Center for Advanced Studies in
Measurement and Assessment

CASMA Research Report

Number 9

Manual for BB-CLASS: A Computer Program that uses the Beta-Binomial Model for Classification Consistency and Accuracy

Version 1.0

Robert L. Brennan

December 2004

Disclaimer of Warranty

No warranties are made, express or implied, that BB-CLASS is free of error, that it is consistent with any particular standard, or that it will meet the requirements of any particular application. The author disclaims any direct or consequential damages resulting from use of this program.

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

Abstract

BB-CLASS is an ANSI C computer program that uses the *B*eta-*B*inomial model (and its extensions) for *CLASS*ification consistency and accuracy. It is intended to provide results for both the Hanson and Brennan (1990) and Livingston and Lewis (1995) procedures, although BB-CLASS has some capabilities that slightly extend these procedures. Both Macintosh and PC versions of BB-CLASS are available.

Introduction

BB-CLASS is an ANSI C computer program that uses the *B*eta-*B*inomial model (and its extensions) for *CLASS*ification consistency and accuracy. It is intended to provide results for both the Hanson and Brennan (1990) and Livingston and Lewis (1995) procedures, although BB-CLASS extends these procedures somewhat.[1] The theoretical underpinning of both sets of procedures is the beta-binomial model and its extensions originally introduced by Lord (1964, 1965). (See Hanson, 1991, for a very detailed consideration of method of moments estimates of parameters for these procedures.) Both Macintosh and PC versions of BB-CLASS are available. Although it is assumed that users are familiar with both Hanson and Brennan (1990) and Livingston and Lewis (1995), important features of these procedures are summarized next.

Hanson and Brennan (1990) Procedures

For the Hanson and Brennan (1990) procedures, the data modeled are raw scores for the sum of n equally weighted, dichotomously-scored items. In general, the probability that the raw score random variable X equals i ($i = 0, \ldots, n$) is

$$\Pr(X = i) = \int_0^1 \Pr(X = i|\tau)\, g(\tau)\, d\tau, \tag{1}$$

where τ is the proportion-correct true score, $g(\tau)$ is the distribution of true scores, and $\Pr(X = i|\tau)$ is the conditional error distribution. In the Hanson and Brennan (1990) procedures, $g(\tau)$ can be either the two- or four-parameter beta distribution, and $\Pr(X = i|\tau)$ can be the binomial distribution or Lord's (1965) two-term approximation to the compound binomial distribution. The two-parameter beta consists of two shape parameters, α and β. The four-parameter beta consists of two shape parameters as well as lower (l) and upper (u) limits of the true score distribution.[2]

The raw score random variables, X_1 and X_2, for two independent administrations have a bivariate pdf given by

$$\Pr(X_1 = i, X_2 = j) = \int_0^1 \Pr(X_1 = i|\tau) \Pr(X_2 = j|\tau)\, g(\tau)\, d\tau. \tag{2}$$

The corresponding bivariate cdf is given by

$$\Pr(X_1 <= i, X_2 <= j) = \int_0^1 \Pr(X_1 <= i|\tau) \Pr(X_2 <= j|\tau)\, g(\tau)\, d\tau. \tag{3}$$

[1] Strictly speaking, Hanson and Brennan (1990) discuss classification in the context of only two categories, although they note that the extension to more than two categories is straightforward. Lee, Hanson, and Brennan (2002) compare beta-binomial and IRT procedures for multiple categories.

[2] Technically, for the four-parameter beta, in Equation 1 (and subsequent equations) it would be better if the limits of integration were specified as l and u, rather than 0 and 1, respectively.

For simplicity, assume there are only $K = 2$ categories labeled 0 and 1, and each examinee is classified into one of them according to the following rule: classify the examinee into category 0 if the examinee's raw score is less than x_0; classify the examinee into category 1 if the examinee's raw score is greater than or equal to x_0. It follows that the probability of a consistent classification is

$$\begin{aligned} p &= \sum_{i=0}^{x_0-1} \sum_{j=0}^{x_0-1} \Pr(X_1 = i, X_2 = j) + \sum_{i=x_0}^{n} \sum_{j=x_0}^{n} \Pr(X_1 = i, X_2 = j) \\ &= \Pr(X_1 <= x_0 - 1, X_2 <= x_0 - 1) + \Pr(X_1 >= x_0, X_2 >= x_0), \end{aligned}$$

the probability of a consistent classification by chance is

$$\begin{aligned} p_c &= \left[\sum_{i=0}^{x_0-1} \Pr(X_1 = i)\right] \left[\sum_{j=0}^{x_0-1} \Pr(X_2 = j)\right] + \\ &\qquad \left[\sum_{i=x_0}^{n} \Pr(X_1 = i)\right] \left[\sum_{j=x_0}^{n} \Pr(X_2 = j)\right] \\ &= \Pr(X_1 <= x_0 - 1) \Pr(X_2 <= x_0 - 1) + \Pr(X_1 >= x_0) \Pr(X_2 >= x_0), \end{aligned}$$

and coefficient κ is given by

$$\kappa = \frac{p - p_c}{1 - p_c}.$$

Classification consistency indices are based on a comparison of scores for two administrations of a test—i.e., two *observed* score distributions. By contrast, classification accuracy is based on a comparison of observed scores and true scores. Let τ_0 be a true cut score such that examinees pass if they have true scores greater than or equal to τ_0, and they fail otherwise. Then, traditionally, classification accuracy (or more correctly inaccuracy) is quantified by false positive and false negative error rates. The false positive rate is

$$\int_0^{\tau_0} \sum_{i=x_0}^{n} \Pr(X = i|\tau)\, g(\tau)\, d\tau = \int_0^{\tau_0} \Pr(X >= x_0|\tau)\, g(\tau)\, d\tau,$$

and the false negative rate is

$$\int_{\tau_0}^{1} \sum_{i=0}^{x_0-1} \Pr(X = i|\tau)\, g(\tau)\, d\tau = \int_{\tau_0}^{1} \Pr(X <= x_0 - 1|\tau)\, g(\tau)\, d\tau.$$

Often the true cut score is set equal to the proportion-correct version of the observed cut score—i.e., $\tau_0 = x_0/n$. There is nothing in the theory that requires this equality, however.

Livingston and Lewis (1995) Procedures

The Hanson and Brennan (1990) procedures assume that a test consists of n equally weighted, dichotomously-scored items. By contrast, suppose (a) items

are not equally weighted and/or (b) some or all of the items are polytomously scored. The Livingston and Lewis (1995) procedures are intended to handle these and other "complex" situations through an ad hoc extension of the beta-binomial procedures discussed by Lord (1965) and Hanson and Brennan (1990). The essence of the extension is that Livingston and Lewis (1995) substitute a so-called "effective test length" (denoted $\tilde{n}$ here) for an actual number of dichotomously-scored items. The formula for effective test length that they suggest is

$$\tilde{n} = \frac{(\mu_x - X_{min})(X_{max} - \mu_x) - r\sigma_x^2}{\sigma_x^2(1-r)},$$

where X_{min} is the lowest score for X, X_{max} is the highest score, μ_x is the mean, σ_x^2 is the variance, and r is the reliability. In this case, X refers to reported scores (rounded to integers), which need not be raw scores in the sense of numbers of items correct, numbers of points earned, etc. Since $\tilde{n}$ can be a non-integer, the actual value used by BB-CLASS is rounded to the nearest integer, which is denoted $\tilde{n}'$ here.[3]

As noted above, the Hanson and Brennan (1990) procedures were specified for either the two- or the four-parameter beta true score distribution, and for either the binomial or the compound-binomial error distribution. Livingston and Lewis (1995) confined their discussion to the binomial error model and the four-parameter beta true score distribution. The implementation of the Livingston and Lewis (1995) procedures in BB-CLASS always uses the binomial error model, but it permits either the two- or the four-parameter beta true score distribution.

Another difference between the Hanson and Brennan (1990) and Livingston and Lewis (1995) procedures involves the observed score distributions that are used for estimating classification consistency and accuracy results. Hanson and Brennan (1990) always use the observed score distributions predicted from the model. By contrast, for classification consistency, Livingston and Lewis (1995) compare the actual observed score distribution with the observed score distribution for an "alternate form" predicted from the model, and for classification accuracy they compare the actual observed score distribution with the true score distribution predicted from the model. BB-CLASS provides results for both approaches for both sets of procedures.

Executing BB-CLASS

To execute BB-CLASS, the user double-clicks the BB-CLASS icon. BB-CLASS then prompts the user for the name of the file containing the control cards. This file must be in the same folder as the BB-CLASS application, or the full pathname for the control cards must be specified. After the user types a return, BB-CLASS executes. When execution is complete (usually only a second or

[3]The Livingston and Lewis (1995) use of the phrase "effective test length" should not be confused with other, more traditional uses of this phrase (see, for example, Feldt & Brennan, 1989, p. 111).

two), the message "Successful execution" is printed in the same window used to specify the name of the control cards file.

In the text of this manual, variables are put in italics, user input is put in typewriter type style, and file names are enclosed by double quotation marks. Note, however, that the name of the file does not include the quotation marks.

Control Cards

A run of BB-CLASS requires a file containing a set of three control cards, and a file containing either raw data or a frequency distribution. All files should be in text-only format.

For each control card, all parameters are separated from each other by any number of spaces and/or tabs. Unless otherwise specified, the order in which parameters are provided is fixed. BB-CLASS looks for a linebreak (newline or return) character at the end of each line, which is generated by typing a return. Note that the linebreak produced by hitting the return key generates different ASCII code under Macintosh and PC/Windows/DOS operating systems. Therefore, a control cards file generated using a Macintosh computer will usually not work as input for a PC.[4]

Card 1: Procedures

type — type of procedure. Use `HB` for the Hanson and Brennan (1990) procedures; use `LL` for the Livingston and Lewis (1995) procedures.

r — reliability. If `HB` is specified and $r = 0$ then the binomial error model is used. If `HB` is specified and $r > 0$, then Lord's (1965) two-term approximation to the compound-binominal model is used. If `LL` is specified, then r *must* be greater than 0, because a value for r is required to estimate "effective test length."

nparm — set to `2` for the two-parameter beta true score distribution; set to `4` for the four-parameter beta true score distribution.

The following fields are optional and may occur in any order.

`xfit` — `xfit` followed by an integer, say a, means that a Pearson chi-square value is computed for cells that have a fitted frequency greater than a (default is 0).

`Gquad`, `EHquad`, or `EDquad` — type of quadrature procedure used for numerical integration (default is `EDquad`). (See discussion on pages 5 and 20.)

[4]Some text editors for a Macintosh allow users to generate DOS newlines, and some text editors for a PC allow users to generate Macintosh newlines.

`npts` — `npts` followed by an integer, say b, means use b equally spaced quadrature points if `EHquad` or `EDquad` is specified (default is 1000).

`check` — `check` means compute the sum of the bivariate probabilities in Equation 2 (default is no computation). (See discussion at the bottom of this page and on 21.)

Quadrature

Quadrature is a process for doing numerical integration, in which the integration is replaced with a summation. BB-CLASS permits the user to choose among three procedures for performing numerical integration:

- `Gquad`: 64 point Gauss quadrature, which was used by Hanson and Brennan (1990, p. 349), who suggest that it works well when both of the shape parameters of the beta distribution are greater than 1. The quadrature points range from the lower limit to the upper limit of π, but the points are not equally spaced for Gauss quadrature.

- `EHquad`: quadrature with a specific number of *equally* spaced points for π. For each of the points the "proportional height" of the true score distribution is determined (i.e., the sum of the heights equals 1). Livingston and Lewis (1995) propose using this procedure with 100 points, which is specified by using `"npts 100"` (without the quotes) in the Procedures control card. This author's experience suggests that 1000 points is a better choice, which is the default when `EHquad` is chosen. (Additional comments about `EHquad` are provided on page 20 in the section entitled *Other Issues*.)

- `EDquad`: quadrature with a specific number of *equally* spaced points for π using the density in the interval for each point, as opposed to the proportional height at the mid-point of each interval (as in `EHquad`). For example, if there are 1000 points, the lower limit is 0, and the upper limit is 1, then the interval for the i-th point is $(i/1000 - .0005, i/1000 + .0005)$, and the density in this interval is used rather than the proportional height at $i/1000$. The default procedure is `EDquad` with 1000 points, which was chosen as the default because it gives results that are generally closer to certain results known by the author to be true a priori. Obviously, this does not guarantee that `EDquad` will always give absolutely correct results. (Additional comments about `EDquad` are provided on page 21 in the section entitled *Other Issues*.)

Bivariate Checking

One way to examine how well a quadrature procedure is working is to determine whether it leads to a result that is known a priori. For example, the sum of the

terms given by Equation 2 must equal 1; i.e,

$$\sum_{i=0}^{n}\sum_{j=0}^{n}\Pr(X_1 = i, X_2 = j) = 1. \tag{4}$$

(For `LL`, n is replaced by the rounded value of $\tilde{n}$, which we denote $\tilde{n}'$.) If the user specifies `check`, then this sum is determined for the particular quadrature procedure chosen. Doing so can take a few seconds, depending primarily on the value of n (or $\tilde{n}'$).

The numerical procedures used by BB-CLASS do not actually require computing all of the $n \times \ n$ terms in Equation 4, because results are needed only at the cut scores. So, in a sense, the bivariate checking that is done is a more stringent than necessary.

Card 2: Input Data

`data[]` name of file containing input data. The filename must be enclosed in double quotation marks; the quotation marks are not part of the filename. The file must be located in the same folder as the application, or the full pathname must be provided. The input data filename must be different from the filename containing the control cards.

`input_data` specify `R` or `r` if the input file contains a listing of raw scores; specify `F` or `f` if the input file contains a frequency distribution of raw scores. Here, the phrase "raw scores" simply means the reported scores on the test, which need not be number of items correct, number of points obtained, etc.

BB-CLASS can also be executed using as input the first four raw-score moments, as opposed to reading in the full observed score distribution. To do so, specify `input_data` as `M` or `m`. Using raw-score moments as input is discussed on page 19 in the *Other Issues* section of this manual.

`xcol` column (an integer) for reading scores; not required if `input_data` is `M` or `m`. Columns must be delimited with white space (e.g., blanks and/or tabs).

`fcol` column (an integer) for reading frequencies; required only if `input_data` is `F` or `f`. Columns must be delimited with white space (e.g., blanks and/or tabs).

Card 3: Cut Scores

K number of categories.

$xcut[*]$ $K - 1$ raw cut scores.

$tcut[*]$ $K-1$ true cut scores in the proportion-correct metric (optional). If the true cut scores are not specified, they are set to the proportion-correct raw cut scores,

$$tcut[*] = (xcut[*] - X_{min})/(X_{max} - X_{min}),$$

which are simply $tcut[*] = xcut[*]/n$ for the Hanson and Brennan (1990) procedures.

Metric Conventions

In representing and discussing the beta-binomial model (and its extensions), it is typical practice to use the raw-score metric (e.g., number of items correct) for observed scores and the proportion-correct metric (i.e., a number between 0 and 1) for true scores. This conventional practice is followed in BB-CLASS.

Also, note that when `input_data` is `F` or `f`, usually actual frequencies would be used as input such that the sum of the frequencies is the total number of examinees. However, proportions could be substituted for frequencies such that the proportions sum to 1.[5] The output from BB-CLASS does not actually depend upon the number of examinees.

Output Files

BB-CLASS generates three output files. Letting "cc" (without the quotes) be a generic designator for the name of the file containing the control cards, these output files are identified as:

- "cc out",
- "cc true_dist", and
- "cc observed_dist",

(without the quotes). The file "cc out" contains the principal output, which is described in the context of two examples in the next two sections.

The file "cc true_dist" provides the true score distribution. More specifically, for each of the `npts` intervals (default is 1000), the following are provided: midpoint, relative frequency (height) at the midpoint, pdf value for the interval, and cdf value at the upper limit of the interval.

The file "cc observed_dist" provides the observed score distribution. More specifically, for each of the $n+1$ possible observed scores, the following are provided: raw score, raw proportion, fitted proportion, raw frequency, and fitted frequency. The raw/fitted proportion is simply the raw/fitted frequency divided by the sample size. For the Livingston and Lewis (1995) procedures, the observed scores are defined to be $0, \ldots, \tilde{n}'$, where $\tilde{n}'$ is the rounded value of $\tilde{n}$ (effective test length).

[5] If proportions are used, of course, the number of examinees printed in the output will not be the actual number of examinees in the user's data.

Hanson and Brennan (1990) Example

Table 1 provides a listing of control cards and data for a run of BB-CLASS that provides Hanson and Brennan (1990) results for an example distributed with Hanson's (1995) *Class Consistency* program. In the first control card

- `HB` means perform Hanson and Brennan (1990) computations,
- 0. means use the binomial error model,
- 4 means use the four-parameter beta distribution for true scores, and
- `check` means perform bivariate checking.

The second control card says that the data are in a file named "act288m" (without the quotes) which contains a frequency distribution (i.e., `f`) with scores in column 1 and frequencies in column 2. The third control card says that there are 2 categories with a raw cut score of 24; i.e., examinees with 23 or fewer items correct are classified into the first category, and examinees with 24 or more items correct are classified into the second category.

The number of items is not specified in the control cards. BB-CLASS determines the number of items from the data. In this case the data are provided in the form of a frequency distribution, which means that the number of records (or lines) in "act288m" must be the number of items plus one, which is 41 for this example. In Table 1 to save space the frequency distribution is provided in four columns. The file itself, however, must have 41 lines with two numbers (a score followed by a frequency) in each line.

The output for the run of BB-CLASS in Table 1 is provided in Tables 2–4. In Table 2, the BB-CLASS header is followed by a listing of the control cards. Note that the control cards were in a file named "ccHB" (without the quotes). Because the input reliability was set at 0, Lord's k is 0, which means that the binomial error model will be used. Both the raw cut score and the true cut score are listed. Since there was no true cut score provided in the control cards, it is set to $24/40 = .6$.

The middle part of the output is provided in Table 3. Recall that the control cards request that the four-parameter beta distribution be used, but in this case only three parameters could be fit using the method of moments (see Hanson, 1991).

Provided next are the raw-score, fitted raw-score, and true-score moments. The first three moments for the raw-score and fitted raw-score distributions are the same, which is a direct reflection of the previously noted fact that three moments were fit. Two chi-square statistics are provided for evaluating the similarity of the raw-score and fitted raw-score distributions. If a chi-square value is exceptionally large (for its degrees of freedom), then there is reason to doubt that the model is appropriate.

An estimate of reliability and the overall standard error of measurement (SEM) are provided based on the true-score and raw-score moments. For this

Table 1: Control Cards and Frequency Distribution for Hanson and Brennan (1990) Example

```
HB 0. 4  check
"act288m" f 1 2
2   24.


0 0         10 9597     20 4367     30 1967    40 294
1 23        11 9809     21 4083     31 1874
2 98        12 9674     22 3651     32 1710
3 384       13 8971     23 3333     33 1581
4 986       14 8033     24 3191     34 1503
5 2161      15 7384     25 2899     35 1349
6 3722      16 6758     26 2644     36 1181
7 5623      17 6004     27 2597     37 994
8 7533      18 5463     28 2287     38 827
9 8817      19 4896     29 2197     39 585
```

example,

$$\text{Reliability} = \frac{7.517^2}{8.049^2} = .872$$

and

$$\text{SEM} = 8.049\sqrt{1 - .872} = 2.878.$$

The manner in which numerical integration was performed is specified next. The control cards do not specify which procedure to use; consequently, the default procedure, `EDquad`, was used. Since the control cards specify `check`, the sum of the bivariate probabilities is provided. The sum is 1.00000 for this example, which gives us confidence that the quadrature procedure is working well.

The final part of the output is in Table 4. The first two contingency tables provide classification accuracy and consistency results, respectively, of the type used by Hanson and Brennan (1990). The last two contingency tables provide classification accuracy and consistency results, respectively, of the type used by Livingston and Lewis (1995). More specifically:

- For the first contingency table, rows represent category true scores and columns represent *expected* observed scores for categories under the model (in this case the three-parameter beta binomial model). The column and row identifiers for the body of the contingency table are always specified as `x0 ... x(K-1)` and `t0 ... t(K-1)`, respectively. The probability of a

correct classification (sum of diagonal entries), the false positive error rate, and the false negative error rate are provided below the contingency table.

- For the second contingency table, both rows and columns represent *expected* observed scores for categories under the model. The column and row identifiers for the body of the contingency table are always specified as `x0` ... `x(K-1)`. The usual classification consistency indices are provided below the contingency table. That is, `pc` is the proportion of consistent decisions, `pchance` is the chance proportion of consistent decisions, and `kappa` is the kappa statistic. The probability of a misclassification (sum of off-diagonal entries) is also provided.

- For the third contingency table, rows represent category true scores and columns represent *actual* observed scores. Note that the column marginals match the "category proportions in original data" reported earlier in the output (see Table 2).

- For the fourth contingency table, rows represent *expected* observed scores for categories under the model, and columns represent *actual* observed scores for categories.

Table 2: Control Cards and Frequency Distribution for Hanson and Brennan (1990) Example

```
************************************************************************
*** BB-CLASS:  Beta-Binomial Classification Consistency and Accuracy ***
***                        Version 1.0                               ***
***                                                                  ***
***                      Robert L. Brennan                           ***
***                            CASMA                                 ***
***                     University of Iowa                           ***
***                                                                  ***
***                        December 2004                             ***
***                                                                  ***
***                     All Rights Reserved                          ***
************************************************************************

*** Hanson and Brennan Results ***

*** Listing of Control Cards in ccHB ***

HB 0. 4  check
"act288m" f 1 2
2   24.

*********************************************

            Number of examinees =  151050.00000
              Input reliability =       0.00000
                    Test length =      40
                       Lord's k =       0.00000   (binomial error model)

Number of Categories = 2

Cut Scores:   xcut[] =  24.00000
              tcut[] =   0.60000

Category proportions in original data:
               0.80351   0.19649
```

Table 3: Control Cards and Frequency Distribution for Hanson and Brennan (1990) Example (continued)

```
***Parameter Estimates for Beta Distribution***

      alpha        beta   low limit   upp limit

   0.523779    1.625693    0.223172    1.000000

Number of moments fit:        3

***Moments (Raw, Fitted Raw, True)***

                  Mean        S.D.        Skew        Kurt

       Raw   16.498709    8.048720    0.829364    2.965899
Fitted Raw   16.498709    8.048720    0.829364    2.925241
      True   16.498709    7.516707    1.021434    3.158296

Likelihood Ratio Chi-Square =  339.84519 (with df = 38)
Pearson Chi-Square =  344.66484 (with df = 38)
  for cells with fitted frequencies greater than  0.00

Reliability (from above moments) =    0.87217
SEM (from above moments) =    2.87767

***Numerical Integration***

Numerical integration performed using 1000 equally-spaced quadrature points
  and the true-score density for each interval

Sum of bivariate probabilities =    1.00000
```

Table 4: Control Cards and Frequency Distribution for Hanson and Brennan (1990) Example (continued)

```
***ACCURACY RELATIVE TO EXPECTED OBSERVED SCORES GIVEN MODEL***

            x0        x1      marg
  t0   0.78247   0.03795   0.82042
  t1   0.01778   0.16180   0.17958

marg   0.80026   0.19974   1.00000

probability of correct classification = 0.94427
false positive rate = 0.03795;  false negative rate =  0.01778

***CONSISTENCY USING EXPECTED OBSERVED SCORES GIVEN MODEL***

            x0        x1      marg
  x0   0.76127   0.03898   0.80026
  x1   0.03898   0.16076   0.19974

marg   0.80026   0.19974   1.00000

pc = 0.92204;  pchance = 0.68031;  kappa = 0.75613
probability of misclassification = 0.07796

***ACCURACY RELATIVE TO ACTUAL OBSERVED SCORES***

            x0        x1      mar
  t0   0.78565   0.03733   0.82298
  t1   0.01785   0.15916   0.17702

marg   0.80351   0.19649   1.00000

probability of correct classification = 0.94482
false positive rate = 0.03733;  false negative rate =  0.01785

***CONSISTENCY USING EXPECTED (row) VS. ACTUAL (column) OBSERVED SCORES***

            x0        x1      marg
  x0   0.76437   0.03835   0.80272
  x1   0.03914   0.15814   0.19728

marg   0.80351   0.19649   1.00000

pc = 0.92251;  pchance = 0.68375;  kappa = 0.75498
probability of misclassification = 0.07749
```

Livingston and Lewis (1995) Example

Table 5 provides a listing of control cards and data for a run of BB-CLASS that provides Livingston and Lewis (1995) results for a hypothetical example. In the first control card

- `LL` means perform Livingston and Lewis (1995) computations,
- 0.9 is the estimated reliability,
- 4 means use the four-parameter beta distribution for true scores, and
- `check` means perform bivariate checking.

The second control card says that the data are in a file named "LL data" (without the quotes) which contains a frequency distribution (i.e., `f`) with scores in column 1 and frequencies in column 2. The third control card says that there are 3 categories with raw cut scores of 140 and 160, and true cut scores of .4 and .6.

In Table 5 the frequency distribution is provided in four pairs of columns to save space, with scores ranging from 121 to 190. The "LL data" file itself, however, must have $190 - 121 + 1 = 70$ lines with two numbers (a score followed by a frequency) in each line.

The format of the output for an `LL` run of BB-CLASS is very much like that of the `HB` example, but an `LL` run of BB-CLASS contains some additional information, as indicated in Table 6. For example, the effective test length ($\tilde{n}$) and its rounded value ($\tilde{n}'$) are reported. Also reported are the first four moments, the minimum score, and the maximum score for:

- raw scores, X (i.e., the reported scores),
- raw scores transformed to a scale of 0 to 1 (proportional scores):

$$p = \frac{X - X_{min}}{X_{max} - X_{min}},$$

  and

- raw scores transformed to the 0 to $\tilde{n}'$ metric:

$$X' = \tilde{n}'p = \tilde{n}' \frac{X - X_{min}}{X_{max} - X_{min}}.$$

These scores, and the reasons for computing them, are discussed by Livingston and Lewis (1995). Basically, the model is applied to the X' scores using $\tilde{n}'$ is the number of dichotomously-scored items.

Table 7 provides the middle part of the output. The interpretation of these results is the same as that for the results in Table 3 for the `HB` example, keeping in mind that in Table 7 "Raw" scores are now X' scores.

Table 5: Control Cards and Frequency Distribution for Livingston and Lewis (1995) Example

```
LL 0.9 4    check
"LL data" f 1 2
3  140. 160.   .4  .6


121 3      141 5      161 14     181 8
122 5      142 20     162 17     182 3
123 8      143 11     163 17     183 9
124 5      144 14     164 23     184 0
125 3      145 15     165 29     185 7
126 9      146 21     166 19     186 5
127 2      147 13     167 16     187 0
128 2      148 12     168 33     188 2
129 9      149 10     169 12     189 1
130 18     150 18     170 34     190 1
131 10     151 18     171 16
132 11     152 17     172 21
133 13     153 8      173 17
134 12     154 21     174 32
135 10     155 6      175 0
136 11     156 33     176 32
137 16     157 32     177 22
138 11     158 7      178 14
139 16     159 17     179 8
140 15     160 36     180 25
```

The final part of the output is provided in Table 8. The interpretation of these results is the same as that for the results in Table 4, with the obvious understanding that this `LL` example has three categories, not two. With more than two categories, the false positive rate is defined as the sum of the upper off-diagonal elements (e.g., .01090 + .00000 + .01146 = .02236); similarly, the false negative rate is defined as the sum of the lower off-diagonal elements. The probability of a correct classification is the sum of the diagonal elements.

Table 6: Control Cards and Frequency Distribution for Livingston and Lewis (1995) Example

```
************************************************************************
*** BB-CLASS:  Beta-Binomial Classification Consistency and Accuracy ***
***                         Version 1.0                              ***
***                                                                  ***
***                      Robert L. Brennan                           ***
***                            CASMA                                 ***
***                     University of Iowa                           ***
***                                                                  ***
***                        December 2004                             ***
***                                                                  ***
***                     All Rights Reserved                          ***
************************************************************************

*** Livingston and Lewis Results ***

*** Listing of Control Cards in ccLL ***

LL 0.9 4    check
"LL data" f 1 2
3  140. 160.   .4  .6

*********************************************

            Number of examinees =    1000.00000
              Input reliability =       0.90000
         Effective test length =      49.70252
Effective test length (rounded) =      50
                       Lord's k =       0.00000   (binomial error model)

Number of Categories = 3

Cut Scores:   xcut[] = 140.00000 160.00000
         xprimecut[] =  21.91011  33.14607
              tcut[] =   0.40000   0.60000

Category proportions in original data:
               0.21400   0.31300   0.47300

***Moments used by Livingston and Lewis Procedure***

              Mean        S.D.        Skew        Kurt         Min         Max

     x  155.244000   17.921006   -0.499583    2.543831  101.000000  190.000000
     p    0.609483    0.201360   -0.499583    2.543831    0.000000    1.000000
Raw=x'   30.474157   10.067981   -0.499583    2.543831    0.000000   50.000000
```

Table 7: Control Cards and Frequency Distribution for Livingston and Lewis (1995) Example (continued)

```
***Parameter Estimates for Beta Distribution***

      alpha        beta   low limit   upp limit

   2.666934    1.302899    0.000000    0.907239

Number of moments fit:        3

***Moments (Raw, Fitted Raw, True)***

                  Mean        S.D.        Skew        Kurt

       Raw   30.474157   10.067981   -0.499583    2.543831
Fitted Raw   30.474157   10.067981   -0.499583    2.515916
      True   30.474157    9.554546   -0.546516    2.561804

Likelihood Ratio Chi-Square =  158.33372 (with df = 48)
Pearson Chi-Square =  151.27795 (with df = 48)
  for cells with fitted frequencies greater than  0.00

Reliability (from above moments) =    0.90061
SEM (from above moments) =    3.17410

***Numerical Integration***

Numerical integration performed using 1000 equally-spaced quadrature points
  and the true-score density for each interval

Sum of bivariate probabilities =    1.00000
```

Table 8: Control Cards and Frequency Distribution for Livingston and Lewis (1995) Example (continued)

```
***ACCURACY RELATIVE TO EXPECTED OBSERVED SCORES GIVEN MODEL***

           x0        x1        x2      marg
  t0   0.14951   0.01090   0.00000   0.16042
  t1   0.06173   0.20016   0.01146   0.27335
  t2   0.00045   0.12324   0.44255   0.56624

marg   0.21169   0.33430   0.45401   1.00000

probability of correct classification = 0.79222
false positive rate = 0.02236;  false negative rate =  0.18542

***CONSISTENCY USING EXPECTED OBSERVED SCORES GIVEN MODEL***

           x0        x1        x2      marg
  x0   0.16625   0.04479   0.00066   0.21169
  x1   0.04479   0.22121   0.06830   0.33430
  x2   0.00066   0.06830   0.38505   0.45401

marg   0.21169   0.33430   0.45401   1.00000

pc = 0.77251;  pchance = 0.36269;  kappa = 0.64304
probability of misclassification = 0.22749

***ACCURACY RELATIVE TO ACTUAL OBSERVED SCORES***

           x0        x1        x2      marg
  t0   0.15114   0.01021   0.00000   0.16135
  t1   0.06240   0.18741   0.01194   0.26174
  t2   0.00046   0.11539   0.46106   0.57690

marg   0.21400   0.31300   0.47300   1.00000

probability of correct classification = 0.79961
false positive rate = 0.02215;  false negative rate =  0.17824

***CONSISTENCY USING EXPECTED (row) VS. ACTUAL (column) OBSERVED SCORES***

           x0        x1        x2      marg
  x0   0.16806   0.04193   0.00068   0.21068
  x1   0.04527   0.20712   0.07116   0.32355
  x2   0.00066   0.06395   0.40116   0.46577

marg   0.21400   0.31300   0.47300   1.00000

pc = 0.77634;  pchance = 0.36667;  kappa = 0.64685
probability of misclassification = 0.22366
```

Other Issues

BB-CLASS makes use of several functions in Press, Teukolsky, Vetterling, and Flannery (1992). Many other functions were written by Bradley A. Hanson, with some revisions made by the author. Other functions were written entirely by the author.

As noted previously, it is especially important that the control cards and frequencies files use the type of linebreak appropriate to the application. That is, when BB-CLASS is used with a Macintosh, the linebreaks should be the Macintosh type, and when BB-CLASS is used with a PC/Windows/DOS operating system, the linebreaks should be the DOS type.

Users should note that there is very little literature that examines how well the Livingston and Lewis (1995) ad hoc procedures actually work. Consequently, no claim is made here about the adequacy of these procedures. For example, to this author, it seems likely that the adequacy of the procedures depends in part on the extent to which dichotomous and polytomous items are measuring the same construct. Whether or not this speculative statement is correct, it seems reasonable to suggest that the Livingston and Lewis (1995) procedures be subjected to considerably more research. Note also that the Livingston and Lewis do not provide a computer program for their procedures.[6]

Using Raw-Score Moments as Input

In the discussion of the second control card on page 6, it was noted that BB-CLASS can be run using raw score moments as input, as opposed to the full distribution of raw scores. To do so, the `input_data` variable should be specified as `M` or `m`, and the input data file (`data[]`) should contain (in order) the sample size (number of examinees), the first four raw-score moments (mean, standard deviation, skewness, and kurtosis), the minimum raw score, the maximum raw score, and the proportions of examinees in the K categories in the raw data. These values may be in a single record (i.e., line) or split over as many records as desired.[7]

For example, the results for the Hanson and Brennan (1990) example can be obtained if the control cards are

```
HB 0. 4  check
"HBdata moments" m
2   24.
```

and the "HBdata moments" file contains

```
151050    16.498709 8.048720 .829364 2.965899    0  40 .80351 .19649
```

[6] The Livingston and Lewis (1995) paper does not always provide the level of detail needed to write a computer program to implement the procedures. For example, the paper does not distinguish clearly between $\tilde{n}$ and $\tilde{n}'$. Consequently, occasionally the author of BB-CLASS relied on his own judgement about implementation procedures.

[7] For purposes or reading input, the fscanf() function in C treats a newline or return character just like a space or tab (i.e., "white space").

Note that for an `HB` run of BB-CLASS the minimum raw score must be 0, and the maximum raw score is the number of items.

Similarly, the results for the Livingston and Lewis (1995) example can be obtained if the control cards are

```
LL 0.9 4    check
"LL data moments" m
3  140. 160.   .4  .6
```

and the "LL data moments" file contains

```
1000
155.244000
17.921006
-.499583
2.543831
101
190
.214
.313
.473
```

When moments are used as input, BB-CLASS requires that the user input the number of examinees. Strictly speaking, however, results do not depend on the number of examinees. Still, for documentation purposes it is best to specify the actual number of persons, if it is known; if not, choose some arbitrary positive number.

In the BB-CLASS output, the last two contingency tables provide classification accuracy and consistency results, respectively, of the type used by Livingston and Lewis (1995). This is the only output that depends on the proportions of examinees in the K categories in the raw data. Consequently, if the user is not interested in these two contingency tables, arbitrary values for the K proportions can be used, provided they sum to 1.

With one exception, the output using raw-score moments as input will be identical to that produced using the full raw-score distribution as input—assuming, of course, that the moments are provided with sufficient accuracy (six decimal digits are suggested). The one exception is that chi-square values are not provided.

More about Quadrature in BB-CLASS

Consider Equation 3. The quadrature procedure `EHquad` replaces this equation with

$$\Pr(X_1 <= i, X_2 <= j) \doteq \sum_{m=1}^{\text{npts}} \Pr(X_1 <= i|\tau_m) \Pr(X_2 <= j|\tau_m)\, g(\tau_m), \quad (5)$$

where $\tau_m = m/\text{npts} - 1/(2\,\text{npts})$, and $g(\tau_m)$ is the proportional height of the true score distribution at τ_m. For example, if npts = 100, it follows that

$\tau_m = .005, .015, \ldots, .995$. Livingston and Lewis (1995, p. 184) suggest using an additional linear interpolation adjustment. Such an adjustment is used in BB-CLASS when `LL` is specified.

The quadrature procedure `EDquad` replaces Equation 3 with

$$\Pr(X_1 <= i, X_2 <= j) \doteq \sum_{m=1}^{\text{npts}} \Pr(X_1 <= i|\tau_m) \Pr(X_2 <= j|\tau_m) [G(\tau_{m_u}) - G(\tau_{m_l})], \quad (6)$$

where $G(*)$ designates the true-score cdf, and τ_{m_l} and τ_{m_u} are the lower and upper limits, respectively of the interval with a midpoint of τ_m. For example, if npts = 100, then the second interval is $(\tau_{2_l}, \tau_{2_u}) = (01, .02)$ with a midpoint of $\tau_2 = .015$. For the two parameter beta-binomial model, $G(*)$ is the incomplete beta distribution, which is relatively easy (and quick) to evaluate. Also, for the two-parameter beta binomial model, $\Pr(X <= i|\tau)$ can be expressed in terms of the incomplete beta distribution, which makes it relatively easy (and quick) to evaluate, too. This means that, for the two-parameter beta binomial model, all of the terms in Equation 6 can be evaluated using the incomplete beta. Computation is somewhat more complicated for the four-parameter beta binomial or compound binomial models.

The above discussion of quadrature procedures uses Equation 3 for illustrative purposes. Similar discussions could be provided, of course, for any of the other equations cited previously that involve integration.

Computational Accuracy

Hanson's (1995) provided a computer program entitled *Class Consistency* for computing two- and four- parameter beta binomial and compound binomial results of the type discussed by Hanson and Brennan (1990).[8] *Class Consistency* uses the method-of-moments estimation procedures discussed extensively by Hanson (1991). In particular, *Class Consistency* uses explicit closed-form formulas rather than quadrature procedures. (The above discussion of `EDquad` hints at these closed-form formulas.) For this reason, it seems sensible to use results from *Class Consistency* as a standard for evaluating the various quadrature procedures in BB-CLASS. In the author's experience, `EDquad` with 1000 intervals gives results that generally match results from *Class Consistency* to at least four decimal places. Note, however, that there are two practical limitations of *Class Consistency*. It is available for Macintosh computers, only, and it can handle only two categories. These limitations, as well as the need for a program to compute the Livingston and Lewis (1995) results, were motivating reasons for programming BB-CLASS.

For the two-parameter beta binomial model, Huynh (1976) provides recursive formulas for obtaining the marginal distribution of X in Equation 1 (the so-called negative hypergeometric distribution) and the bivariate distribution

[8] *Class Consistency* is available from the author of BB-CLASS.

in Equation 2 (the so-called bivariate negative hypergeometric distribution). These recursive formulas can be used to obtain very precise results, which can be compared with results using the quadrature procedures in BB-CLASS. Such comparisons made by the author suggest that `EDquad` with 1000 intervals generally gives very accurate results.

Finally, by specifying `check` in the first control card, the user can determine if Equation 4 is satisfied, although BB-CLASS takes a few seconds to obtain the sum of the $n \times n$ bivariate terms. The author suggests that the few seconds is time well spent. However, even if the sum is not quite 1, BB-CLASS may be working adequately because the principal results given by BB-CLASS depend on the accuracy of the quadrature procedures at the cut scores, only.

References

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes.* ACT Research Report 91-5. Iowa City, IA: ACT, Inc.

Hanson, B. A. (1995). Class consistency—A program for computing classification consistency indexes [Computer software and manual.] Available from Center for Advanced Studies in Measurement and Assessment, University of Iowa, Iowa City, IA.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345–359.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253–264.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*(4), 412–432.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: American Council on Education and Macmillan. (Currently published by Greenwood).

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179–197.

Lord, F. M. (1964). *A strong true-score theory, with applications.* Educational Testing Service Research Bulletin 64-19. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30*, 239–270.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C* (2nd ed.). New York: Cambridge University Press.