Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 15, Number 2, January 2010 ISSN 1531-7714

Use of Robust z in Detecting Unstable Items in Item Response Theory Models

Huynh Huynh, University of South Carolina
Patrick Meyer, University of Virginia

The first part of this paper describes the use of the robust z_R statistic to link test forms using the Rasch (or one-parameter logistic) model. The procedure is then extended to the two-parameter and three-parameter logistic and two-parameter partial credit (2PPC) models. A real set of data was used to illustrate the extension. The linking results illustrate the efficacy of the robust z_R vis-à-vis some of the most commonly used processes such as the Stocking and Lord (1983) linking process.

Educational and psychological assessments often require the use of different forms of the same test. These forms usually have a set of common items that are used to link all resulting scores to a common scale. Item response theory (IRT) software such as the Rasch WINSTEPS (Linacre, 2006) and PARSCALE (Muraki & Bock, 1997) have been used for linking purposes. Although there are many ways to accomplish this task, most statewide assessment programs rely on the "anchor test" design for yearly linking. In many situations, this design calls for a separate calibration of each form and then a process to link all these forms together though a set of "stable" common items. The selection of the common items (aka linkers) is usually based on several criteria. During test form construction, potential linkers are chosen to reflect overall test content and the range of item difficulty level. Linkers are typically chosen to be not too easy or too difficult. Once the tests are administered, final linkers are oftentimes restricted to common items which are reasonably stable across the forms to be linked. It may be noted that when a common test form is administered to two different groups of examinees (i.e. all items are treated as linking items), then stability in item parameters is prerequisite for the invariance of the construct under measurement between these two groups.

For technical work based on the Rasch model, the robust z_R statistic (Huynh, 2000; Huynh & Rawls, 2009) has been used widely in large-scale assessment programs (including South Carolina, Arkansas, Maryland, Minnesota, and New Mexico) in detecting items that are unstable (outliers) in yearly linking/equating. The statistic has also been used in studying stability of item parameters across gender and ethnicity groups (Kim & Huynh, 2009). The purpose of this study is to extend the use of the robust z_R statistic to the three- parameter logistic (3PL) and two-parameter partial credit (2PPC) models. A real set of data will be used to illustrate the extension. The linking results also will be used to illustrate the efficacy of the robust z_R vis-à-vis some of the most commonly used processes such as the Stocking and Lord (1983) linking process.

Use of Robust z Statistic in Detecting Outliers

The robust z statistic originates from robust/resistant statistical procedures. (see Hogg, 1979; Huber, 1964; and Huynh, 1982). Huynh took note that most procedures for detecting outliers are based on a "robustification" of the traditional z statistic. Let D be a variable. Traditionally the z statistic is defined as z = (D – mean)/standard deviation. However, both the mean and standard deviation (SD) are influenced by outlying observations. So in order to pinpoint precisely at the outliers, it may be more efficient to look for a z-like statistic that is *not* affected by the outliers. Let Md and IQR be the median and inter-quartile range. For the

normal distribution, the IQR is equal to 1.35 × SD or SD = 0.74 × IQR or 0.74(IQR). With the quantity 0.74(IQR) emulating the standard deviation, a robust version of the traditional z statistic can be taken as the ratio $z_R = (D - Md)/[0.74(IQR)]$. When the D values come from a normal distribution, the robust z_R statistic follows (asymptotically) a normal distribution with zero mean and unit standard deviation. A level of significance (two-tailed alpha) may be selected and a positive critical value z* may be set. Items with a robust z_R smaller than z* in absolute value will be declared "stable" and other items with a robust z_R greater than or equal to z* in absolute value will be declared as "unstable." It may be noted that some traditional definitions of "outliers" can be framed within the robust z_R context. For example, Agresti and Finlay (2009; p. 54) suggest that an observation is an outlier if it falls more than 1.5(IQR) above the upper quartile or more than 1.5(IQR) below the lower quartile. Assuming that the median is equidistant from the upper and lower quartiles, it can be verified that the above definition corresponds to the critical value z* = 2.7.

Use of Robust z_R in Detecting Unstable Items in the Rasch Model

Now let R_1 and R_2 be the Rasch item difficulties obtained from two separate calibrations. There are two sets of Rasch item difficulties for the linkers. Let $D = R_1 - R_2$ and z_R be the robust statistic associated with each D discrepancy. The South Carolina (SC) linking protocols (2001) calls for two quality indices for Rasch linking: ratio (RSD) of the standard deviations of R_1 and R_2 and the correlation (CORR) between R_1 and R_2. Under perfect conditions for linking, the two sets of Rasch difficulties differ by a constant; thus the optimal values for CORR and RSD are exactly 1. However, due to sampling fluctuations in the calibration process, these values tend to depart from the optimal values. The SC protocol for "acceptable linking results" calls for two criteria: (a) the correlation CORR to be at least .95 and (b) the RSD to be within .9 and 1.1. The next paragraph summarizes the justifications provided by Huynh (2009) for these benchmark values.

According to Huynh (2009), the benchmark value for CORR is a result of a study by Yen (1987) who found that the correlation between the true value of the item location parameter and its estimate was better than 0.97 in many situations. Within the context of classical test theory, the correlation between true and estimated values can be treated as a validity coefficient (r_{val}) whereas the correlation between two estimated values can be treated as a reliability coefficient (r_{rel}). Since $r_{val} < \sqrt{r_{rel}}$, we have $r_{rel} > r_{val}^2$. Since r_{val} is at least 0.97, the other coefficient r_{rel} is at least $(0.97)^2$ or about 0.95. Also according to Huynh, the two bounds for the ratio RSD are the results of the significance test that the population value of this ratio is one. When the standard deviations of R_1 and R_2 are equal (i.e. when RSD = 1), Pearson correlation between the quantities $(R_1 + R_2)$ and $(R_1 - R_2)$ is exactly zero. The traditional t test for the null hypothesis (H_o) of zero correlation can be used to check that the hypothesis that RSD is equal to 1. Assuming a correlation of 0.95 between R_1 and R_2 and at the 5% level of significance, the null hypothesis that true value of RSD is 1 is acceptable if the observed value of RSD is between 0.9 and 1.1.

When both criteria for "acceptable linking results" are satisfied, then all items are treated as "stable" and all potential linkers are used in the linking process. However, if one of the criteria does not hold, then some very unstable items will be deleted, starting with items with largest robust z_R statistic. (The z_R statistics are computed only once.) This process is stopped when either the criteria are met or 20% of potential linkers have already been deleted. Huynh noted that, in assessment situations like the South Carolina Basic Skills Assessment Program (BSAP), each test has several strands (subtests), each with about five items, and it would not be desirable to delete more than one item (20%) for each subtest.

Extension of Methodology Based on Robust z to 3PL Models

Let a_1 and b_1 be the slope and location of a given item in the 3PL model for one group of students (first calibration). Let a_2 and b_2 be the same parameters for another group of students (second calibration). Linking the second set of item parameters to the first set requires determination of two constants A and B that set the following two transformations:

$$a_2 = Aa_1, \text{ and}$$

$$b_2 = \frac{(b_1 - B)}{A}.$$

Note that these transformation equations do not involve the pseudo-chance ("c") item parameter. When the two sets of calibrated item parameters are perfectly linked, the two constants A and B are identical for all linkers. However, due to sampling variations, the

equations usually do not hold for all these items. So a statistical process has to be used to find to constants A and B that fit (in some statistical sense) the calibrated parameters of all items. Once the fitting constants A and B are found, then can be used to find the estimated parameters of items what are not part of the linking items.

There are a variety of methods to fit the linking constants A and B to the calibrated item data of the linkers. Among them are the Mean-Mean, Mean-Sigma, Haebra, and Stocking-Lord methods (Haebara, 1980; Loyd & Hoover, 1980; Marco, 1977; Stocking & Lord, 1983). A new method that is based on robust z_R statistics for both the "a" (slope) and "b" (location) parameters is described next.

The general process of the robust z_R method is described as follows. Equation 1 can be written as: $\log(A) = \log(a_2) - \log(a_1)$. By writing $R_1 = \log(a_1)$ and $R_2 = \log(a_2)$, it can be seen that the robust z_R procedure described for the Rasch model can be extended to identify items that are unstable in the "log (a)" (slope) parameter and needs to be deleted in the linking process. The value of $\log(A)$ can then be taken as the mean of the discrepancy $\log(a_2) - \log(a_1)$ of all surviving linkers. The value of A will then be computed via the formula $A = \exp[\log(A)]$. Applying this A value to Equation 2 for all linkers that have survived so far, Equation 2 now becomes $B = Ab_2 - b_1$. By writing $R_1 = b_1$ and $R_2 = Ab_2$, it can be seen again that the robust z_R procedure described for the Rasch model can be extended to identify items that are unstable in the "b" (location) parameter and needs to be deleted in the linking process. Those that are stable in location may then be used to find the linking constant B. This constant is the mean of the discrepancy $(Ab_2 - b_1)$ taking over all surviving linking items.

As noted at the beginning of this section of the paper, the linking constants A and B are defined theoretically using the slope ("a") and location ("b") item parameters. The robust z_R does not rely on the pseudo-chance item parameter ("c"), and conduct the data analysis sequentially, first with the slope ("a") and then follow this up with the location ("b") parameter. It may be noted that the "c" parameters are hard to estimate, especially when the sample size is small. In fact, in a number of situations, the "c" parameter has to be fixed at a certain value or certain range to allow to calibration process to converge. So it seems to make sense to set aside the "c" parameter in fitting the linking constants A and B to the item data. It may also be noted that there are more sampling fluctuations for the slope ("a") than for the location ("b") estimates. This seems to justify the use of the robust z_R for the slope parameter first.

Extension to Mixed-Format 2PPC Items

Although this expression was developed for 3PL binary (multiple-choice) items, it can also be applied to partial credit items that follow a two-parameter partial credit (2PPC) model (Muraki, 1992, 1993). It may be noted that a 2PPC item with k score categories can be explicitly defined as (k-1) conditional 2PL binary items. The j-th conditional binary item is the "item" that has only (j-1) and j as scores. (see Masters, 1982, for a formulation of conditional items in the Rasch model.) All conditional binary items have identical slope. The location parameter of each conditional item is often called "threshold" parameters in IRT software such as PARSCALE (Muraki & Bock, 1997).

As for a 3PL model, for a set of 2PPC linking items, the robust z_R procedure starts with detecting items that are unstable along the slope dimension and need to be set aside in the linking process. Then the robust z_R procedure is applied to detect any threshold that is unstable along the location dimension and needs to be deleted. Theoretically only thresholds that are unstable need to be deleted from the calibration process. However, operational testing programs such as the PACT of South Carolina delete an entire item even if it has only one unstable threshold. It seems justified in the sense that one threshold is only one part of the item; therefore if it is "unstable" then the entire item should be considered as "unstable."

For the illustrative purposes of this paper, the software PARSCALE was used to calibrate the test for each group, setting the prior distribution to be the unit normal distribution for each group. Mean/mean, mean/sigma, Haebara, and Stocking-Lord linking constants, A and B, were obtained via the software STUIRT (Kim & Kolen, 2004) using item parameter estimates from PARSCALE. Equating constants obtained from the robust z_R method were compared to those obtained by the mean/mean, mean/sigma, Haebara, and Stocking-Lord procedures.

An Illustration

The performance of the robust z procedure will be illustrated using a set of archival data from a large-scale state assessment program. The data came from the

administration of a math test to 5th grade students in 2006. The math test has 40 multiple-choice items and two constructed response items (with four score categories 0, 1, 2, and 3). Group 1 (N = 4045) is comprised of one-third of all male students who had free or reduced priced lunch. Group 2 is comprised of one-third of all female students who had to pay their lunches (N = 3692). As can be seen from Table 1, these two groups differ considerably in terms of ability.

Table 1: Descriptive Statistics

Student Group	N	Mean	SD	Alpha
Group 1	4045	21.09	8.22	.86
Group 2	3692	28.67	8.30	.87

PARSCALE was used to estimate the item parameters for each group separately. Each calibration set the mean at zero and the standard deviation at one. The items parameters (a_1, a_2; b_1, b_2; and c_1, c_2) for Group1 and Group 2 are reported in the Appendix A. Items with ID from 1 to 40 are for the 40 multiple-choice items. Data listed for the codes 41A, 41B, and 41C are the common slope and threshold parameter for the three non-zero scores 1, 2, and 3 of the first CR items. Similarly, data listed for the codes 42A, 42B, and 42C are the common slope and threshold parameter for the three non-zero scores 1, 2, and 3 of the second CR items.

The first set of robust z statistics were first computed for the difference $A_{diff} = \log(a_1) - \log(a_2)$. Using the cut score 1.96 for the robust z_R, two items (ID = 26 with zR = 4.261; and 38 with z_R = -2.88) were found to be "unstable" along the slope dimension. Based on the remaining 44 data items, the linking constant (on the log scale) is -0.19609, which is transformed back to A = 0.822 on the original slope scale.

The second set of robust z_R statistics was computed for the difference $B_{diff} = Ab_2 - b_1$ using the 44 "stable" items. Using also the cut score 1.96 for the robust z_R, six items were found to be "unstable" along the location dimension. They are listed as follows: ID = 17 (z_R = 2.624); ID = 21 (z_R = -2.37); ID = 28 (z_R = -2.58); ID = 33 (z_R = 2.399); ID = 35 (z_R = 3.924); and ID = 44 (z_R = 1.987). Based on the remaining "stable" items, the linking constant was found to be B = 1.072.

Appendix B lists a sample SAS program for the robust z analysis.

For illustration purposes, the software STUIRT was also used to find the linking constants A and B for the same set of data. Table 2 reports the results of various linking methods.

Table 2: Results of six linking processes

Method	A	B
Mean/Mean	0.823	1.165
Mean/Sigma	0.770	1.130
Haebara	0.802	1.036
Stocking-Lord	0.823	1.078
Robust z	0.822	1.072

It is interesting to see that the robust z_R results are almost identical to those obtained from the Stocking-Lord procedure, yet the robust z_R method is computationally much simpler than the Stocking-Lord procedure. This finding applies only to this data set and should be considered as suggestive of an interesting possibility that needs further research.

Educational Importance

IRT models are widely used linking educational assessment. The robust z_R statistic has proved useful in detecting unstable items in the Rasch model. The statistic is intuitively appealing and simple to use. This paper extends the robust z_R to the 3PL binary items and 2PPC partial credit items. An illustration based on real data indicates that the linking constants A and B obtained from the robust z_R procedure are strikingly similar to those obtained from the well-known Stocking-Lord procedure. This observation applies only to this data set and should be considered as suggestive of an interesting possibility that needs further research.

References

Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences*, 4th Ed. Upper Saddle River, NJ: Pearson-Prentice Hall.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144-149.

Hogg, R. V. (1979). Statistical robustness: One view on its use in applications today. *The American Statisticians, 33*, 108-115.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*, 73-101.

Huynh, H. (1982). A comparison of four approaches to robust regression. *Psychological Bulletin*, *92*, 505-512.

Huynh, H. (2000, June). *Guidelines for Rasch Linking for PACT*. Memorandum to Paul Sandifer on June 18, 2000. Columbia, SC: Available from Author.

Huynh, H. (March 22, 2009). *The Golden Numbers in Rasch Linking Protocol.* Personal Communication to Technical Advisory Committee on Rasch Linking Protocols.

Huynh, H, & Rawls, A. (2009). A comparison between robust z and 0.3-logit difference procedures in assessing stability of linking items for the Rasch model. In Everett V. Smith Jr. & Greg E. Stone (Eds.) *Applications of Rasch Measurement in Criterion-Referenced Testing: Practice Analysis to Score Reporting*. Maple Grove, MN: JAM Press.

Kim, D. H., & Huynh, H. (2009). Transition from paper-and-pencil to computer-based testing: examining stability of Rasch latent trait across gender and ethnicity. In Everett V. Smith Jr. & Greg E. Stone (Eds.) *Applications of Rasch Measurement in Criterion-Referenced Testing: Practice Analysis to Score Reporting*. Maple Grove, MN: JAM Press.

Kim, S., & Kolen, M. J. (2004). STUIRT: A computer program for scale transformation under unidimensional item response theory models [Computer software and manual]. Retrieved from http://www.education.uiowa.edu/casma/computer_programs.htm#irt.

Linacre, J., M. (2006). *A user's guide to WINSTEPS/MINISTEP Rasch-model computer programs.* Chicago: www.winsteps.com.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(*2*), 139-160.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(*2*), 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(*2*), 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 14*(*4*), 351-363.

Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating scale data* [Computer software]. Chicago: Scientific Software.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, 52, 275-291.

APPENDIX A

Item Calibration Data for Group 1 and Group 2

ID	a1	a2	b1	b2	c1	c2
1	0.729	0.650	1.585	0.676	0.134	0.110
2	0.846	0.782	0.635	-0.525	0.304	0.316
3	0.909	0.816	-0.378	-1.749	0.267	0.161
4	0.818	0.787	-0.100	-1.092	0.176	0.149
5	0.742	0.611	-0.195	-1.619	0.215	0.145
6	0.890	0.888	0.749	-0.406	0.194	0.200
7	1.741	1.192	1.246	-0.132	0.267	0.243
8	0.907	0.589	1.016	0.006	0.159	0.059
9	1.487	1.211	-0.234	-1.352	0.095	0.081
10	1.228	0.742	0.537	-0.872	0.197	0.075
11	0.672	0.526	0.070	-1.242	0.089	0.028
12	1.007	0.690	1.985	0.873	0.272	0.267
13	1.016	0.996	1.101	0.239	0.229	0.242
14	0.776	0.816	-0.742	-2.038	0.159	0.189
15	0.921	0.781	0.463	-0.487	0.162	0.184
16	0.550	0.507	-0.060	-1.372	0.100	0.121
17	0.624	0.378	0.477	-1.492	0.259	0.000
18	0.984	0.976	1.084	0.214	0.167	0.170
19	0.506	0.473	-2.340	-4.537	0.000	0.000
20	0.594	0.364	1.068	0.220	0.242	0.151
21	0.687	0.585	-0.055	-0.686	0.323	0.383
22	0.541	0.566	-1.045	-2.394	0.000	0.000
23	0.691	0.511	1.859	0.747	0.196	0.195
24	0.843	0.718	0.645	-0.467	0.189	0.177
25	0.530	0.354	-0.689	-3.629	0.000	0.000
26	0.462	1.080	-2.583	-5.000	0.000	0.000
27	1.007	0.840	1.922	0.927	0.334	0.352
28	0.825	0.865	0.709	0.305	0.538	0.647
29	0.608	0.528	0.499	-0.839	0.125	0.116
30	1.177	0.814	1.973	1.270	0.511	0.501
31	0.900	0.555	0.104	-1.618	0.192	0.000
32	0.861	0.701	0.809	-0.091	0.353	0.286
33	0.843	0.530	0.640	-1.228	0.103	0.000
34	1.404	1.220	0.247	-1.019	0.241	0.248
35	0.446	0.344	0.820	-1.453	0.245	0.064
36	1.014	0.966	1.837	1.090	0.118	0.150
37	1.632	1.044	2.129	1.743	0.155	0.126
38	0.831	0.358	1.012	-1.436	0.132	0.000
39	1.560	1.192	1.774	1.024	0.215	0.187
40	0.798	0.615	0.095	-1.358	0.148	0.007
41A	0.561	0.486	0.784	-0.539	0	0
41B	0.561	0.486	-0.113	-1.489	0	0
41C	0.561	0.486	1.166	-0.052	0	0
42A	0.745	0.737	3.687	2.599	0	0
42B	0.745	0.737	2.506	1.250	0	0
42C	0.745	0.737	-0.001	-1.209	0	0

APPENDIX B

Sample SAS Program for Robust z Analysis for Slope and Location Parameters

```
%let CV = 1.96;
Data raw;
Input   ID  $ a1  a2  b1  b2 ;
a1log = log(a1);
a2log = log(a2);
Adiff = a2log-a1log;
Kode  = 1;
Cards;

* Datelines are here;
;

proc univariate data=raw noprint ; var Adiff; output out=summary Median= MD Mean=Mean
Qrange=IQR;
data new1; set summary; kode=1;
data new2; merge raw new1; by kode;
data new3; set new2; zrobust=(Adiff-MD)/(0.74*IQR);
code ="A-unstable";
if abs(zrobust) le &CV then code ="A-stable";
proc print data=new3 noobs; var ID a1 a2 a1log a2log Adiff zrobust code;
format a1 a2 a1log a2log Adiff zrobust 5.3;
run;

data new4; set new3; if code="A-stable";
proc means data=new4 noprint; var Adiff; output out=LC1 mean=AlogCon;
proc print data=LC1; var AlogCon;

data LC2; set LC1; Kode=1;
Acon=exp(AlogCon);
proc print data=LC2; var Acon;

data new5; merge new4 LC2; by Kode;
data new6; set new5; Bdiff=b1-Acon*b2;
keep ID Kode a1 a2 b1 b2 Bdiff;
proc univariate data=new6 noprint ; var Bdiff; output out=summary2 Median= MD Mean=Mean
Qrange=IQR;
data summary3; set summary2; kode=1;

data new8; merge new6 summary3; by kode;
data new9; set new8; zrobust=(Bdiff-MD)/(0.74*IQR);
code ="B-unstable";
if abs(zrobust) le &CV then code ="B-stable";
proc print data=new9 noobs; var ID b1 b2 Bdiff zrobust code;
format b1 b2 Bdiff zrobust 5.3;
data new10; set new9; if code="B-stable";
proc means data=new10; var Bdiff;
run;
```

Citation

Huynh, Huynh & Meyer, Patrick (2010). Use of Robust z in Detecting Unstable Items in Item Response Theory Models. *Practical Assessment, Research & Evaluation*, 15(2). Available online: http://pareonline.net/getvn.asp?v=15&n=2.

Authors

Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208.
Email: hhuynh [at] mailbox.sc.edu.

J. Patrick Meyer
Curry School of Education
University of Virginia
P.O. Box 400277
405 Emmet Street South
Charlottesville, VA 22904.
Email: meyerjp [at] virginia.edu