Running head: ITEM ORDERING ON HISTORY AND GOVERNMENT

Chronological Item Ordering:

Does It Make a Difference on a State History and Government Assessment?

Pui Chi Chiu and Patrick M. Irwin

University of Kansas

Abstract

The purpose of this study is to examine the effect of three different ways of item ordering (by content standards, chronologically − past to present, and present to past) on students' performance in the statewide assessment system. This study focused on high school students, grades 9 to 11, who took the History and Government assessment. The average percent correct scores and average correct responses on dated and non-dated items were calculated and compared across test forms. Item characteristic functions were also calculated and compared across test forms using the Mantel-Haenszel method. The results suggest that students answered significantly more dated items correct when they took a test form where the items were ordered from past to present. However, students answered significantly more non-dated items correct on the test form ordered by content standards.

Chronological Item Ordering:

Does It Make a Difference on a State History and Government Assessment?

In testing situations, the use of alternate test forms or forms constructed with the same items presented in different order is one of the strategies for deterring cheating and enhancing test security in test administration. Scrambling, or the rearrangement of the same set of items to create additional test forms, is often used to discourage examinee copying (Harris, 1991). However, the research has shown that varied item and section ordering can affect item and section characteristics (such as difficulty) and as a result have unintended effects on test scores (Pommerich & Harris, 2003; Zwick, 1991). These effects can make claims of test form interchangeability questionable and possibly violate testing industry standards (Moses, Yang, & Wilson, 2007).

Newman and colleague (1988) found that students (enrolled in an undergraduate educational psychology class) who received the forms with items in an increasing cognitive order scored higher on hard items, no matter what order of statistical difficulty; while students who received items in an decreasing cognitive order and statistical difficulty orders scored the highest on medium difficulty level items. Hambleton and Traub (1974) studied 11th graders' performance on an Algebra II Mathematics Test. They discovered the average number of correct answers for test questions arranged from easy-to-difficult was significantly higher than the test questions arranged from difficult-to-easy.

The previous research available mainly focuses on ordering items based on either item difficulty or cognitive level, the research focused on ordering items chronologically is nonexistent. This study focused on high school students, grades 9 to 11, who took the History and Government assessment. The History and Government assessment is part of a statewide assessment system in a Midwestern state. Items were arranged in three different ways: (1) by the states content standards, (2) chronologically from past to present, and (3) chronologically from present to past, to create three different test forms with the exact same items. The goal of the study was to investigate the effect of item orderings on students' performance on high school History and Government assessment.

Method

Participants

A total of 19,479 high school students (grades 9 to 11) took the History and Government assessment. The students were randomly assigned to one of the three test forms resulting in approximately 6,500 students on each test form. Of the 19,479 individual students, 49% were male; 27% qualified for free or reduced lunch support; and 76% were Caucasian, 10% were Hispanic, 7% were African American, 3% were Asian, 1% were Native American and 3% were classified as other. Special educational students (except gifted students) and students who were provided the read aloud accommodation were removed from the study. Table 1 illustrates students' demographic information for each of the test forms.

Assessment

The history and government assessment consisted of 30 items, which focused on history, geography, and economics. The item format for the assessment was multiple-choice, with one correct answer to be selected from four response options. There were three parallel test forms, and each form consisted of exactly the same items, but they were ordered in three different ways: (1) by content standards, (2) chronologically from past (the earliest date) to present (the latest date), and (3) chronologically from present (the latest date) to past (the earliest date). There were two types of items in each test form: dated items (year(s) specified) and non-dated items. All forms were administered via computer-based testing (CBT) and were randomly assigned to students during registration. When the students are signed up or registered to take a state assessment via CBT, they are randomly assigned to one of the three test forms.

Table 2 illustrates how items were ordered in each of the three test forms. For instance, on the form ordered by the content standards (Form 1), the first item (Question 1) is the 1st item (Question 1) on the form ordered from past to present (Form 2) and is the 30th item (Question 30) on the form ordered from present to past (Form 3).

Analyses

The effect of item ordering on students' performance was examined by looking at average percent correct scores of three test forms, average correct responses on each of the categories (all dated items, dated items in history, dated items in geography, dated item in economics, and non-dated items), items' proportion corrects (*p*-values), and item characteristic functions.

The average percent correct scores were calculated for each test form, and they were compared across forms using a univariate analysis of variance (ANOVA). The average correct responses of all dated items, dated items by content standards, and non-dated items for each test form were also calculated and compared across test forms using a multivariate analysis of variance (MANOVA). Follow-up tests were conducted to evaluate pairwise differences among these correct responses. The Tamhane's T2 procedure which does not assume equal variances across test forms was used to control for Type I error across the three pairwise comparisons.

P-values were calculated and plotted for each item, and the item characteristic functions were compared across test forms using the Mantel-Haenszel method to assess the presence of differential item functioning (DIF). Differential Item Functioning Analysis System (DIFAS)

(Penfield, 2005) was used to calculate the following: Mantel-Haenszel chi-square (MH CHI) (Mantel & Haenszel, 1959), Mantel-Haenszel common log-odds ratio (MH LOR), standard error of the Mantel-Haenszel common log-odds ratio (LOR SE), Breslow-Day chi-square (BD) (Breslow & Day, 1980), and the Educational Testing Service (ETS) categorization scheme (Zieky, 1993).

A critical value of 6.63 at the 0.01 significance level was used for the statistic tests of the Mantel-Haenszel chi-square and the Breslow-Day chi-square. The Mantel-Haenszel common log-odds ratio is asymptotically normally distributed; positive values indicate DIF in favor of the reference group, and negative values indicate DIF in favor of the focal group. The Breslow-Day chi-square statistic test has been shown to be effective at detecting non-uniform DIF; the calculations are similar to the Mantel-Haenszel chi-square statistic test. The ETS categorization scheme categorizes items as having small (A), moderate (B), and large (C) levels of DIF.

Results

A total of 19,479 students took the History and Government assessment. Of those, 6,502 students took Form 1, 6,489 students took Form 2, and 6,488 students took Form 3. Across the three test forms, the average percent correct score obtained on the assessment was 56.0 (*SD* = 15.7), which is about 17 items answered correctly. For the 19 dated items, the average correct response was 10.7 (*SD* = 3.4), which is about 11 items answered correctly. For the 11 non-dated items, the average correct response was 6.1 (*SD* = 2.2), which is about 6 items answered correctly. Summary statistics for the entire History and Government assessment and the three test forms are given in Tables 3 and 4 respectively.

A one-way ANOVA was conducted to explore differences in students' percent correct scores across test forms. The results indicate a statistically significant differences in students' percent correct scores across test forms, $F_{(2,\ 19,476)} = 4.890$, $p < 0.01$, partial eta squared (η^2) = 0.001. Tamhane's T2 post hoc simple effect test suggests students obtained significantly higher scores on Form 3 compared to Form 1, $p < 0.01$. However, there were no statistically significant differences between Forms 1 and 2, and between Forms 2 and 3.

A one-way MANOVA was also conducted to evaluate differences in students' correct responses on all dated items, dated items in history, dated items in geography, dated item in economics, and non-dated items across test forms. The results indicate there were statistically significant differences in all of the students' correct responses across test forms, $p < 0.001$, see Table 5. For all of the dated items, Tamhane's T2 post hoc test suggests that students correctly answered significantly more items (about two items) on Form 2 compared to Forms 1 and 3, $p < 0.001$. However, there were no statistically significant differences between Forms 1 and 3.

For the dated items in history, Tamhane's T2 post hoc test indicates that students answered significantly more items correctly (about two items) on Form 2 compared to Forms 1 and 3, $p < 0.001$. However, there were no statistically significant differences between Forms 1 and 3.

For the dated items in geography, Tamhane's T2 post hoc test suggests that students who took Forms 1 and 3 obtained significantly higher average correct responses (less than one item) than the students who took Form 2, $p < 0.001$. Also, students who took Form 3 obtained significantly higher average correct responses than the students who took Form 1, $p < 0.001$.

For the dated item in economics, though Tamhane's T2 post hoc test indicates that students obtained significantly lower average correct response on Form 2 compared to Forms 1 and 3, $p < 0.001$, there was only one item in the assessment. Also, there were no statistically significant differences between Forms 1 and 3.

For the non-dated items, Tamhane's T2 post hoc test suggests that students who took Form 3 obtained significantly higher average correct responses than the students who took Forms 1 and 2, $p < 0.001$; students who took Form 3 answered two more items correctly than the students who took Form 2, but there was less than one item difference between Forms 1 and 3. The Tamhane's T2 post hoc test also suggests that students answered significantly more items correctly (about two items) on Form 1 compared to Form 2, $p < 0.001$.

In additional to looking at the average percent correct scores and average correct responses, *p*-value for each item was also calculated and evaluated across test forms, see Table 6. Figures 1 through 6 show item's *p*-value for each test form under different categories; the item numbers that were used in the figures were based on the item numbers on Form 1, and the *p*-values shown in the figures were sorted from lowest to highest based on the *p*-values of the Form 1. The Mantel-Haenszel method was used to examine item parameters' differences across test forms. Tables 7 through 9 summarize the results obtained from DIFAS.

Three pairwise comparisons were conducted to assess the presence of DIF. The results of the Mantel-Haenszel chi-square statistics indicate that nine items were identified as uniform DIF when comparing Form 1 to Form 2. In comparing Form 1 to Form 3, there were 16 items identified as uniform DIF. In comparing Form 2 to Form 3, there were 16 items identified as uniform DIF.

The Breslow-Day chi-square statistics identified the following: no items were identified as non-uniform DIF when comparing Form 1 to Form 2, a total of eight items were identified as non-uniform DIF (three of which were not identified by the Mantel-Haenszel chi-square statistic) when comparing Form 1 to Form 3, a total of nine items were identified as non-uniform DIF (four of which were not identified by the Mantel-Haenszel chi-square statistic) when comparing Form 2 to Form 3.

Discussion

Over the years testing, cheating, and test security has increased along with the use of high stakes testing. The federal mandate has not yet included history and government, yet some states, districts, and schools have high stakes History and Government assessment. The results of this study not only relate to History and Government items, but they also relate to item ordering in general. The use of multiple test forms constructed with the same items presented in different order (scrambled forms) is one strategy to enhance test security and deter cheating. However, when scrambled forms and the base form are administered at the same time, the question of equity arises (Harris, 1991). Thus, caution should be used when scrambled forms are being administered, if item ordering has an effect on students' performance.

Many ways have been used in ordering items in a test form. The most common way is to order the items based on the order in which material was presented in class (Form 2). History text books and courses are usually organized chronologically, following some sort of time line. Organizing the items on an assessment chronologically should help students retrieve learned information. Gestalt theory and research showed that well-organized material is easier to learn and recall (Katona, 1940). Organized material improves memory because items are linked to it. Recall of one item may prompt the recall of other linked items (Schunk, 2004). Organizing the assessment in the same order in which the material presented and learned should assist the memory system to locate the stored information in their memory network for answering the questions on the assessment.

The results of this study indicate students answered more dated items correctly when the items were ordered from past to present (Form 2), which suggests that students performed better when the items were ordered based on when the events happened. However, students answered more non-dated items correctly when they took a test form where the content standards were placed together (Form 1), which indicates that organizing items based on content standards also assists students. On the other hand, students also answered more non-dated items correctly when the items were placed at the beginning of the test (Form 3), compared to the same items placed at the end of the test (Form 2). In general, regardless of content, the results suggest that fatigue effects may play a role in testing situations; students tended to perform better when the items were placed at the beginning of the test than when the same items were placed toward the end of the test. These results support the research on fatigue effects that test takers have subjective feeling of tiredness, change quantity or quality in work output, and decrease capacity to do work as a direct result of having worked (Spaeth, 1920; Bills, 1937).

Limitations

This study had quite a few limitations to overcome, but the limitation that had the greatest impact was the lack of items' variation in the History and Government assessment. This assessment had only 30 items that covered 15 indicators, across three content standards (history, geography, and economics). The number of items for each standard was also disproportionate; more items were tested in the content standard of history than in geography and economics, especially the items with date(s). In general, this assessment is quite limited in depth and breadth of the overall concept of history and government. Furthermore, this study only looked at high school students in the content area of history and government. The other content areas and grade levels were not examined. In addition to the lack of item variation, the ordering or shuffling of items across test forms was not equally balanced. The item ordering on Forms 1 and 2 had the dated items in front of the test, whereas Form 3 had all of the items that contained dates toward the end of the test. The order was not exactly the same but was very similar on Forms 1 and 2, whereas Form 3 was quite different. See Table 2 for a visual representation of this information.

Future Research

Based on the limitations presented above, future research should look at larger sets of test items, multiple grade levels, and a more complete set of history and government items. Future research should attempt to order a larger set of dated items on a specific event or time period in history (for example, World War II: 1930s to 1950s) and explore the effects of item ordering. This would help to eliminate the majority of the limitations mentioned above, and also produce more generalizable results.

References

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: Volume 1 – The analysis of case-control studies*. Lyon: International Agency for Research on Cancer.

Bills, A. G. (1937). Fatigue in mental work. *Physiological Review, 17*, 436–453.

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *The Journal of Experimental Education, 43*(1), 40-46.

Harris, D. J. (1991). Effects of Passage and Item Scrambling on Equating Relationships. *Applied Psychological Measurement, 15* (3), 247-256.

Katona, G. (1940). Organizing and memorizing. In Schunk, D. H., *Learning theories: an educational perspective* (4th ed., pp. 160 – 161). Upper Saddle River, NJ: Pearson Education.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Moses, T., Yang, W., & Wilson, C. (2007). Using Kernel Equating to Assess Item Order Effects on Test Scores. *Journal of Educational Measurement, 44* (2), 157-178.

Newman, D. L. et al. (1988). Effects of varying item order on multiple-choice test scores: Importance of statistical and cognitive difficulty. *Applied Measurement in Education, 1*(1), 89-97.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement, 29*, 150–151.

Pommerich, M., & Harris, D. J. (2003). *Context Effects in Pretesting: Impact on Item Statistics and Examinee Scores*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Schunk, D. H. (2004). Information Processing. In *Learning theories: an educational perspective* (4th ed., pp. 136 – 189). Upper Saddle River, NJ: Pearson Education.

Spaeth, R. A. (1920). The problem of fatigue. *Journal of Industrial Hygiene, 1*, 22–53.

Zieky, M. (1993). Practical questions use of DIF statistics it item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-364). Hillsdale, NJ: Lawrence Erlbaum.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *Bilog MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.

Table 1

Students' Demographic Information for Each of the Test Forms.

		Form 1	Form 2	Form 3	Total
Gender	**Female**	50.5%	50.8%	50.8%	**51%**
	Male	49.5%	49.2%	49.2%	**49%**
Socioeconomic Status	**Not Qualified for Lunch Support**	71.6%	72.6%	73.4%	**73%**
	Qualified for Free or Reduced Lunch Support	28.4%	27.4%	26.6%	**27%**
Ethnicity	**Caucasians**	75.2%	75.3%	76.8%	**76%**
	Hispanics	10.5%	10.4%	9.4%	**10%**
	African Americans	7.0%	6.8%	6.5%	**7%**
	Asians	2.8%	2.9%	2.9%	**3%**
	Native Americans	1.0%	0.9%	1.0%	**1%**
	Others	3.5%	3.7%	3.4%	**3%**

Table 2

Item Orders for Each of the Test Forms.

Form 1			Form 2			Form 3		
Item Number	**Content Standard**	**Date (Year)**	**Item Number**	**Content Standard**	**Date (Year)**	**Item Number**	**Content Standard**	**Date (Year)**
1	History	1495	1	History	1495	1	Geography	
2	History	1532	2	History	1532	2	Geography	
3	History	1630	3	History	1600	3	Geography	
4	History	1600	4	History	1630	4	Geography	
5	History	1637	5	History	1637	5	Geography	
6	History	1637	6	History	1637	6	Geography	
7	History	1800	7	History	1791	7	Economics	
8	History	1800	8	History	1800	8	Economics	
9	History	1830 - 1940	9	History	1800	9	Economics	
10	History	1791	10	History	1830 - 1940	10	Economics	1900
11	History	1933 - 1945	11	History	1933 - 1945	11	Economics	
12	History	1933 - 1945	12	History	1933 - 1945	12	Economics	
13	History	1960	13	History	1956	13	Geography	1900
14	History	1956	14	Geography	1947 - 1996	14	Geography	1979 - 2000
15	Geography		15	History	1960	15	Geography	1980 - 2000
16	Geography		16	Geography	1980 - 2000	16	History	1960
17	Geography	1947 - 1996	17	Geography	1979 - 2000	17	Geography	1947 - 1996
18	Geography	1980 - 2000	18	Geography		18	History	1956
19	Geography	1979 - 2000	19	Geography		19	History	1933 - 1945
20	Geography		20	Geography	1900	20	History	1933 - 1945
21	Geography	1900	21	Economics	1900	21	History	1830 - 1940
22	Geography		22	Economics		22	History	1800
23	Geography		23	Economics		23	History	1800
24	Geography		24	Economics		24	History	1791
25	Economics		25	Economics		25	History	1637
26	Economics		26	Economics		26	History	1637
27	Economics		27	Geography		27	History	1630
28	Economics		28	Geography		28	History	1600
29	Economics	1900	29	Geography		29	History	1532
30	Economics		30	Geography		30	History	1495

Table 3

Summary Statistics for the History and Government Assessment.

	Entire Assessment	All Dated Items	Dated Items: History	Dated Items: Geography	Dated Item: Economics	Non-Dated Items
Number of Items	30	19	14	4	1	11
Average Correct Responses	16.8	10.7	7.4	2.6	0.7	6.1
Standard Deviation of Average Correct Responses	4.7	3.4	2.7	1.1	0.5	2.2

Table 4

Summary Statistics for Each of the Test Forms.

	Form 1 **(Content Standard)**	**Form 2** **(Past to Present)**	**Form 3** **(Present to Past)**
Number of Examinees	**6,502**	**6,489**	**6,488**
Number of Items: Entire Test	**30**		
Average Percent Correct Score (*SD*)	**55.5** **(15.6)**	**56.0** **(16.0)**	**56.4** **(15.4)**
Average Correct Response (*SD*)	**16.7** **(4.7)**	**16.8** **(4.8)**	**16.9** **(4.6)**
Number of Items: All Dated Items	**19**		
Average Correct Response: All Dated Items (*SD*)	**10.1** **(3.2)**	**11.7** **(3.3)**	**10.2** **(3.3)**
Number of Items: Dated Items in History	**14**		
Average Correct Response: Dated Items in History (*SD*)	**6.7** **(2.5)**	**8.8** **(2.6)**	**6.6** **(2.6)**
Number of Items: Dated Items in Geography	**4**		
Average Correct Response: Dated Items in Geography (*SD*)	**2.7** **(1.0)**	**2.3** **(1.1)**	**2.8** **(1.0)**
Number of Items: Dated Items in Economics	**1**		
Average Correct Response: Dated Items in Economics (*SD*)	**0.7** **(0.5)**	**0.6** **(0.5)**	**0.7** **(0.5)**
Number of Items: Non-Dated Items	**11**		
Average Correct Response: Non-Dated Items (*SD*)	**6.6** **(2.0)**	**5.1** **(2.2)**	**6.8** **(1.9)**
Reliability: Coefficient Alpha	**0.72**	**0.74**	**0.72**

Table 5

Results of the MANOVA in Students' Average Correct Responses.

Source	Average Correct Responses	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Test Forms	All Dated Items	11,192.67	2	5,596.33	525.65	0.000	0.05
	Dated Items in History	19,695.56	2	9,847.78	1,532.00	0.000	0.14
	Dated Items in Geography	984.14	2	492.07	441.12	0.000	0.04
	Dated Items in Economics	15.20	2	7.60	34.26	0.000	0.00
	Non-Dated Items	11,016.16	2	5,508.08	1,306.13	0.000	0.12
Error	All Dated Items	207,352.76	19,476	10.65			
	Dated Items in History	125,192.65	19,476	6.43			
	Dated Items in Geography	21,725.23	19,476	1.12			
	Dated Items in Economics	4,320.11	19,476	0.22			
	Non-Dated Items	82,132.37	19,476	4.22			
Total	All Dated Items	2,436,295.00	19,479				
	Dated Items in History	1,204,059.00	19,479				
	Dated Items in Geography	157,508.00	19,479				
	Dated Items in Economics	12,966.00	19,479				
	Non-Dated Items	824,407.00	19,479				

Table 6

Items' Proportion Correct by Test Form (Item Number based on Form 1).

Item Number	Content Standard	Date (Year)	Form 1		Form 2		Form 3	
			Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
1	History	1495	**0.53**	0.50	**0.53**	0.50	**0.53**	0.50
2	History	1532	**0.51**	0.50	**0.51**	0.50	**0.41**	0.49
3	History	1630	**0.29**	0.45	**0.32**	0.47	**0.33**	0.47
4	History	1600	**0.36**	0.48	**0.36**	0.48	**0.31**	0.46
5	History	1637	**0.34**	0.47	**0.35**	0.48	**0.36**	0.48
6	History	1637	**0.39**	0.49	**0.39**	0.49	**0.44**	0.50
7	History	1800	**0.49**	0.50	**0.50**	0.50	**0.49**	0.50
8	History	1800	**0.54**	0.50	**0.53**	0.50	**0.49**	0.50
9	History	1830 - 1940	**0.62**	0.49	**0.63**	0.48	**0.60**	0.49
10	History	1791	**0.41**	0.49	**0.46**	0.50	**0.40**	0.49
11	History	1933 - 1945	**0.52**	0.50	**0.50**	0.50	**0.51**	0.50
12	History	1933 - 1945	**0.62**	0.48	**0.63**	0.48	**0.66**	0.47
13	History	1960	**0.48**	0.50	**0.48**	0.50	**0.48**	0.50
14	History	1956	**0.58**	0.49	**0.58**	0.49	**0.63**	0.48
15	Geography		**0.95**	0.22	**0.95**	0.22	**0.95**	0.23
16	Geography		**0.83**	0.37	**0.84**	0.37	**0.85**	0.35
17	Geography	1947 - 1996	**0.61**	0.49	**0.64**	0.48	**0.70**	0.46
18	Geography	1980 - 2000	**0.66**	0.47	**0.67**	0.47	**0.69**	0.46
19	Geography	1979 - 2000	**0.62**	0.49	**0.61**	0.49	**0.61**	0.49
20	Geography		**0.49**	0.50	**0.49**	0.50	**0.52**	0.50
21	Geography	1900	**0.85**	0.36	**0.85**	0.36	**0.85**	0.36
22	Geography		**0.68**	0.47	**0.67**	0.47	**0.71**	0.45
23	Geography		**0.86**	0.35	**0.85**	0.36	**0.87**	0.33
24	Geography		**0.53**	0.50	**0.51**	0.50	**0.58**	0.49
25	Economics		**0.33**	0.47	**0.31**	0.46	**0.32**	0.46
26	Economics		**0.53**	0.50	**0.52**	0.50	**0.55**	0.50
27	Economics		**0.34**	0.47	**0.34**	0.47	**0.34**	0.47
28	Economics		**0.57**	0.49	**0.61**	0.49	**0.61**	0.49
29	Economics	1900	**0.68**	0.47	**0.71**	0.45	**0.69**	0.46
30	Economics		**0.45**	0.50	**0.47**	0.50	**0.44**	0.50

Table 7

Form 1 vs. Form 2: DIF Statistics.

Item Number	Content Standard	Date (Year)	MH CHI	MH LOR	LOR SE	BD	ETS
1	History	1495	0.17	0.02	0.04	5.65	A
2	History	1532	0.33	0.02	0.04	0.00	A
3	History	1630	9.26*	-0.12	0.04	0.00	A
4	History	1600	0.01	0.00	0.04	5.38	A
5	History	1637	0.23	-0.02	0.04	4.67	A
6	History	1637	1.34	0.05	0.04	0.23	A
7	History	1800	0.01	0.00	0.04	2.49	A
8	History	1800	7.84*	0.11	0.04	0.02	A
9	History	1830 - 1940	0.36	-0.02	0.04	0.05	A
10	History	1791	29.80*	-0.21	0.04	4.20	A
11	History	1933 - 1945	10.97*	0.13	0.04	0.03	A
12	History	1933 - 1945	0.03	0.01	0.04	0.32	A
13	History	1960	0.70	0.03	0.04	0.01	A
14	History	1956	0.00	0.00	0.04	2.24	A
15	Geography		1.05	0.09	0.09	0.24	A
16	Geography		0.05	-0.01	0.05	0.25	A
17	Geography	1947 - 1996	7.75*	-0.11	0.04	2.22	A
18	Geography	1980 - 2000	1.71	-0.05	0.04	0.39	A
19	Geography	1979 - 2000	0.76	0.03	0.04	0.17	A
20	Geography		1.43	0.05	0.04	3.50	A
21	Geography	1900	0.48	-0.04	0.05	0.10	A
22	Geography		1.15	0.05	0.04	0.31	A
23	Geography		1.08	0.06	0.05	0.04	A
24	Geography		9.21*	0.12	0.04	0.28	A
25	Economics		15.77*	0.16	0.04	0.55	A
26	Economics		2.50	0.06	0.04	0.37	A
27	Economics		0.61	0.03	0.04	6.46	A
28	Economics		20.88*	-0.17	0.04	1.30	A
29	Economics	1900	11.21*	-0.15	0.04	1.30	A
30	Economics		0.99	-0.04	0.04	0.86	A

*Significant at 0.01 alpha level and the corresponding critical value is 6.63
Note: Form 1 as the reference group

Table 8

Form 1 vs. Form 3: DIF Statistics.

Item Number	Content Standard	Date (Year)	MH CHI	MH LOR	LOR SE	BD	ETS
1	History	1495	0.33	0.02	0.04	6.13	A
2	History	1532	151.06*	0.47	0.04	0.93	B
3	History	1630	20.89*	-0.18	0.04	0.63	A
4	History	1600	37.45*	0.24	0.04	0.01	A
5	History	1637	0.99	-0.04	0.04	0.71	A
6	History	1637	25.48*	-0.20	0.04	14.63*	A
7	History	1800	0.90	0.03	0.04	0.94	A
8	History	1800	67.61*	0.33	0.04	0.03	A
9	History	1830 - 1940	7.87*	0.11	0.04	18.62*	A
10	History	1791	4.83	0.09	0.04	0.01	A
11	History	1933 - 1945	7.71*	0.11	0.04	0.94	A
12	History	1933 - 1945	10.97*	-0.14	0.04	11.70*	A
13	History	1960	4.44	0.08	0.04	10.10*	A
14	History	1956	18.46*	-0.16	0.04	0.01	A
15	Geography		7.87*	0.24	0.08	0.00	A
16	Geography		4.94	-0.12	0.05	0.49	A
17	Geography	1947 - 1996	93.25*	-0.39	0.04	4.97	A
18	Geography	1980 - 2000	9.53*	-0.13	0.04	0.22	A
19	Geography	1979 - 2000	5.07	0.09	0.04	4.27	A
20	Geography		5.86	-0.09	0.04	0.22	A
21	Geography	1900	0.54	0.04	0.05	7.29*	A
22	Geography		12.80*	-0.15	0.04	6.66*	A
23	Geography		2.87	-0.09	0.05	10.25*	A
24	Geography		29.30*	-0.22	0.04	0.11	A
25	Economics		9.88*	0.13	0.04	3.95	A
26	Economics		2.56	-0.06	0.04	1.42	A
27	Economics		0.90	0.04	0.04	1.44	A
28	Economics		12.70*	-0.13	0.04	19.21*	A
29	Economics	1900	0.72	0.04	0.04	4.92	A
30	Economics		4.37	0.08	0.04	3.92	A

*Significant at 0.01 alpha level and the corresponding critical value is 6.63
Note: Form 1 as the reference group

Table 9

Form 2 vs. Form 3: DIF Statistics.

Item Number	Content Standard	Date (Year)	MH CHI	MH LOR	LOR SE	BD	ETS
1	History	1495	0.03	0.01	0.04	0.05	A
2	History	1532	135.19*	0.44	0.04	0.56	B
3	History	1630	2.41	-0.06	0.04	0.67	A
4	History	1600	36.84*	0.24	0.04	5.31	A
5	History	1637	0.25	-0.02	0.04	1.45	A
6	History	1637	38.73*	-0.25	0.04	11.14*	A
7	History	1800	1.23	0.04	0.04	0.28	A
8	History	1800	27.83*	0.21	0.04	0.26	A
9	History	1830 - 1940	10.93*	0.12	0.04	19.70*	A
10	History	1791	59.10*	0.30	0.04	3.78	A
11	History	1933 - 1945	0.31	-0.02	0.04	0.72	A
12	History	1933 - 1945	12.54*	-0.15	0.04	16.33*	A
13	History	1960	1.64	0.05	0.04	10.05*	A
14	History	1956	18.77*	-0.17	0.04	2.55	A
15	Geography		2.89	0.15	0.08	0.21	A
16	Geography		3.75	-0.10	0.05	1.57	A
17	Geography	1947 - 1996	46.14*	-0.28	0.04	0.87	A
18	Geography	1980 - 2000	3.23	-0.08	0.04	0.01	A
19	Geography	1979 - 2000	1.80	0.05	0.04	3.07	A
20	Geography		12.19*	-0.13	0.04	6.10	A
21	Geography	1900	1.83	0.07	0.05	9.02*	A
22	Geography		21.21*	-0.19	0.04	4.33	A
23	Geography		7.84*	-0.15	0.05	12.77*	A
24	Geography		70.85*	-0.34	0.04	0.01	A
25	Economics		0.73	-0.04	0.04	6.73*	A
26	Economics		10.37*	-0.12	0.04	0.53	A
27	Economics		0.03	0.01	0.04	2.01	A
28	Economics		1.22	0.04	0.04	30.39*	A
29	Economics	1900	17.18*	0.18	0.04	0.84	A
30	Economics		9.41*	0.11	0.04	8.79*	A

*Significant at 0.01 alpha level and the corresponding critical value is 6.63
Note: Form 2 as the reference group

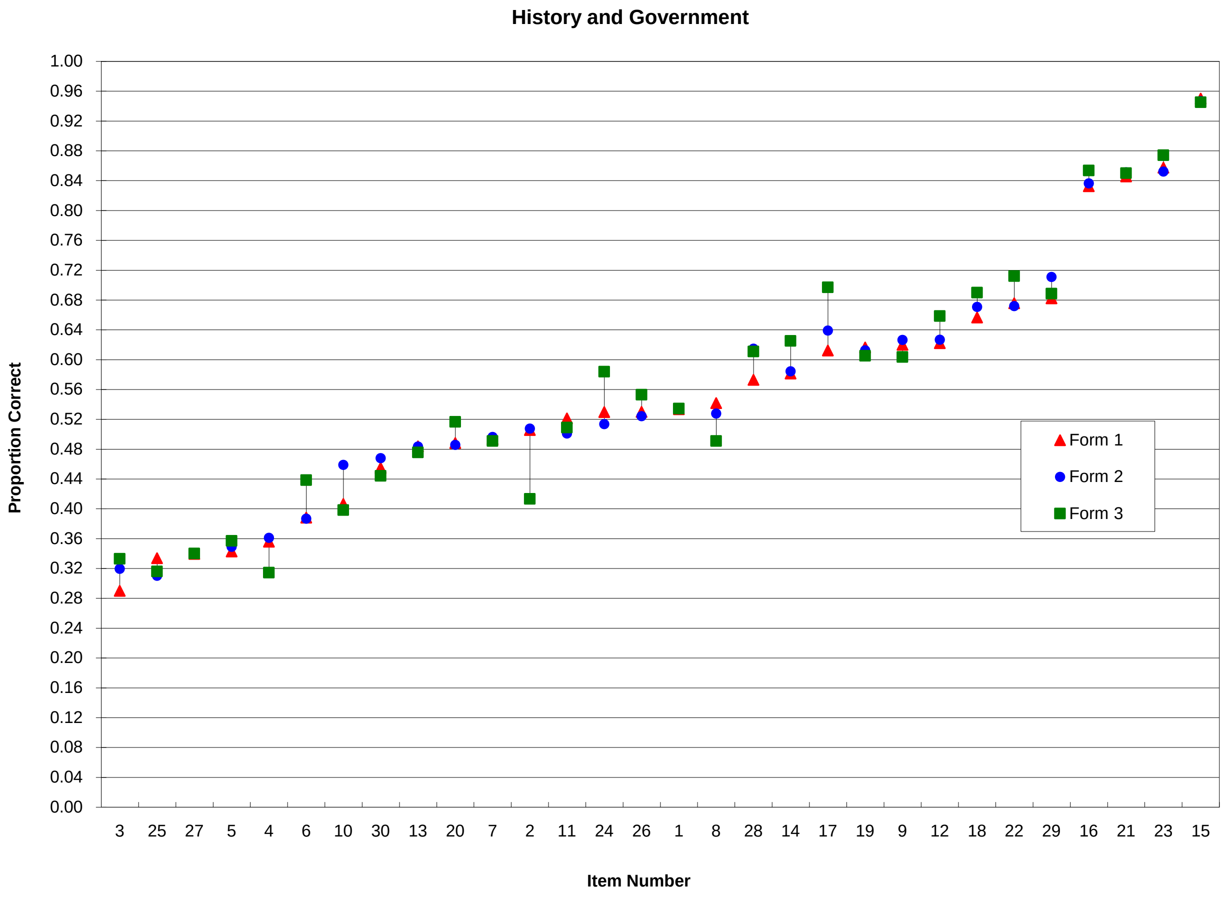

Figure 1. Entire Assessment: Plot of Item Proportion Correct (ordered from lowest to highest) by Test Form.

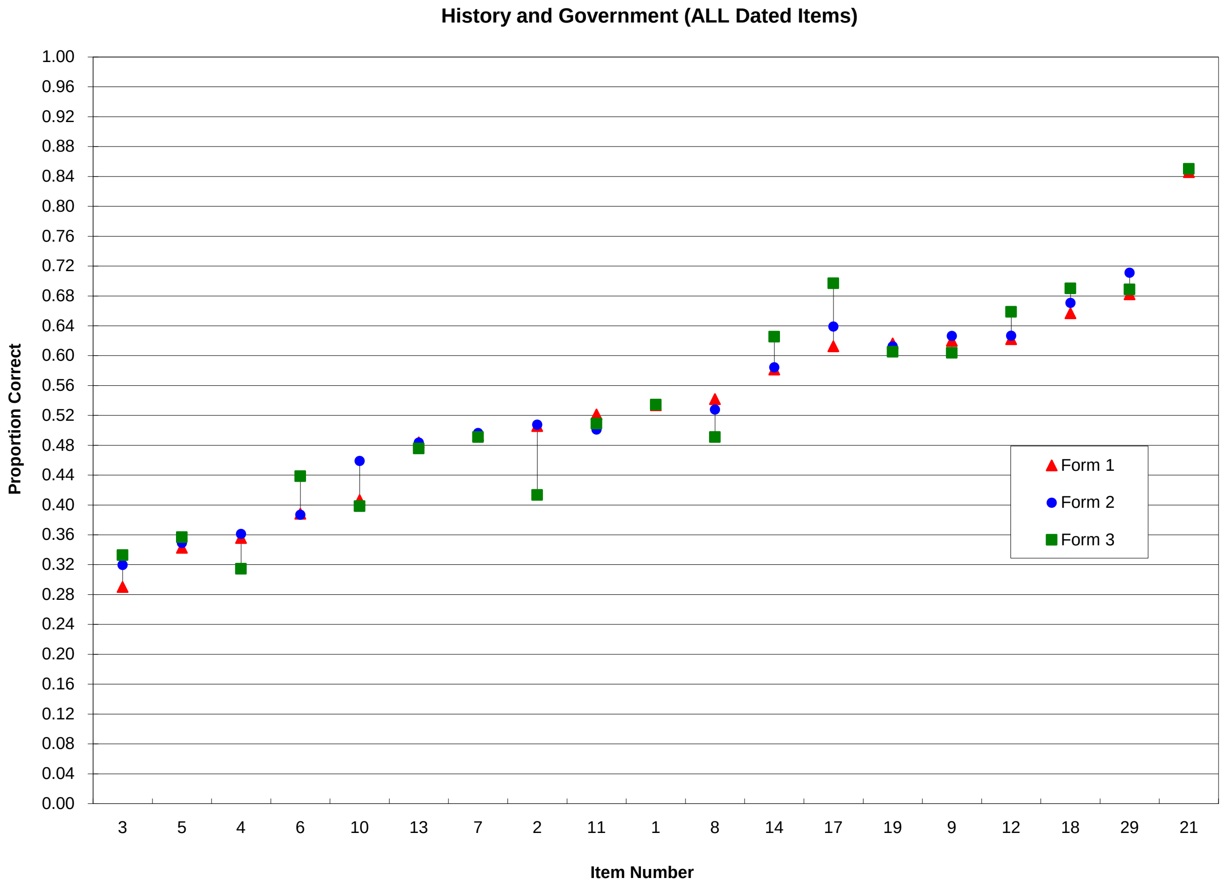

Figure 2. All Dated Items: Plot of Item Proportion Correct (ordered from lowest to highest) by Test Form.

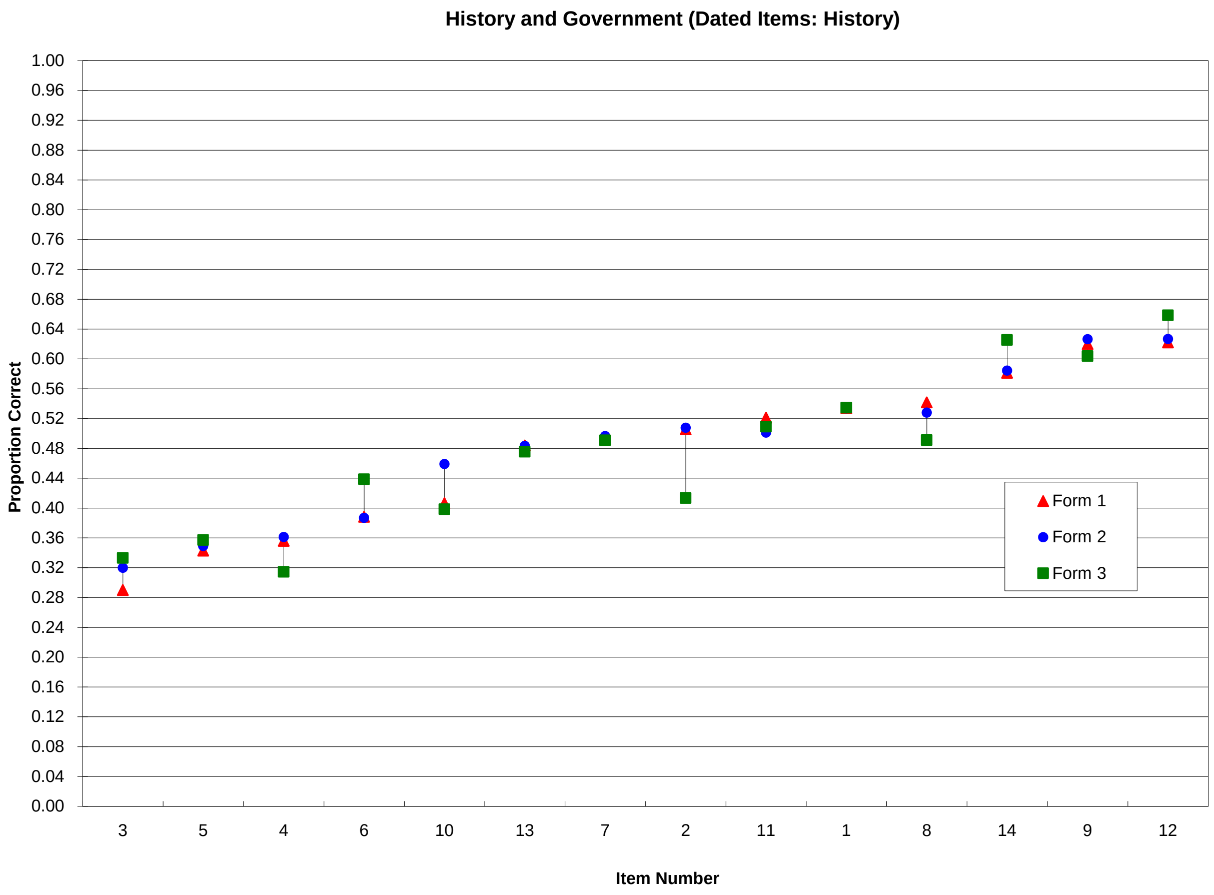

Figure 3. *Dated Items in History: Plot of Item Proportion Correct (ordered from lowest to highest) by Test Form.*

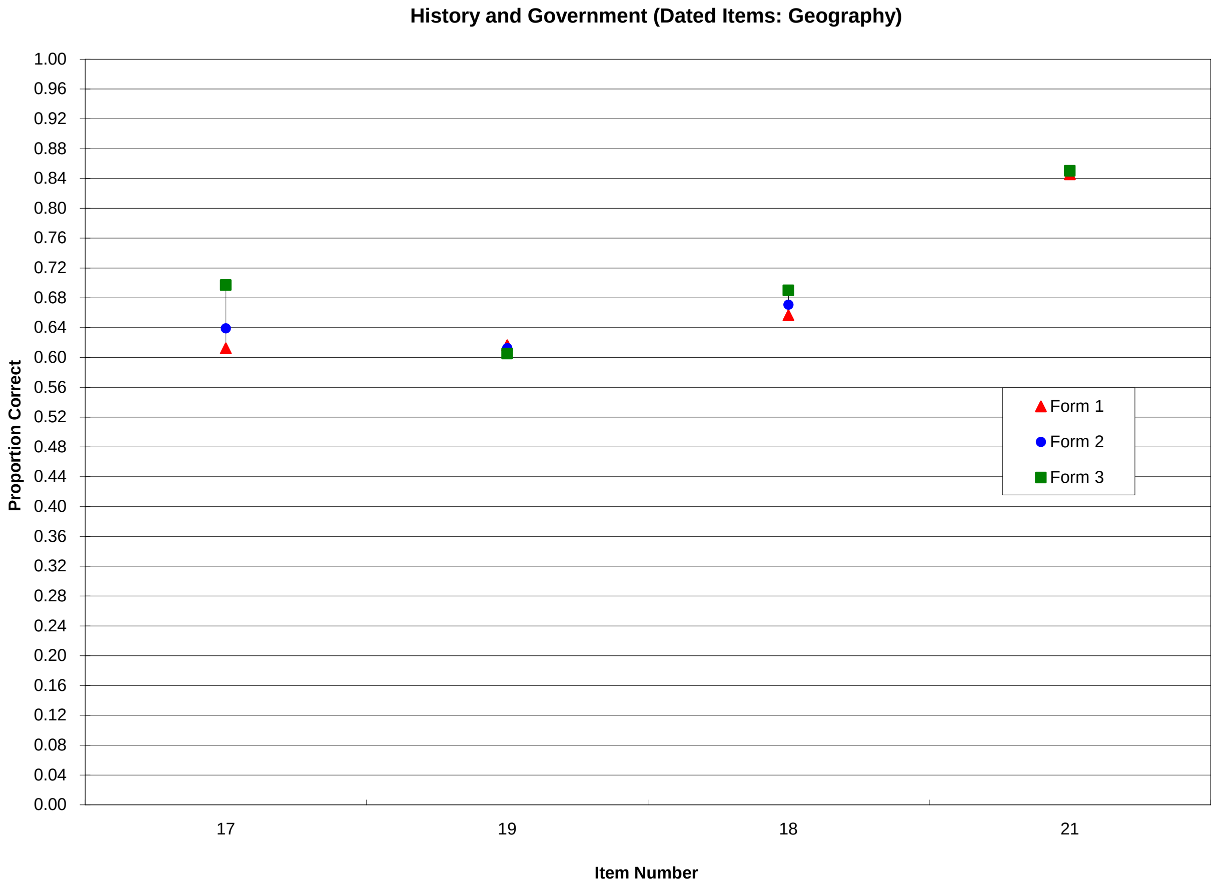

Figure 4. Dated Items in Geography: Plot of Item Proportion Correct (ordered from lowest to highest) by Test Form.

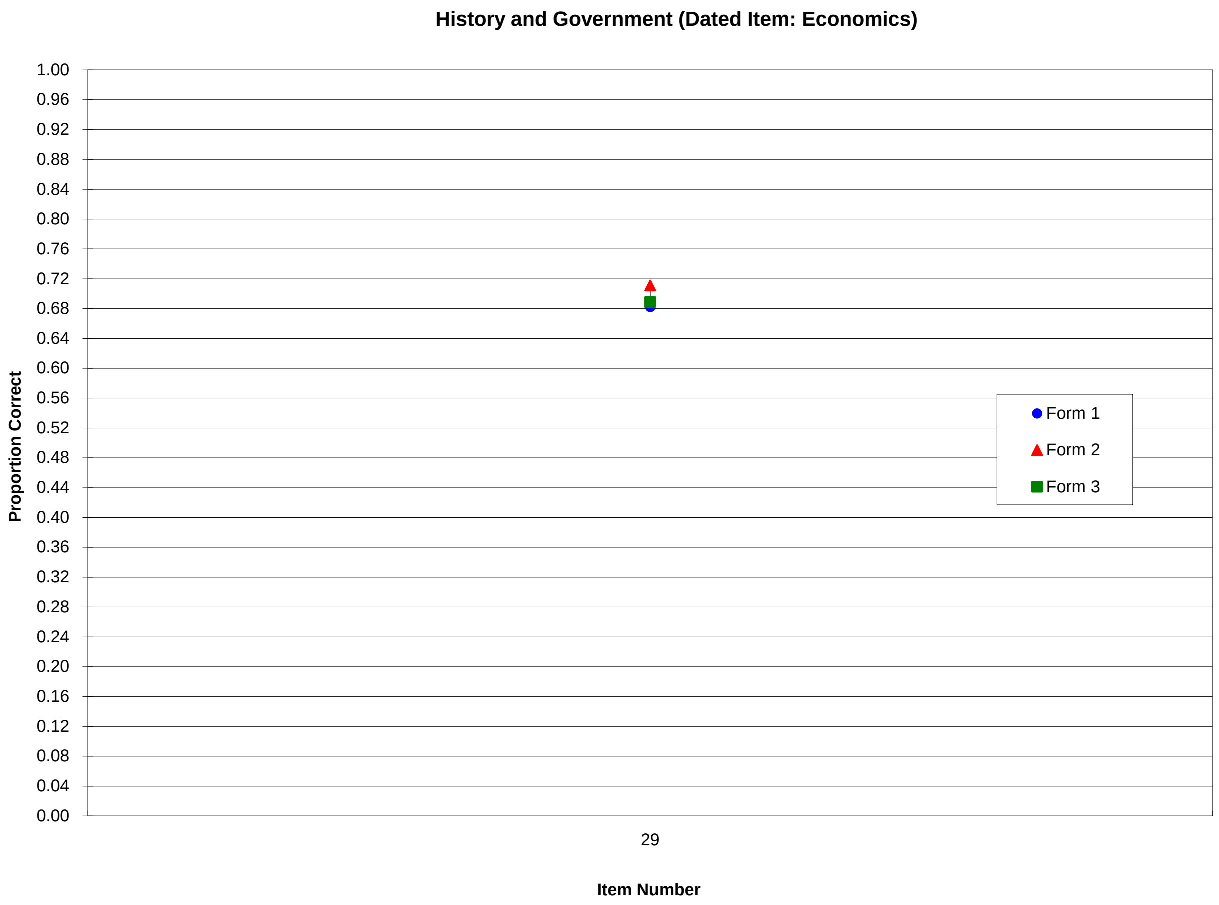

Figure 5. Dated Item in Economics: Plot of Item Proportion Correct (ordered from lowest to highest) by Test Form.

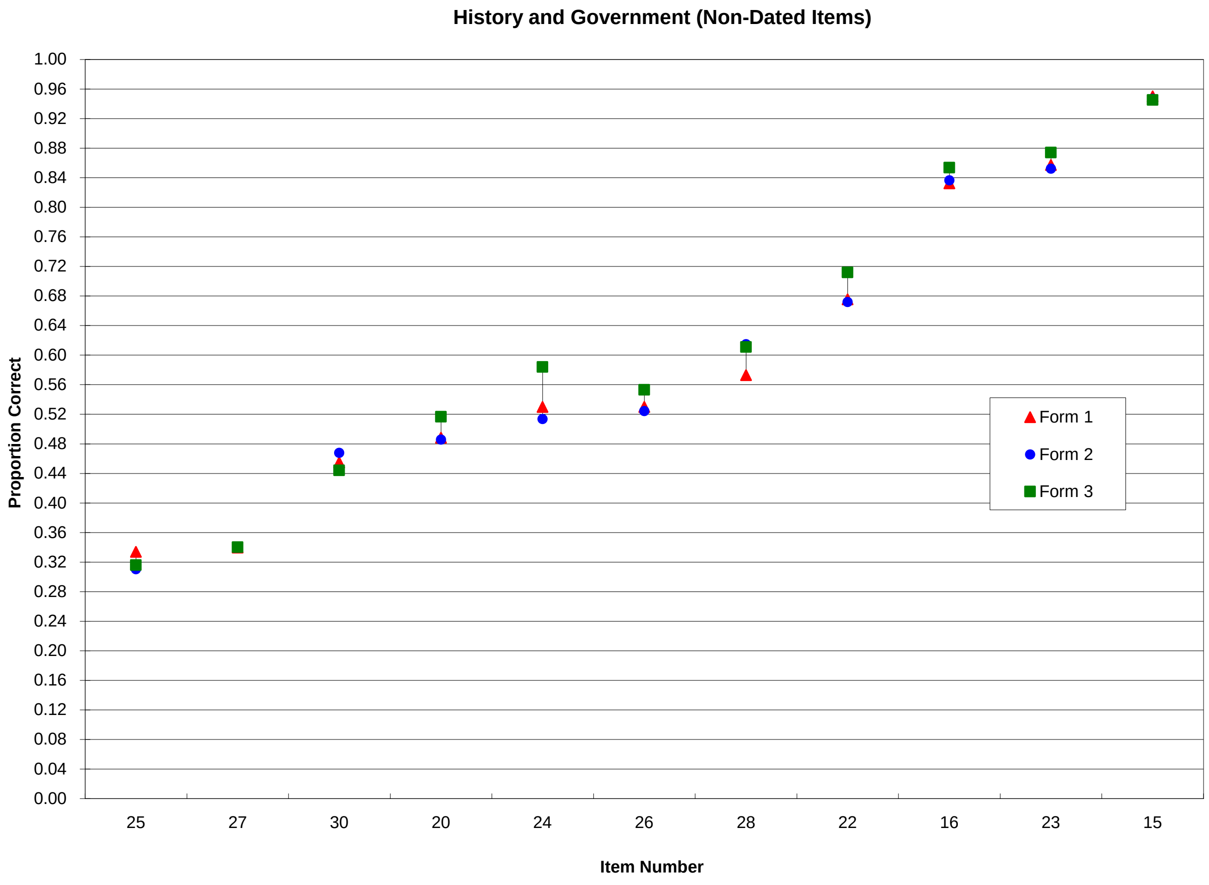

Figure 6. Non-Dated Items: Plot of Item Proportion Correct (ordered from lowest to highest) by Test Form.