Fall 10-23-2008

A Confirmatory Factor Analytic Study Examining the Dimensionality of Educational Achievement Tests

Nina Deng
University of Massachusetts at Amherst, ndeng@educ.umass.edu

Craig Wells
University of Massachusetts at Amherst, cswells@educ.umass.edu

Ronald Hambleton
University of Massachusetts at Amherst, rkh@educ.umass.edu

Running head: EXAMINING TEST DIMENSIONALITY

A Confirmatory Factor Analytic Study

Examining the Dimensionality of Educational Achievement Tests

Nina Deng

Craig S. Wells

Ronald K. Hambleton

University of Massachusetts at Amherst

The paper was presented at the meeting of the NERA, Rocky Hill, Connecticut, October, 2008.

Abstract

Along with the increasing popularity of item response theory (IRT) in testing practices, it is important to check a fundamental assumption of most of the popular IRT models, which is unidimensionality. Nevertheless, it is hard for educational and psychological tests to be strictly unidimensional. The tests studied in this paper are from a standardized high-stake testing program. They feature potential multidimensionality by presenting various item types and item sets. Confirmatory factor analyses with one-factor and bifactor models, and based on both linear structural equation modeling approach and nonlinear IRT approach were conducted. The competing models were compared and the implications of the bifactor model for checking essential unidimensionality were discussed.

A Confirmatory Factor Analytic Study

Examining the Dimensionality of Educational Achievement Tests

Item response theory (IRT) and its applications has been widely used in various educational and psychological testing practices, including test construction, ability estimation, score reporting, equating, and computer adaptive testing. One of the fundamental assumptions for the most popular IRT models is unidimensionality (Hambleton, Swaminathan, & Rogers, 1991), namely only one single latent variable accounts for the item responses. Technically, it can be defined that item responses are independent when a single latent variable is controlled for. To justify the uses of IRT, unidimensionality needs to be demonstrated before any unidimensional IRT modeling is applied. Applying IRT models without the justification of unidimensionality may greatly undermine the advantages of IRT and result in negative consequences. However, it is more often than not that the educational tests are not strictly unidimensional. This could be due to factors such as content variety, construct complexity, and varying item formats. For most tests, a dominant factor rather than only one factor is the goal. The dimensionality analysis described in this paper is therefore to check whether the item responses are essentially unidimensional or "unidimensional enough" for proper IRT applications.

Through years of research, a number of different methods have been developed to assess test dimensionality. The main approaches include linear and non-linear factor analysis, nonparametric procedures, multidimensional scaling and structural equation modeling, which is a type of linear factor analysis. In view of the importance of essential unidimensionality, this paper adopted different methods and approaches, including the parallel analysis and bifactor analysis suggested in an earlier paper (see Deng & Hambleton, 2007) and applied to an educational achievement testing program. These tests feature potential multidimensionality by presenting

various item types and item sets. The purposes of the study were to (1) assess the test dimensionality in both exploratory and confirmatory ways, (2) check the item type effect and check essential unidimensionality, and (3) apply parallel analysis and bifactor analysis methods, and discuss their implications.

Methods

Data

Four data sets consisting of two achievement tests and two years (2007-2006) per test were assessed in this study. In Test 1, there were 85 items consisting of 80 multiple-choice (MC) items and 5 free response (FR) items. All MC items were discrete and independent from each other (i.e., no testlets), and five options for each item. The FR items have some missing values by design. The first FR item was mandatory for each examinee. Then the examinees are required to choose a second item to answer from the second and third FR items, and a third item to answer from the forth and the fifth items. Thus each examinee answered three FR items, leaving the other two blank.

In Test 2, there were 98 MC items in 2006, 50 were discrete and 48 were MC set members, that is, a member of a item set consisting of 3 or 4 MC items which share a common stimulus. Similarly, there were 99 MC items in 2007, 49 discrete and 50 MC set members. For the FR items, there were 4 FR items and all were mandatory. The sample size for each of the original datasets was 20,000.

To deal with the missing and not-reached data, some rules were applied prior to the dimensionality analyses. For Test 1, only examinees answering all of the three FR items were retained for the analysis. Since there are unoverlapping missing values due to the test design, the missing responses in FR were replaced by the average across the non-missing FR responses.

Keeping in mind that the purpose was to assess the effect of possible item type factor, this method was chosen to minimize the effects of MC responses on the FR items. For Test 2, the examinees answering none of the four FR items were dropped from the data. And the missing values in FR items in Test 2 were treated as zero. For MC items in both tests, all the missing and unreached values were replaced randomly with a probability of 20% of being 1, and a probability of 80% in being 0. This replacement came from the idea that the examinees have a probability of 20% in guessing the answer right, given five options for each MC item. After recoding, the sample sizes drop a very small amount and varied from 19151 to 19999 for the four tests. Our view was that none of the decisions about handling missing data would impact to any substantial extent on the final results.

Methods for Assessing Dimensionality

Parallel Analysis. Principle component analysis (PCA) together with eigenvalue plots (see Lord & Novick, 1968) is a common way to evaluate test dimensionality and has been used for decades (see Hattie, 1985). The percentage of total variance explained by the first principle component is often regarded as an index of unidimensionality. The higher percentage of total variance the first principle component accounts for, the closer the test is to unidimensionality. One downside with eigenvalue plots is that there is no statistical index available to decide the number of underlying dimensions. Various criteria have been proposed to solve the problem. Reckase (1979) recommended that a percentage of 20% or more of the total variance explained by the first principle component is necessary for the data to be viewed as unidimensional. Lord (1980) suggested checking the ratio of the first to the second eigenvalue, and compare that with the ratio of the second to any of the other eigenvalues. Kaiser (1970) suggested retaining any components with eigenvalues larger than 1. Drasgow and Lissak (1983) introduced parallel

analysis with baseline plots in the hope of deciding the number of dimensions in a more meaningful way. In the parallel analysis, the eigenvalue plot of actual data is compared with the baseline plot from an inter-item correlation matrix of the random data, which are generated from uncorrelated variables. If the test data are unidimensional, the eigenvalue plot and the baseline plot should look similar except that the first eigenvalue of the real data is much bigger than the first eigenvalue of the random data. The remaining eigenvalues should be close since they are expected from random errors.

The parallel analysis with baseline plots has been very helpful in interpreting test dimensionality (Hambleton, Swaminathan, & Rogers, 1991) but it is not so well known or widely used. It was our decision in this paper to conduct the parallel analysis. In this study, parallel analysis was conducted by computing the eigenvlaues of the random data, which were generated from the matrix with the same sample size, the same number of variables, and the same p-values of each item as that in the real data. Then they were compared with the eigenvalues of the real data. The eigenvalues of real data which are substantially larger than the first eigenvalue of the random data indicate possible secondary dimensions.

*Bifactor Analysis***.** As mentioned earlier, it is common that many educational and psychological assessments are not strictly unidimensional. The reasons could be due to different item formats, various item contents, and complex underlying constructs. Nevertheless, the items are usually designed to be accounted by one main factor, which is the latent trait the test is intended to measure. Since most popular IRT models requires the fundamental assumption of unidimensionality, the very practical and compelling issue in IRT applications is not to assess whether a test is strictly unidimensional, but rather whether the test is "unidimensional enough" to be properly analyzed using IRT models. If we can confirm a test is essentially unidimensional

in spite of potentially small factors, it will greatly help assure the proper usage of IRT models to these tests, and thus avoid negative consequences when applying unidimensional IRT models to potential multidimensional data.

The bifactor model was proposed (Holzinger & Swineford,1937; Gibbons & Hedeker,1992; Reise et al., 2007) with the intention of checking the essential unidimensionality and the possible consequences when applying unidimensional models to multidimensional data. In a bifactor model, which is different from traditional one-factor or multi-factor models, each item has factor loadings on two and only two factors on the same level: one general factor, and one group factor. The general factor stands for the latent trait measured in the test and explains the covariance shared by all the items. The group factor stands for the secondary factors possibly existing in particular items, and account for the covariance independent of the general factor. The general factor and group factors are uncorrelated and account for the covariance simultaneously and independently for each item. The idea of the bifactor model is to look at how much of the total covariance explained by the group factors compared to that explained by the general factor. If most of the covariance is explained by the general factor and only small amount is remaining for the group factors, and if the factor loadings on the general factor in the bifactor model are comparable with that in the one-factor model, the test is regarded as essentially unidimensional, or "unidimensional enough", since the group factors account for little of the covariance compared to the general factor, and have a small impact on the factor loadings. The bifactor model is especially useful when the data feature a complex structure.

The bifactor model was applied to the data using two approaches in this study, one with structural equation modeling (SEM) and one with a nonlinear IRT-based approach. The approach of SEM was carried out in LISREL8.8 (Jöreskog & Sörbom, 1981, 1999). Polychoric and

tetrachoric correlation matrices along with asymptotic covariance matrices were calculated for the dichotomous and polytomous data using PRELIS 2.8, which is embedded in the LISREL program. Weighted-Least Squares method (WLS) was used to estimate the model. Both one-factor and bifactor models were conducted and the results were compared.

Full-information Item Bifactor Analysis. The other approach based on IRT models is called full-information (FI) item bifactor analysis. It was first proposed for dichotomous data (Bock, Gibbons & Muraki, 1988; Gibbons & Hedeker, 1992) and later extended for graded response data (Gibbons et al., 2007). Adopting the idea of the bifactor model, FI item bifactor analysis uses the non-linear IRT models and marginal maximum likelihood (MML) estimation method (Bock & Aitkin, 1981). Output includes item parameters (e.g. factor loadings, thresholds) and standard errors, and likelihood chi-square statistics are provided too. The number of identifiable factors can be statistically determined by testing the difference between the likelihood ratio chi-square values of competing models. FI item bifactor analysis was chosen because it has the advantages over the linear factor analysis in that the nonlinear regression is more suitable for the dichotomous and ordinal data, and it has less problems of model identification in parameter estimation.

For this paper, the software package POLYBIF (Gibbons & Hedeker, 2007) was used. It can be freely downloaded from http://www.uic.edu/labs/biostat/bifactor.html. There are, however, some restrictions applied to this program. It can handle up to only 5000 examinees, 100 items, and 10 categories for free-response (FR) items. Therefore, 5000 examinees were randomly sampled from the original data sets, and the first few MC items were deleted to make the number of items within 100, and lastly, the scored category of 10 for FR items was collapsed with the 9. The scores at the bottom scale were collapsed because the numbers of students in category of 9

and 10 were the fewest compared to the other categories, ranging from 0.1%-5.7% of the total number of students across all FR items.

Results

Parallel Analysis

Eignevalue plots of the data, together with baseline plots of generated random data, were produced and compared in Figure 1 to Figure 4 for each data set. Table 1 to Table 4 display the largest ten eigenvalues and the percentage of variance explained for each data set. Throughout the tests, the first eigenvalues explained over 20% of the total variance, and were 8 to 9 times larger than the second eigenvalues, which is well beyond what is often used as the bench mark: 4 or 5 to 1. This suggested a dominant factor for these tests. Furthermore, besides for the first eigenvalue, there were 2 to 4 eigenvalues which were also larger than the first eigenvalue of the random data, and 1 to 2 of them were substantially larger. This indicated that one or two small secondary factors may exist. Nevertheless, the secondary factors looked small compared to the biggest factor. And the plots showed strong unidimensionality.

Exploratory Factor Analysis

Exploratory factor analysis (EFA) was carried out with the Mplus software (Muthén & Muthén, 1998-2004) using the weighted least square method (WLSM) as a preliminary analysis. Based on the results of PCA and parallel analysis, which indicated a strong dominant factor, one-factor model was adopted. The model fit statistics are summarized in Table 5. The fit statistics from EFA showed good fit of the one-factor model to the data. CFI and TLI were based on the ratio of the explained variance to the observed variance. A value greater than 0.95 indicates a good fit.

RMSEA and *SRMR* were based on residuals, or the discrepancies between the implied and observed covariance matrices. A smaller value indicates good fit. The *RMSEA* was less than 0.05 and the SRMR was less than 0.08, which was also interpreted as acceptable fit to the data.

Confirmatory Factor Analysis

Since EFA is only a preliminary and exploratory step in dimensionality analysis, a confirmatory analysis is often desired. In confirmatory factor analysis (CFA), different from EFA in which we do not assign observed variables (items in this study) to any particular factor, we define specifically which item is explained by which factor, based on our hypothesis or theory. Therefore we have a much clearer hypothesis compared to EFA and this hypothesis can be tested in a confirmatory way.

Traditional one-factor model. Both Mplus and LISREL were used to assess the data with one-factor model. Their statistics were compared in Table 6 and Table 7. Interestingly, the fit index CFI was higher in Mplus than in LISREL, and RMSEA was lower in LISREL than in Mplus. Although both programs assumed to use the same method called Weighted Least Square (WLS), apparently they adopted different estimation algorithms, since they produced different values for the indices. The fit statistics were worse compared to EFA, although Test 1 did not exhibit poor fit and it fit apparently better than Test 2. It appeared that the first factor in Test 1 was more dominant than in Test 2.

Bifactor model. The results from applying the one-factor model indicated that there were secondary factors existing in the tests besides the dominant factor. Some previous studies also suggest that item type is one of the most common unintended factors in educational tests. Therefore the bifactor model was subsequently used to assess the effect of item type factor on the dominant factor. In the bifactor model, all the items were assigned to the dominant factor, called

the general factor, in addition, all the MC items were assigned to the first group factor called MC factor, and all the FR items were assigned to the second group factor called the FR factor. Each item had loadings on two factors simultaneously, one general factor, and one MC or FR factor.

In addition to the item type effect, we found that Test 2 had a poorer fit with the one-factor model than Test 1. Since Test 2 not only has discrete MC items but also MC set items (i.e., items that share a common stimulus), the errors between items within one set may not be viewed as independent. Therefore, to improve the model fit, the errors of the MC set items within one set in Test 2 were specified as correlated. Bifactor model (with errors correlated for MC sets items in Test 2) was conducted using LISREL (Mplus was tried too but had convergence problems using WLSM). Results showed that the new models fit much better than the former ones. The selected summary statistics provided by the program are shown in Table 8.

The fit with Test 1 was improved with CFI and NFI larger than 0.95 and RMSEA and SRMR lower than 0.05. The fit with Test 2 would be considered adequate, with CFI and NFI larger than 0.9, RMSEA lower than 0.05 and SRMR lower than 0.08. Overall, the new models had greatly improved fit.

The factor loadings of the bifactor model were compared with that of the one-factor model (see Table 9 for an example of the factor loadings in Test 1). The loadings on the general factor in the bifactor model were comparable with that of the one-factor model, and the loadings on the group factors in bifactor model were comparatively small and not meaningful. This indicates that most of the covariance in the test was explained by the general factor rather than the group factors. Again the comparability between the loadings on the general factor in the bifactor model and that in the one-factor model suggested that the group factors had small effects on the general factor.

Full-Information Bifactor Analysis

Similarly, as the confirmatory factor analysis using SEM, both the one-factor and bifactor models were tried using full-information bifactor analysis. This method was chosen because it is particularly suitable for dichotomous and polytomous data and has shown efficiency in assessing test dimensionality in recently studies (Immekus & Imbrie, 2008). The statistics provided by the program are displayed in Table 10. Two times the differences between log-likelihood statistics (LL) resembles a chi-square distribution, and the degrees of freedom for these chi-square statistics is the difference of the numbers of parameters estimated in the competing models. By comparing the differences of -2LL between the two models, the significance of model fit improvement can be statistically tested. The results showed that the bifactor model had significantly improved the model fit over the one-factor model across all the four tests. Further checking the factor loadings for each item (see Table 11 for an example of Test 1 in 2007) provided compelling evidence that in the bifactor model, most of the covariances were explained by the general factor. This finding was consistent with the linear confirmatory factor analysis results discussed earlier in the paper. One point may be worthy of notice is that the improvement of the bifactor model is more substantial in Test 1 than in Test 2, which is not seen in the SEM results. This might be due to the fact that Test 2 has MC set items therefore the local independence may not be preserved. In POLYBIF, the FI bifactor analysis was based on Samejima's graded response model (1997), which should not have taken the related errors into account, whereas in SEM, the correlated errors were considered and thus the model had a better fit for Test 2 in SEM.

Conclusions and Discussions

A variety of methods and approaches were tried in this research assessing the dimensionality of a set of educational tests which feature potential multidimensionality by presenting various item types, formats and item sets. The purpose of the study was to check whether the tests present multidimensionality by having various item types, and whether the tests are essentially unidimensional so that they can be appropriately analyzed using IRT models. In addition to the common practices in principle component analysis, exploratory factor analysis and confirmatory factor analysis in SEM, two not widely known approaches, parallel analysis and bifactor analysis, were suggested and applied. The methods applications in dimensionality analysis were checked and the results were discussed. Several conclusions and suggestions can be drawn:

1. Although almost unavoidably, the educational and psychological tests are not strictly unidimensional, the essential unidimensionality is suggested being checked prior to unidimensional IRT analysis. In this paper, through various analyses the four tests presented essential unidimensionality by presenting a strong first factor. Test 1 appeared to have stronger unidimensionality than Test 2 but both tests appeared strongly unidimensional.
2. Bifactor analysis is suggested as a good and useful way in checking essential unidimensionality. In this study, item type effect was examined since it is often the case that it shows up as a secondary factor in educational tests. The results of bifactor analysis show that the item type does not seem to post a great impact on the main factor. Most of the covariance is explained by the main factor, and the loadings on the main factor were not changed much due to the secondary item type factors. Nevertheless, Test 1 did show

a stronger item type effect than Test 2 by having larger loadings on the group factor in bifactor model. In addition to bifactor analysis, parallel analysis was found helpful in explaining the eigenvalues. However, it was suggested that having a number of repetitions in generating the random data, so that the comparison between the real data and the random data can be more accurate.

3. Although all the tests showed a sign of essential unidimensionality, the degree varied among the tests. Test 2 definitely exhibited a more complex structure than Test 1 and presented more secondary factors. Firstly, there was a testlet effect in Test 2. Since some MC items share a common stimulus, their responses may be correlated even after controlling for the construct of interest. Secondly, the test length was longer (85 items in Test 1 vs. 102/103 items in Test 2). The factor of speededness effect is possible. This can be seen from the comparatively larger group factor loadings with the bifactor model for the last few MC items than the other MC items. Thirdly, the free response items in Test 2 have two formats, one with text only and another with art format. This may result in another nuisance factor. All these potential nuisance factors deserve further research.
4. By applying bifactor analysis, both Test 1 and Test 2 showed a dominant factor in spite of secondary factors. Nevertheless, it is recommended for Test 2 to use IRT models taking account of the testlet effect in future analysis so that the local dependence with MC set items could occur. The testlet effect was proven to greatly and negatively affect the model fit if the model did not consider the correlation between the MC items with a common stimulus.
5. The tests feature other complex issues such as missing data, various item content, etc., which may affect the test dimensionality as well, and deserve further analysis. Besides,

the methods used in this study would be desirable to be repeated using a cross-validation study by dividing the data into two groups. Sample sizes in this study were more than sufficient to consider conducting a cross-validation study.

In conclusion, unidimensionality is a key and fundamental assumption that needs to be checked prior to various IRT applications. It is suggested in this paper that various methods and statistics are used when assessing test dimensionality, and essential unidimensionality is checked for complex tests which feature potential multidimensionality. Empirical studies were conducted using parallel analysis and bifactor analysis, both of which were shown useful in checking the essential unidimensionality. Although all the tests showed a dominant first factor in spite of the potential item type effect, some tests showed stronger unidimensionality whereas others showed weaker ones due to the possible item type effect, testlet effect and the speededness effect.

References

Bock, R. D., Gibbons, R., & Muraki, E. J. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Deng, N., & Hambleton, R. K.. (2007, October). *Checking test dimensionality: New methods and applications*. Paper presented at the meeting of the Northeastern Educational Research Association, Rocky Hill, Connecticut.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*(3), 363-373.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D., et al. (2007). Full-information item bi-factor analysis of graded response data. *Applied Psychological Measurement, 31*, 4-19.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika, 57*, 423-426.

Gibbons, R. D., & Hedeker, D.(n.d.). *POLYBIF [Computer software]*. Chicago: Center for Health Statistics, University of Illinois. Retrieved January 22, 2008, from http://www.uic.edu/labs/biostat/projects.html.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41-54.

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using the full-information

item bifactor analysis for graded response data: An illustration with the state metacognitive inventory. *Educational and Psychological Measurement, 68*, 695-709.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood.* Chicago: National Educational Resources.

Jöreskog, K. G., & Sörbom, D. (1999) *LISREL 8: Structural Equation Modeling withthe SIMPLIS Command Language*. Lincolnwood: Scientific Software International.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*, 401-415.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. *Reading*. MA: Addison-Wesley.

Lord, F. M. (1980). *Applications of item response theory to practical testing programs*. Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, L. K., & Muthén, B. O. (1998-2004*). Mplus user's guide (3rd ed.)*. Los Angeles, CA: Muthén & Muthén.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*(Supplement 1), 19-31.

Samejima, F. (1997). Graded response model. In W.J. van der Linden & Ronald K. Hambleton (Eds), *Handbook of item response theory* (pp. 85-100). New York: Springer.

Table 1

Eigenvalues (largest ten) of Real and Random Data for Test 1, 2007

	Test 1, 2007		Random data	
Component	Eigenvalue	% of variance	Eigenvalue	% of variance
1	**18.99**	22.34	**1.49**	1.75
2	**1.89**	2.22	1.44	1.69
3	**1.61**	1.89	1.42	1.67
4	1.45	1.71	1.37	1.62
5	1.39	1.64	1.36	1.60
6	1.29	1.52	1.35	1.59
7	1.26	1.48	1.34	1.58
8	1.21	1.42	1.33	1.57
9	1.19	1.40	1.31	1.54
10	1.17	1.38	1.29	1.52

Table 2

Eigenvalues (largest ten) of Real and Random Data for Test 1, 2006

	Test 1, 2006		Random data	
Component	Eigenvalue	% of variance	Eigenvalue	% of variance
1	**17.76**	20.89	**1.46**	1.72
2	**1.86**	2.19	1.42	1.67
3	**1.60**	1.89	1.39	1.64
4	1.46	1.72	1.38	1.63
5	1.31	1.54	1.37	1.61
6	1.27	1.49	1.36	1.60
7	1.25	1.47	1.33	1.56
8	1.22	1.44	1.33	1.56
9	1.19	1.39	1.31	1.54
10	1.17	1.37	1.30	1.53

Table 3

Eigenvalues (largest ten) of Real and Random Data for Test 2, 2007

	Test 2, 2007		Random data	
Component	Eigenvalue	% of variance	Eigenvalue	% of variance
1	**25.47**	24.73	**1.53**	1.48
2	**3.19**	3.10	1.51	1.47
3	**2.05**	1.99	1.48	1.44
4	**1.87**	1.82	1.44	1.40
5	**1.65**	1.60	1.43	1.39
6	1.55	1.50	1.41	1.37
7	1.51	1.46	1.40	1.36
8	1.38	1.34	1.39	1.35
9	1.32	1.28	1.39	1.35
10	1.29	1.25	1.36	1.32

Table 4

Eigenvalues (largest ten) of Real and Random Data for Test 2, 2006

	Test 2, 2006		Random data	
Component	Eigenvalue	% of variance	Eigenvalue	% of variance
1	**24.91**	24.42	**1.52**	1.49
2	**2.66**	2.61	1.51	1.48
3	**1.92**	1.88	1.44	1.41
4	**1.75**	1.71	1.43	1.40
5	1.50	1.48	1.43	1.40
6	1.40	1.37	1.41	1.38
7	1.35	1.32	1.40	1.37
8	1.31	1.28	1.39	1.36
9	1.25	1.23	1.38	1.35
10	1.23	1.21	1.35	1.32

Table 5

Model Fit Statistics of the Exploratory Factor Analysis using Mplus

Statistic*	Test 2		Test 1	
	2007	2006	2007	2006
CFI	0.967	0.978	0.981	0.976
TLI	0.966	0.977	0.981	0.975
RMSEA	0.029	0.023	0.023	0.025
SRMR	0.042	0.036	0.034	0.035

Note. * CFI = Comparative Fix Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

Table 6

Model Fit Statistics of Confirmatory Factor Analysis with 1-factor Model using Mplus

Statistic*	Test 2		Test 1	
	2007	2006	2007	2006
CFI	0.891	0.926	0.945	0.934
TLI	0.966	0.977	0.981	0.975
RMSEA	0.029	0.023	0.023	0.025
WRMR	1.957	1.654	1.654	1.781

Note. * CFI = Comparative Fix Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

Table 7

Model Fit Statistics of Confirmatory Factor Analysis with 1-factor Model using LISREL

Statistics*	Test 2		Test 1	
	2007	2006	2007	2006
Chi-square (d.f.)	51173.62 (5150)	36737.72 (5049)	16795.31 (3485)	18142.41 (3485)
CFI	0.79	0.83	0.93	0.94
NFI	0.77	0.81	0.92	0.93
NNFI	0.78	0.82	0.93	0.94
GFI	0.98	0.98	0.99	0.99
RMSEA	0.021	0.018	0.014	0.015
SRMR	0.16	0.11	0.062	0.072

Note. * CFI = Comparative Fix Index; NFI = Normed Fit Index; NNFI = Non-Normed Fit Index; GFI = Goodness of Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.

Table 8

Summary Statistics of Bifactor Model using LISREL

Statistics[1]	Test 2 (errors correlated for MC set items)		Test 1	
	2007	2006	2007	2006
Chi-square[2] (d.f.)	19346.14 (4986)	18129.68 (4894)	10077.84 (3400)	10090.35 (3400)
CFI	0.93	0.93	0.97	0.97
NFI	0.91	0.90	0.95	0.96
NNFI	0.93	0.92	0.97	0.97
GFI	0.99	0.99	0.99	0.99
RMSEA	0.012	0.012	0.010	0.010
SRMR	0.058	0.070	0.038	0.033

Note. 2-factor model was conducted for Biology with errors of MC set item correlated. It turned out that 2-factor model had worse fit than bifactor model, with CFI of 0.91(vs. 0.93 in bifactor model), NFI of 0.89 (vs. 0.91), RMSEA of 0.014 (vs. 0.012), and SRMR of 0.072 (vs. 0.058).

Note. [1] CFI = Comparative Fix Index; NFI = Normed Fit Index; NNFI = Non-Normed Fit Index; GFI = Goodness of Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual.
[2] The improvement of Chi-square statistics of bifactor is significant ($p < .001$) over 1-factor model.

Table 9
Factor Loadings of 1-Factor and Bifactor Models for Test 1 2007 Using LISREL

Item	1-Factor model	Bi-factor model		
		General	MC	FR
VAR 1	0.46	0.45	0.11	
VAR 2	0.59	0.58	0.17	
VAR 3	0.56	0.55	-0.07	
VAR 4	0.52	0.52	0.05	
VAR 5	0.53	0.55	0.34	
VAR 6	0.37	0.37	0.09	
VAR 7	0.47	0.46	0.03	
VAR 8	0.54	0.54	0.07	
VAR 9	0.52	0.52	0.04	
VAR 10	0.24	0.23	-0.12	
VAR 11	0.69	0.68	0.09	
VAR 12	0.51	0.50	0.06	
VAR 13	0.61	0.60	0.14	
VAR 14	0.65	0.64	-0.01	
VAR 15	0.59	0.56	-0.22	
VAR 16	0.58	0.57	0.08	
VAR 17	0.35	0.32	-0.09	
VAR 18	0.66	0.65	0.00	
VAR 19	0.31	0.32	0.06	
VAR 20	0.36	0.35	-0.01	
VAR 21	0.43	0.40	-0.09	
VAR 22	0.56	0.55	0.04	
VAR 23	0.56	0.56	0.19	
VAR 24	0.39	0.39	0.11	
VAR 25	0.59	0.58	-0.10	
VAR 26	0.39	0.38	-0.04	
VAR 27	0.56	0.54	-0.09	
VAR 28	0.63	0.63	0.08	
VAR 29	0.58	0.56	-0.06	
VAR 30	0.65	0.64	0.08	
VAR 31	0.39	0.38	-0.06	
VAR 32	0.54	0.53	0.07	
VAR 33	0.44	0.43	0.04	
VAR 34	0.56	0.55	0.05	
VAR 35	0.45	0.43	-0.18	
VAR 36	0.27	0.25	0.04	
VAR 37	0.42	0.41	-0.04	
VAR 38	0.33	0.32	-0.07	
VAR 39	0.52	0.49	-0.13	
VAR 40	0.28	0.27	-0.06	
VAR 41	0.61	0.59	-0.08	
VAR 42	0.44	0.43	0.04	

VAR 43	0.41	0.39	-0.11	
VAR 44	0.65	0.63	-0.07	
VAR 45	0.35	0.33	-0.02	
VAR 46	0.65	0.64	-0.10	
VAR 47	0.38	0.36	-0.09	
VAR 48	0.49	0.47	0.16	
VAR 49	0.50	0.49	-0.08	
VAR 50	0.35	0.35	-0.04	
VAR 51	0.53	0.53	0.06	
VAR 52	0.42	0.41	0.07	
VAR 53	0.45	0.43	-0.08	
VAR 54	0.52	0.50	-0.14	
VAR 55	0.59	0.60	0.25	
VAR 56	0.50	0.49	0.04	
VAR 57	0.32	0.31	0.04	
VAR 58	0.49	0.47	-0.08	
VAR 59	0.32	0.32	-0.02	
VAR 60	0.34	0.33	-0.04	
VAR 61	0.66	0.65	0.13	
VAR 62	0.52	0.52	-0.17	
VAR 63	0.47	0.47	-0.08	
VAR 64	0.56	0.55	0.05	
VAR 65	0.37	0.36	-0.15	
VAR 66	0.53	0.53	-0.12	
VAR 67	0.49	0.49	-0.15	
VAR 68	0.53	0.50	-0.13	
VAR 69	0.35	0.32	-0.20	
VAR 70	0.57	0.56	-0.22	
VAR 71	0.44	0.43	-0.10	
VAR 72	0.39	0.36	0.10	
VAR 73	0.58	0.57	0.19	
VAR 74	0.49	0.48	0.00	
VAR 75	0.15	0.15	-0.11	
VAR 76	0.46	0.44	-0.21	
VAR 77	0.33	0.33	-0.15	
VAR 78	0.50	0.50	-0.14	
VAR 79	0.37	0.38	-0.19	
VAR 80	0.31	0.28	-0.11	
VAR 81	0.75	0.65		0.31
VAR 82	0.92	0.74		0.60
VAR 83	0.89	0.71		0.46
VAR 84	0.91	0.75		0.43
VAR 85	0.94	0.76		0.61

Note. The factor loadings for the other three tests are not reported here due to the space limit. They look similar to that of Test 1, 2007. Please write to the first author if you would like to see the complete set of results.

Table 10

Log-likelihood Statistics of 1-Factor and Bifactor Models using POLYBIF

Statistics*	Model	Test 2		Test 1	
		2007	2006	2007	2006
Log Likelihood (LL)	1-Factor	-301808.52	-302467.85	-265596.12	-267970.49
	Bifactor	-299609.78	-301305.83	-262241.81	-264037.62
-2ΔLL	(Bifactor vs. 1-Factor)	4397.47*	2324.04*	6708.62*	7865.74*
ΔD.F.		100	100	85	85

Note. * The chi-square improvement is significant.

Table 11

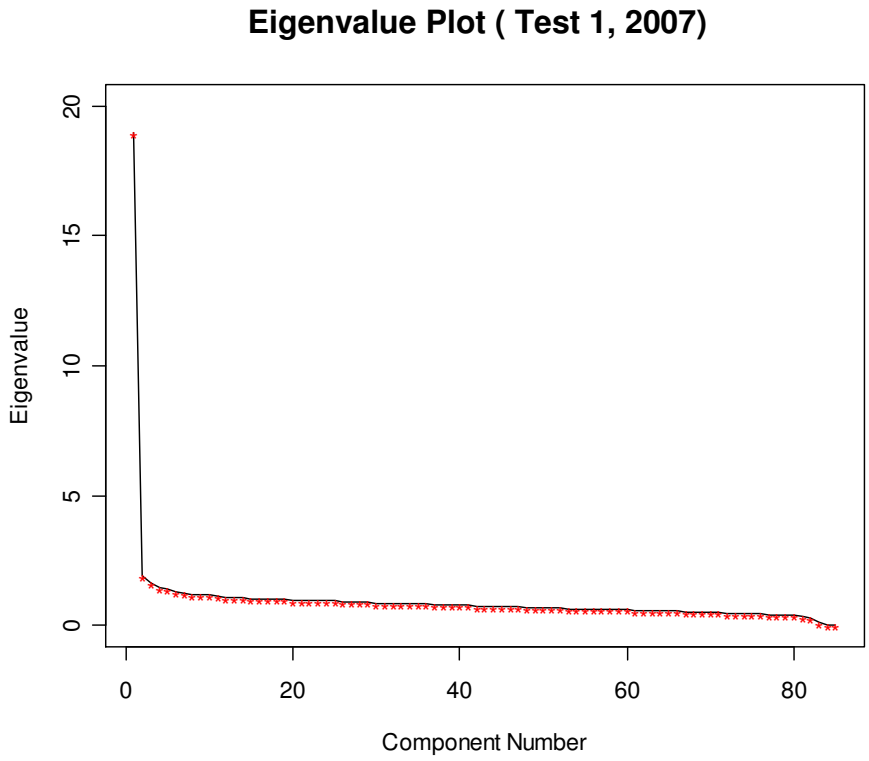
Factor Loadings of 1-Factor and Bifactor Models for Test 1 2007 Using POLYBIF

Item	1-Factor model	Bifactor model		
		General factor	MC	FR
1	0.36	0.44	-0.01	
2	0.44	0.53	-0.03	
3	0.42	0.46	0.13	
4	0.38	0.42	0.22	
5	0.40	0.53	-0.14	
6	0.25	0.28	0.10	
7	0.34	0.39	0.11	
8	0.41	0.43	0.26	
9	0.41	0.43	0.29	
10	0.17	0.19	0.12	
11	0.54	0.63	0.09	
12	0.38	0.41	0.20	
13	0.48	0.54	0.13	
14	0.52	0.57	0.22	
15	0.48	0.50	0.23	
16	0.42	0.47	0.16	
17	0.27	0.30	0.06	
18	0.52	0.60	0.11	
19	0.22	0.23	0.19	
20	0.27	0.31	0.17	
21	0.33	0.35	0.19	
22	0.43	0.45	0.27	
23	0.42	0.52	-0.04	
24	0.26	0.30	0.10	
25	0.47	0.51	0.23	
26	0.30	0.35	0.08	
27	0.42	0.46	0.17	
28	0.51	0.58	0.13	
29	0.40	0.42	0.26	
30	0.50	0.60	0.04	
31	0.29	0.32	0.13	
32	0.43	0.48	0.12	
33	0.33	0.36	0.14	
34	0.41	0.49	0.02	
35	0.34	0.35	0.26	
36	0.18	0.20	0.04	
37	0.31	0.34	0.17	
38	0.27	0.32	0.06	
39	0.41	0.44	0.20	
40	0.21	0.21	0.19	
41	0.47	0.51	0.23	

42	0.31	0.34	0.19	
43	0.31	0.35	0.12	
44	0.52	0.56	0.20	
45	0.26	0.29	0.12	
46	0.50	0.53	0.31	
47	0.28	0.32	0.09	
48	0.36	0.45	-0.07	
49	0.38	0.43	0.10	
50	0.25	0.25	0.22	
51	0.39	0.49	-0.03	
52	0.30	0.35	0.08	
53	0.31	0.34	0.22	
54	0.40	0.42	0.22	
55	0.41	0.48	0.15	
56	0.36	0.44	0.02	
57	0.22	0.25	0.05	
58	0.38	0.40	0.19	
59	0.23	0.24	0.12	
60	0.26	0.33	0.02	
61	0.51	0.57	0.16	
62	0.39	0.38	0.35	
63	0.37	0.38	0.29	
64	0.43	0.51	0.06	
65	0.26	0.28	0.16	
66	0.41	0.45	0.18	
67	0.36	0.38	0.23	
68	0.38	0.42	0.13	
69	0.24	0.27	0.12	
70	0.45	0.45	0.32	
71	0.34	0.35	0.23	
72	0.26	0.33	-0.09	
73	0.44	0.54	-0.07	
74	0.35	0.40	0.11	
75	0.11	0.13	0.08	
76	0.36	0.36	0.28	
77	0.26	0.26	0.25	
78	0.38	0.38	0.33	
79	0.27	0.28	0.24	
80	0.22	0.25	0.08	
81	0.60	0.58		0.44
82	0.72	0.64		0.65
83	0.68	0.63		0.57
84	0.74	0.67		0.55
85	0.75	0.68		0.62

Note. The factor loadings for the other three tests are not reported here due to the space limit. They look similar to Test 1, 2007. Please write to the first author if you would like to see the complete set of results.

.



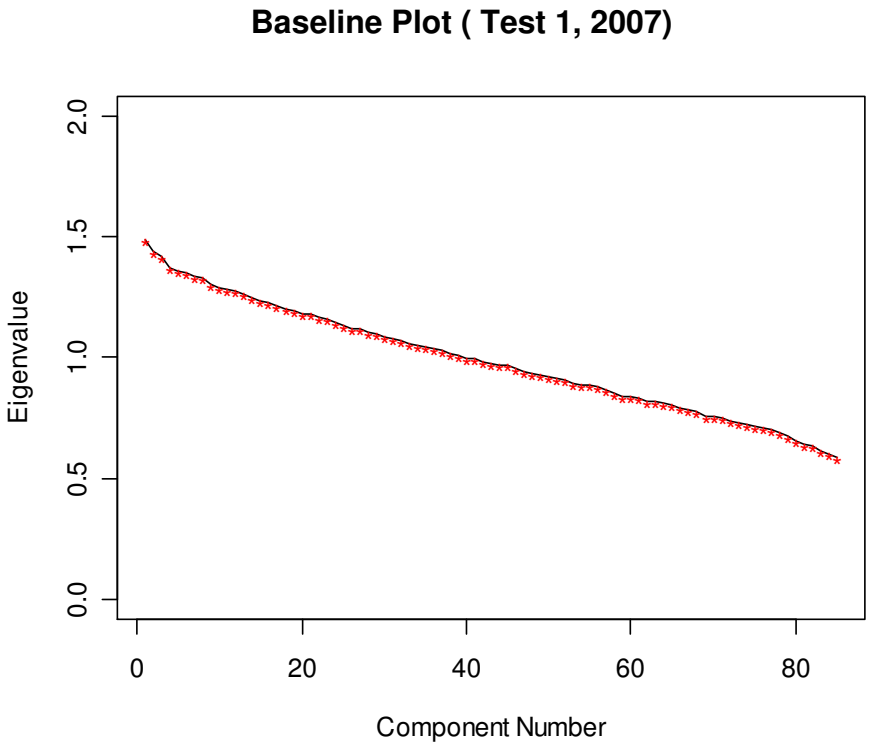


Figure 1. Eigenvalue plot and baseline plot for Test 1, 2007

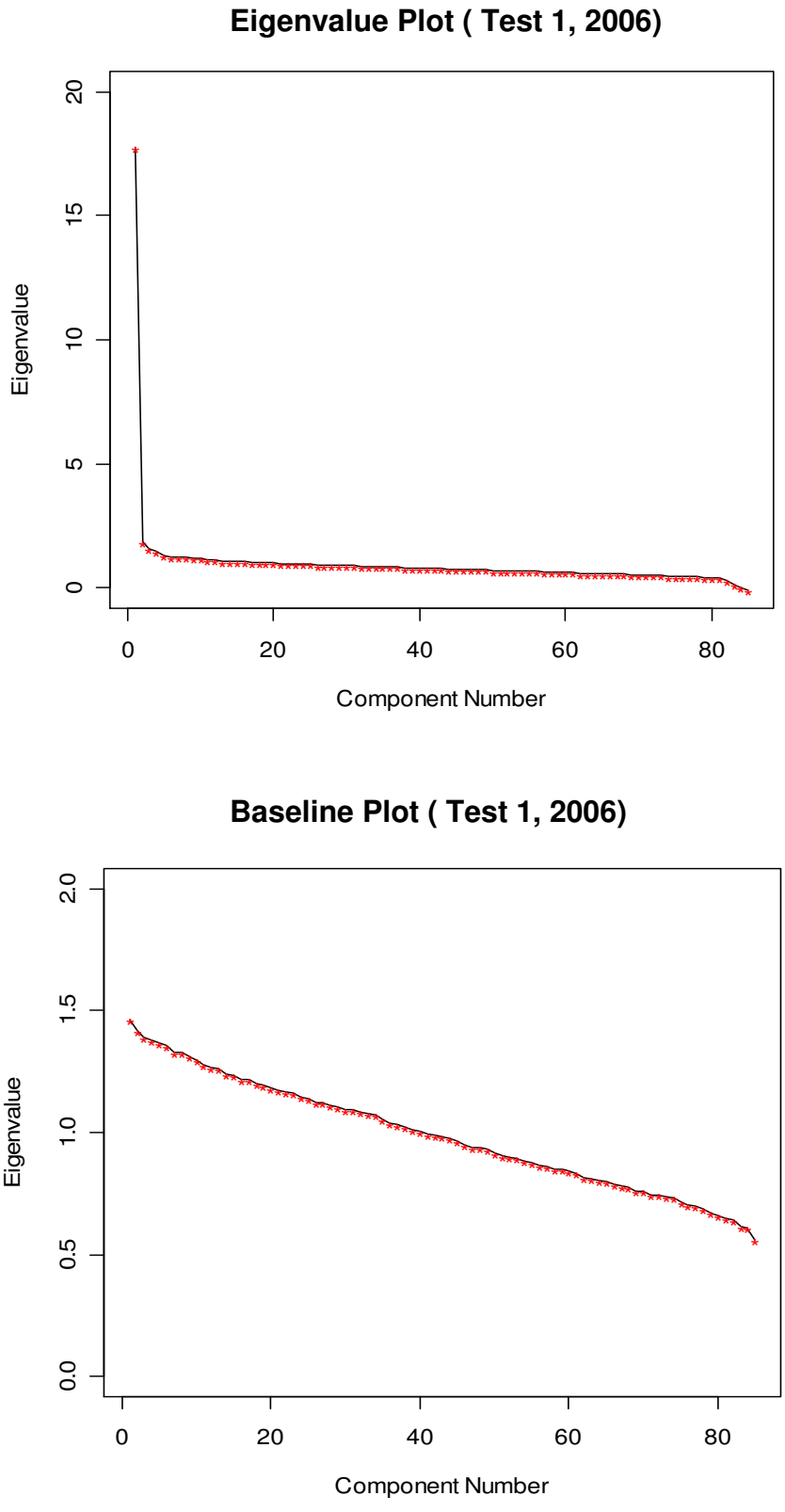
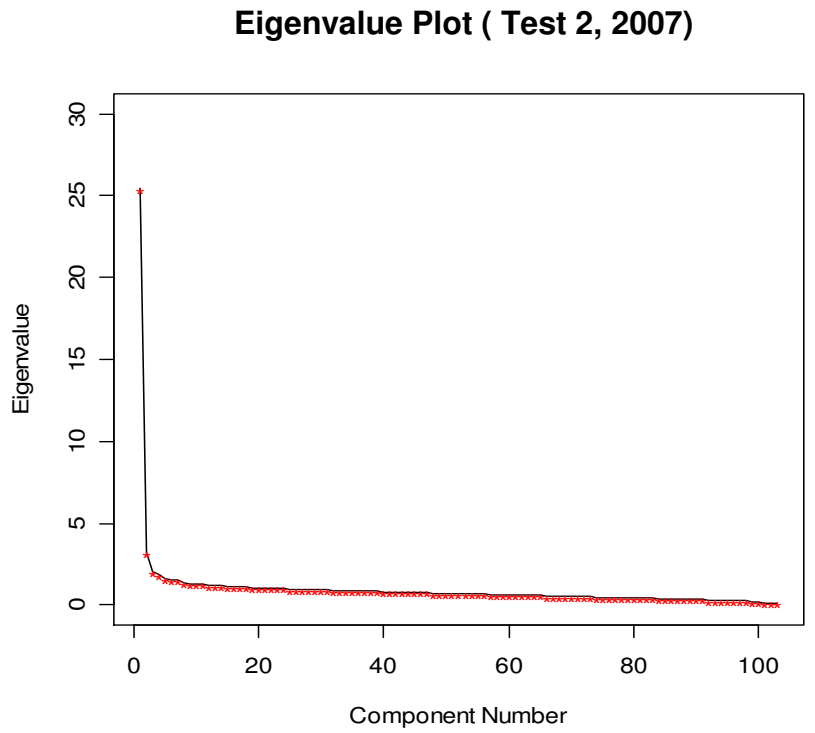


Figure 2. Eigenvalue plot and baseline plot for Test 1, 2006.



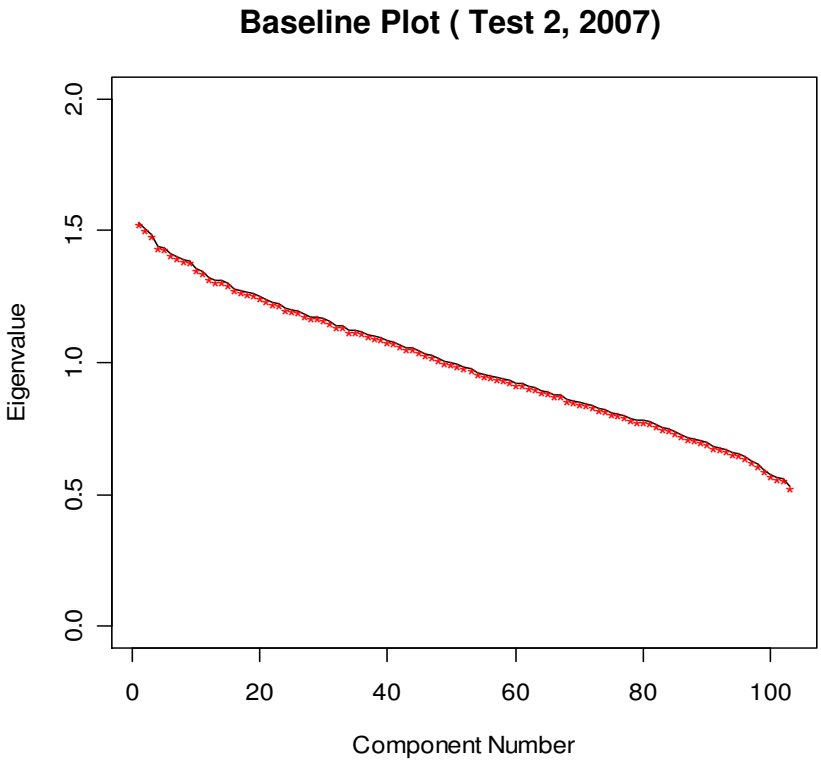


Figure 3. Eigenvalue plot and baseline plot for Test 2, 2007.

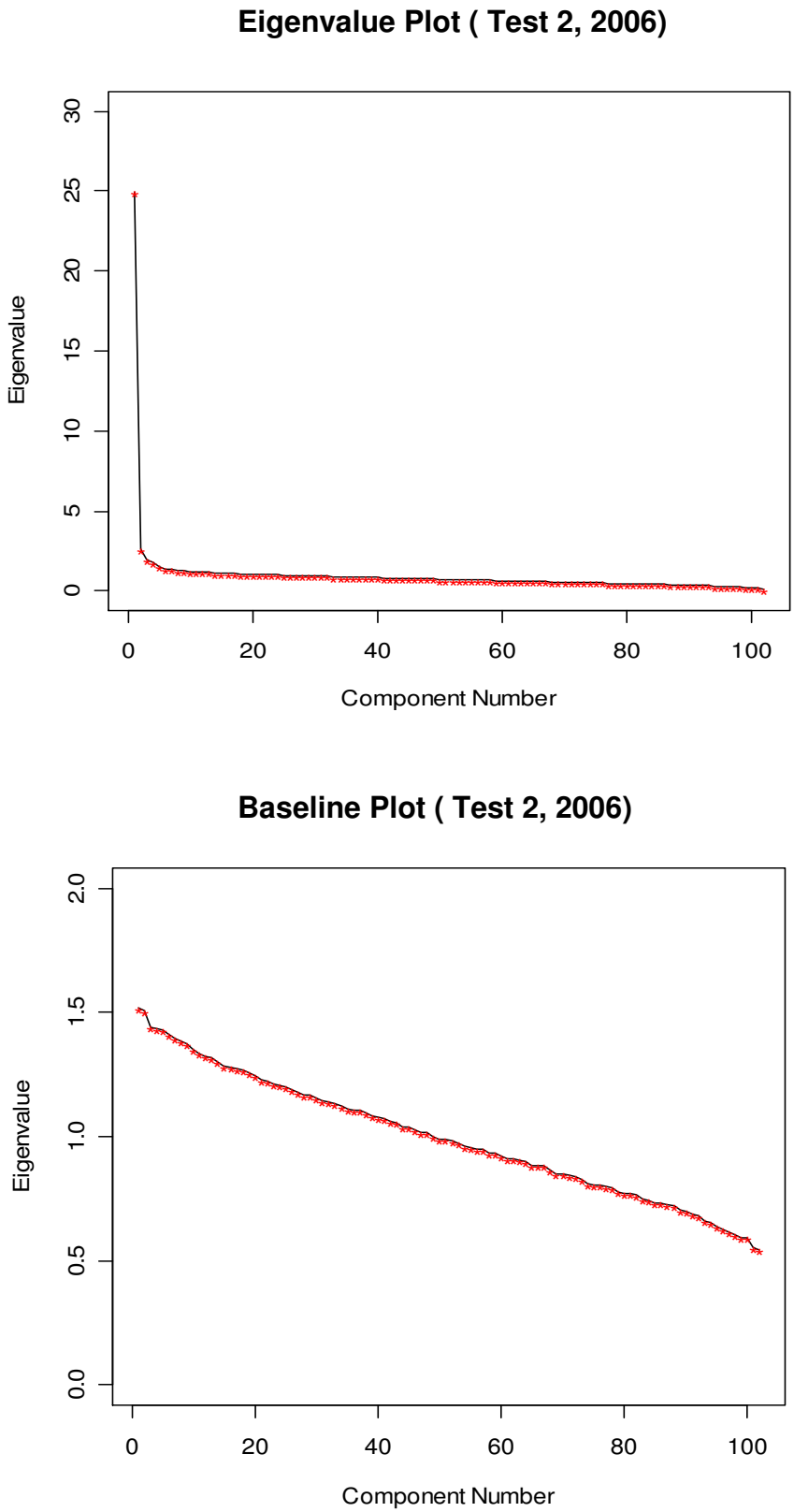


Figure 4. Eigenvalue plot and baseline plot for Test 2, 2006.