distractors)
Luecht (2011).

Requires a CONTROL file(format below).  3PL item parameters are required
in a separate file.  Distractor letters are entered, omit, and not reached
codes are also specified in the CONTROL file. Theta mean, SD, NPersons,
and the Output file stem should be included in the control file.

Press ? for further instructions about the command file. <Esc> to about;
any other key to continue...
Enter CONTROL file:
```

Figure 1. Opening GEN3PL_RawDATA_V2 Screen (Manual Entry)

Drag-and-Drop Program Start. Note: this option may only work for Windows 7® or higher. Use the Manual Control File Entry method for XP or earlier versions of Windows. Prepare the Control File and store it in the same folder (directory) as the GEN3PL_RawDATA_V2.EXE executable file. Using your mouse, drag your Control File name and drop it on top of the GEN3PL_RawDATA_V2.EXE file. The program will run automatically.

Note that GEN3PL_RawDATA_V2.EXE automatically creates a log file named GEN3PL_RAWDATA_V2_Analysis.LOG in the default folder. If the simulation runs and finishes normally, this log file will automatically be named using your specified Output File stem with the file name extension, .OUT. If, however, the simulation fails or aborts due to errors in the Control File, you should locate and open the GEN3PL_RAWDATA_V2_Analysis.LOG file for specific error messages. This log file is overwritten each time that GEN3PL_RawDATA_V2 is run.

Input Files

GEN3PL_RawDATA_V2 requires two input files: (a) a Control File that specifies the Item File and other analysis set-ups and (b) the Item File that contains all of the necessary item parameters and other data used by the simulation software. Both files should be created in a text editor (Unicode text or ASCII text).

The Control File

The Control File contains up to eight entries. Some of the entries are optional (THETA, DLOG, NR, OMIT, and CODES). If these optional entries are excluded from the Control File, default values will be automatically used.

Table 1 describes the eight entries. Note that these Control File entries are prepared with the simple format: **KEYWORD=***parameter values*. Only the first two letters of each keyword in Table 1 are needed. For example, THETA and TH are identically interpreted by the software.

Table 1. Control File Keywords and Parameter Values

Key Word	Values/Parameters	Status	Default Values
ITEMFILE=	*Item.File.Name* (with path prefix, if applicable), text string	Required	- -
THETA=	*θ.mean#, θ.SD#* (mean & SD of the distribution for θ, two numbers comma-delimited)	Optional	0,1
DLOG=	*Normal.ogive.constant#* (set to 1.0 for logistic, 1.7 for normal ogive, one number)	Optional	1.702
NPERSONS=	*Integer.number.examinee* (total number of examinees to simulate)	Required	--
CODES=	*Response.codes* to use for SR or MC items (e.g., ABCDE). Max. width=15	Optional	ABCD
OMIT=	*Omit.code* (code to use for examinee omitted responses)	Optional	X
NREACHED=	*Not.reached.code* (code to use items not reached by a simulated examinee)	Optional	N
OUTFILE=	*Output.File.Stem* (with path prefix, if applicable, text string)	Required	--

Figure 2 contains a sample Control File. All eight keywords are used, even though the parameter values THETA, DLOG, OMIT and NREACH are the default values (see Table 1) and could have been excluded from this Control File.

```
ITEMFILE=SAMPLE10.CSV
THETA=0.0,1.0
DLOG=1.7
NPERSONS=100
CODES=ABCDE
OMIT=X
NREACH=N
OUTFILE=SAMPLE10_RUN
```

Figure 2. A Sample Control File

The item file is named SAMPLE10.CSV. *N*=100 IRT proficiency scores will be sampled from a unit-normal distribution, $\theta \sim (\mu=0.0, \sigma^2=1.0)$. Five response options (A, B, C, D and E) will be used with omits denoted by "X" and not reached items coded as "N". The output file name stem is SAMPLE10_RUN. Extensions will be appended to this file name stem (see Output files).

Item File

The item file should contain n ≥ 1 item records with a minimum of seven **comma-delimited** entries per line. This file must be likewise be saved in ASCII or Unicode text format. Each item record begins with the a_i, b_i, and c_i parameters. (Note: for the 2PL model, set c_i=0.0 for all items; for the 1PL model, set c_i=0.0 and a_i=1.0 or some other constant for all items.). The item parameters are followed by a domain code. This code is included in the outputs to allow items to be assigned to various strands or to other classification schemes that might be used for item analyses or scoring. The domain codes are otherwise ignored by GEN3PL_RawDATA_V2.

The domain codes are followed by two proportions that determine the number of omitted responses (OR) and the number of examinees who fail to reach (not reached or NR) that item. If these values are set to zero, complete response records result (no omitted or not-reached items). If either value is greater than zero, a uniform random sampling mechanism is employed to approximately omit or change to not reached the specified proportion of the examinees for each item. The proportions are bounded [0,1]. If values greater than one are entered, the program divides by 100, assuming a percentage as the entered value. If values less than one are entered, they are changed automatically to zeros (complete data generated).

The final required entry in the Item File is/are the answer key(s) for each item. The answer key(s) for each item must match one of the CODES indicated in the Control File. Up to 15 answer keys can be answered. Answers are assumed to be the union of all single-best-answer responses. For example, if the key is indicated as "AC", either "A"or "C" would be correct for that item.

Following the answer keys, users can specify a set of positive numbers (weights) indicating the popularity of the incorrect distractor options. Each item can therefore proportionally emphasize certain incorrect options for those simulated examinees who get the item wrong. The number of weights must correspond to the number of CODES specified in the Control File. For example, if CODES=ABCDE, there would need to be five comma-delimited weights for each item in the Item File. If the weights are excluded, uniform weighting is applied for each item. These weights are normalized to proportional weights (e.g. $p_{ik}=w_{ik}/\sum_k w_{ik}$) therefore, any positive values can be entered. If a weight of w_{ik}=0 is entered, that item distractor option will not be selected by any simulated examinee. This feature allows simulating poorly constructed multiple-choice items where some of the response choices are easily ruled out as incorrect.

Figure 3 shows a sample Item File with ten item records. This would be the file referred to as ITEMFILE=SAMPLE10.CSV in the Control file (see Figure 2).

```
0.897006088,0.594710535,0.15,T,0,0,B,5,5,5,5,5
1.355796863,-0.472584611,0.15,T,0,0,C,5,5,5,5,5
1.127243502,0.959738834,0.15,T,0,0,D,5,5,5,5,5
1.223897665,0.143034451,0.15,T,0,0,E,25,5,5,5,5
0.725582928,0.045697297,0.15,T,0,0,A,5,5,5,5,50
0.861350454,-0.386298666,0.15,T,0,0,AC,5,5,5,5,5
0.918238504,-0.150489732,0.15,T,0,0,BD,5,5,5,5,5
1.394568932,-0.967417109,0.15,T,0,0,C,5,5,5,5,5
1.093409322,-1.122972701,0.15,T,0.1,0,B,5,5,5,5,5
1.067523393,-0.986604367,0.15,T,0,0.05,E,5,5,5,5,5
```

Figure 3. Sample Item File (10 Item Records with Distractor Option Weights)

There are 10 item records can be denoted by the syntax:

$$a_i, b_i, c_i, dcode_i, p.omit_i, p.nr_i, ans.keys_i, w_{i1}, w_{i2}, \ldots, w_{im}$$

where a_i is an IRT discrimination parameter, b_i is an IRT item difficulty, c_i is an IRT lower asymptote parameter, $dcode_i$ is a domain code, $p.omit_i$ is the approximate proportion of examinees in the sample who will omit for this item, $p.nr_i$ is the proportion of examinees in this sample who will not reach this item, and $w_{i1}, \ldots w_{im}$ are the weights for the distractor options (m=number of distractors as specified by the CODES statement in the Control File).

Output Files

Five output files are generated by the software, each time that the program is run. A summary output file is created with the name *Output.File.Stem*.**OUT**, using the user-supplied *Output.File.Stem* (see OUTFILE= in the Control File description). This .OUT file contains a description of the inputs specified and the generation process, including raw score statistics, counts of omitted and not-reached responses per item (if any were specified), and other aggregated results. The scored dichotomous (0,1) responses are saved to *Output.File.Stem*.**RSP**. Omitted responses have a "9" substituted; not-reached responses are coded as "8". Note that, if *p.omit*=0 and *p.nr*=0 (see Item File), there will be no omitted or not reached items (i.e., the data recodes will have n responses for each simulated examinee). The raw (unscored) responses are stored to *Output.File.Stem*.**DAT.** Omitted and not-reached responses, if any, are coded using the user-supplied codes specified in the Control File (defaults: omitted=X, not reached=N).

The format of the scored and raw response files is as follows

Columns	Contents
9 to 14	Raw number-correct score, $X_j = \sum u_{ij}$
16 to [16+(n–1)]	Scored or raw utem response, u_{ij}, i=1,…,n

If read in FORTRAN, the ID variables, X_j, and u_{ij} can be read using the format statement:

(A8,1X,I5,1X,*n*A1)

where "A" is an alpha-numeric character, *n* is the number of items and "X" denotes a skipped character space.

The fourth file generated by the software is the proficiency score file, *Output.File.Stem*.**THT**. These are the "true" proficiency scores, θ_j, for each simulated examinee, j=1,…,N. The format of this file is (A9, F11.5). The Person ID is in columns 1 to 8 (zero-filled). The sampled θ_j values are in columns 9 to 19 (five decimal places of precision).

The final file created by the software is a item control file, *Output.File.Stem*.**ITM**. This file contains the data that is sometimes used by item analyses (IA) software. The comma-delimited entries for each item record are:

1. Item number (1,…,*n*)
2. Answer keys (concatenated together if there is more than one key)
3. Number of distractor options (integer)
4. Domain code (see d.codei in the Item File)
5. Scored status (set to "Y" for "yes",
6. Item type (set to "M" for multiple-choice)

The author makes no warranties or guarantees, expressed or implied, about the software or its outputs. Users agree to assume full responsibility by downloading and using the software. Citation: Luecht, R. M. (2011). *Gen3PL Raw Data (Version 2)*. Greensboro, NC: [Author]. Send comments to: Richard M. Luecht, PhD, Educational Research Methodology, School of Education, UNC-Greensboro, Greensboro, NC 27402-6170 (rmluecht@uncg.edu)