Robust IRT Scaling:
Considerations in constructing item bank from tests across years

Paper presented at the annual meeting of
the National Council on Measurement in Education,
San Antonio, TX

April 28, 2017

Jungnam Kim, NWEA
Furong Gao, Pacific Metrics
Dong-In Kim, Data Recognition Corporation

Email: jungnam.kim@nwea.org

Introduction

In moving a national licensure exam from its traditional classical testing approaches to item response theory (IRT), one key step is to calibrate and bank all items cumulated over years on the same scale for the continuation of their future use. There are many psychometric issues to be considered in constructing the item bank: base year of scaling, calibration method, scaling and equating errors, etc. One concern is with item parameter drift, as the magnitude of drift may increase and cumulate over time (Kim & Cohen, 1992; Lautenschlager & Park, 1988; Shepard, Camilli, & Williams, 1984). Would different scaling methods cause different parameter drift? If so, what factors need to be considered in selecting the optimal scaling approach? Furthermore, how do the changes in the test designs over the years affect the scaling and equating? Using data collected in test administrations from 1992 to 2014, this study evaluates results from various scaling designs to address these questions.

The items are accumulated from tests across 23 years. Over the years, there have been some changes in the test blueprint and the composition of the item types. In 2009, the blueprints of the exam for nine domains were changed by 0%-4% (Table 1). As for the item type, old test forms were composed of 100 multiple-choice (MC) items and 25 cases which had 4 associated MC items, while new forms were composed of 110 MC items and 10 cases which have 3 polytomous items with 3-points each.

Given that tests share overlapping items, concurrent calibration across administrations seemed reasonable, as equating error or drift would likely be controlled in a single run. However, changes in the test blueprint and item type might introduce multidimensionality in the data. Research studies have shown that concurrent calibration is more accurate when the data fit the IRT model (Kim & Cohen, 1998; Hanson & Béguin, 2002) but less robust to violations of the IRT assumptions due to multidimensionality (Béguin, Hanson, & Glas, 2000; Béguin & Hanson, 2001). An alternative approach was to conduct separate calibrations but with as few equating transformation as possible, by transforming multiple tests simultaneously. This study investigated the impact of three different calibration and equating methods: concurrent calibration, separate calibration with chain equating transformation, and separate calibration with simultaneous equating transformation.

Data

The study constructed an item bank from tests in a national licensure examination administered from 1992 through 2014. The test was given twice a calendar year, in spring (S) and fall (F). Total of 45 tests were equated to the bank, using a common-item nonequivalent groups design.

Methods

Given that the test blueprint and item types have been changed since the 2009F (2009 fall) test, we organized the tests into two blocks: the first with all items in tests from the 2009F through 2014F administrations, and the second block with all items in tests prior to the 2009F administration (1992F through 2009S). As illustrated in Figures 1 – 3, the item bank was constructed in such a way that the first block tests were scaled and equated into the bank first, then the items in the second block were added into the bank. Dichotomous items were calibrated using the three-parameter logistic model (Lord, 1980) and polytomous items were calibrated using the graded response model (Samejima, 1969).

In the concurrent calibration (CR), item responses across tests within a block/sub-block were combined and their parameters were estimated in a single calibration run, as illustrated in Figure 1. Three sub-blocks within the second block were created because a single concurrent calibration failed to run. Items were put to the same scale via the common items between blocks.

In the separate calibration, the item parameters in each test were estimated separately. The chain equating transformation (CH; Figure 2) equated the tests one after another sequentially, as indicated by the arrows in Figure 2. That is, items from the 2014F test were calibrated and put to the bank first. Then at each subsequent step, the number of common items between the bank at that point and each of the remaining tests was checked. The test that shared the highest number of common items with the bank was equated and added to the bank. This was done first for the tests in the first block, then for the tests in the second block till all the tests were added to the bank. The order of the equating and the number of common items at each step, are presented in Figure 2. In the simultaneous equating transformation (SM; Figure 3), once all the tests in the first block were equated to the bank using the chain equating approach, the tests in the second block that had at least 20 common items with the bank were grouped together and

equated to the bank simultaneously. As illustrated in Figure 3, tests in the second block were added to the bank in two steps. In the CH and SM approaches, the common items were evaluated and those with a difference of item parameter estimates larger than the 2 RMSD value were excluded from the anchor set.

To examine the impact of the different scaling and equating methods, examinees' scale scores and passing rates were calculated and compared. Two scoring methods were used to derive the examinees' scale scores: IRT pattern scoring and summed scoring (i.e. scoring table). The two different scoring methods allowed us to examine if one was more robust than the other to item parameter drift and equating error accumulation over multiple years of administrations. At the item level, for each dichotomous item, the root mean squared differences (RMSD) and the difference between two item characteristic curves (ICC) were computed and evaluated using Raju's area measure (Raju, 1988) for each combination of equating methods.

Results

Figure 4 presents the mean scale scores of all the tests in administrations from 1992F to 2014F based on the six different combinations of the equating and scoring methods: three equating procedures (CH, SM, and CR for chained equating transformation, simultaneous equating transformation, and concurrent calibration, respectively), and two scoring methods (S and P for summed and pattern scoring, respectively). Specifically, the CH_S line represents the mean scale scores derived from the chained equating transformation with summed scoring; The CH_P line represents the results of chained equating transformation with pattern scoring; The SM_S line represents the results of simultaneous equating transformation with summed scoring; The SM_P line represents the results of simultaneous equating transformation with pattern scoring; The CR_S line represents the results of concurrent calibration with summed scoring; The CR_P line represents the results of concurrent calibration with pattern scoring). The two lines from the two different scoring methods with the same equating procedure were shown with the same color but different line patter: the solid line for the summed scoring and the dotted line for the pattern scoring. This line and color representation were also used for Figure 5, which shows the percentage of failing examinees. Table 2 reports the scale score summary statistics from the summed soring method, including examinee N-counts, mean, standard deviation (SD),

and the percentage of students who failed the test. Table 3 reports the same scale score summary statistics when pattern scoring was used.

The scale score means fluctuated across administrations, as shown in Figure 4. The comparison of the solid and dotted lines of a same color indicated that the score means were very similar between the two scoring methods regardless of equating procedure. By comparing the three different colored lines, we noted that the score means were similar between CH and SM, but means obtained from the CR calibration were different from these two equating transformations. With the CR method, the score means across years were relatively less fluctuated and consistently lower, more so with the older tests, than those from the separate calibration methods. With the separate calibration, the means of the older tests (i.e. prior to 2002F) fluctuated more than the other tests.

The percentage of failing examinees shown in Figure 5 displays similar patterns as in the scale score means: results from scoring methods within an equating procedure were very similar; the two equating transformation following the separate calibration, CH and SM, yielded similar results; CR produced relatively less fluctuation and higher percentage of failing examinees; larger differences were observed between CR and the two equating transformations for the older tests than the middle (i.e. tests for 2002F and 2009S) or the newer tests (i.e. 2009F and later tests).

The item-level analysis results are reported in Tables 4– 6. Table 4 presents the number of items flagged based on the RMSD for the comparison: CH vs SM, CH vs CR, and SM vs CR. Note that the item parameter estimates for tests 2009F through 2014F were the same in CH and SM, because these two transformations were applied to the tests prior to 2009F only. For tests prior to 2009F, two equating transformations following the separate calibration, CH and SM, yielded similar results, thus less items were flagged, while many items were flagged when item parameter estimates from CR were compared. With two exceptions of the tests 2005S and 2006S, a maximum of 12 items (7%) were flagged when comparing CH and SM. Between CH and CR and between SM and CR, more items were flagged in general (up to 11% of items). It was somewhat expected considering large differences in the scale score statistics with the CR results.

ICCs were compared using Raju's area measure. Items with 0.1 or higher weighted difference were flagged, following the criterion used for PARCC. As presented in Table 5, no

items were flagged between CH and SM but more than half of the items were flagged when items from CR were compared to CH or SM. However, for the newer tests (i.e. 2009F and later tests), only one item was flagged.

Table 6 reported the number of flagged items for both the RMSD and ICC area measure. To see if the number of flagged items were related to the order in which a test was equated to the bank, the order of equating is also included in Tables 4-6. The cells with relatively large difference in the equating order are highlighted in grey. With respect to the RMSD, Table 4 does not show a clear pattern between the equating order and the number of flagged items. For example, the two administrations with the highest number of flagged items between CH and SM were 2005S and 2006S with the equating order of the 21st and 14th in CH and the 10th- order group in SM. With the ICC comparison, Table 5 does not show any clear patter either. More than half of items were flagged by the area measure across administrations. To see if the number of flagged items were related to any content-specific properties, items flagged for both RMSD and area measure, as in Table 6, were further reviewed for any potential content properties, but no meaningful pattern was found.

Summary and Discussions

This study investigated the impact of three different equating methods (CH, SM, and CR) on constructing an item bank using items accumulated over the years. At the test-level, examinees' scale scores and passing rates were compared using both summed scoring and IRT pattern scoring. At the item level, the RMSD and the difference between two ICCs were computed using Raju's area measure were examined for all dichotomous items.

The analyses indicated that scoring methods within an equating procedure produced very similar result. The two equating transformation methods following a separate calibration, CH and SM, yielded similar results, while the CR method produce somewhat different results. The differences between CH/SM and CR were much larger for the older tests than for the relatively recent tests. At the item level, the differences in item parameter estimates were much larger between CH/SM and CR than between CH and SM.

The results may be impacted by the order of each test was scaled to the bank and which set of the item parameter estimates, for the common items, was kept in the bank. Although tests were scaled into the bank in non-chronological order, the parameter estimates (after equating)

from the more recent administrations were always kept in the bank and used in subsequent scoring analyses. That is, for example, the test 2006S was scaled into the bank before the test 2008S (i.e. the CH equating order was 14 and 20 for 2006S and 2008S, respectively). the common items between 2006S and 2008S, however, their parameter estimates from 2008S were kept in the bank as the 2008S estimates were more recent than those from 2006S. Considering that many items were re-used in later administrations, if the older estimates were used in the subsequent scoring analysis, results may be different. Furthermore, a study to investigate why the concurrent calibration produced such different results would be necessary. It would also be interesting to examine the results if the concurrent calibration can be carried out in one run for the second block tests as intended.

References

Béguin, A. A., & Hanson, B. A. (2001, April). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000, April*). Effect of multidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hanson, Bradley A. & Béguin, Anton A. (2002). Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in The Common-Item Equating Design. *Applied Psychological Measurement*. V26, N1.

Kim, S. -H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29,* 51-66.

Kim, S. -H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22,* 131-143.

Lautenschlager, G. J., & Park, D.-G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement, 12,* 365-376.

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

Raju, N. S. (1988). The area between two item characteristic cureves. *Psychometrika,* 53, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement,* 14, 197-207.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 17.

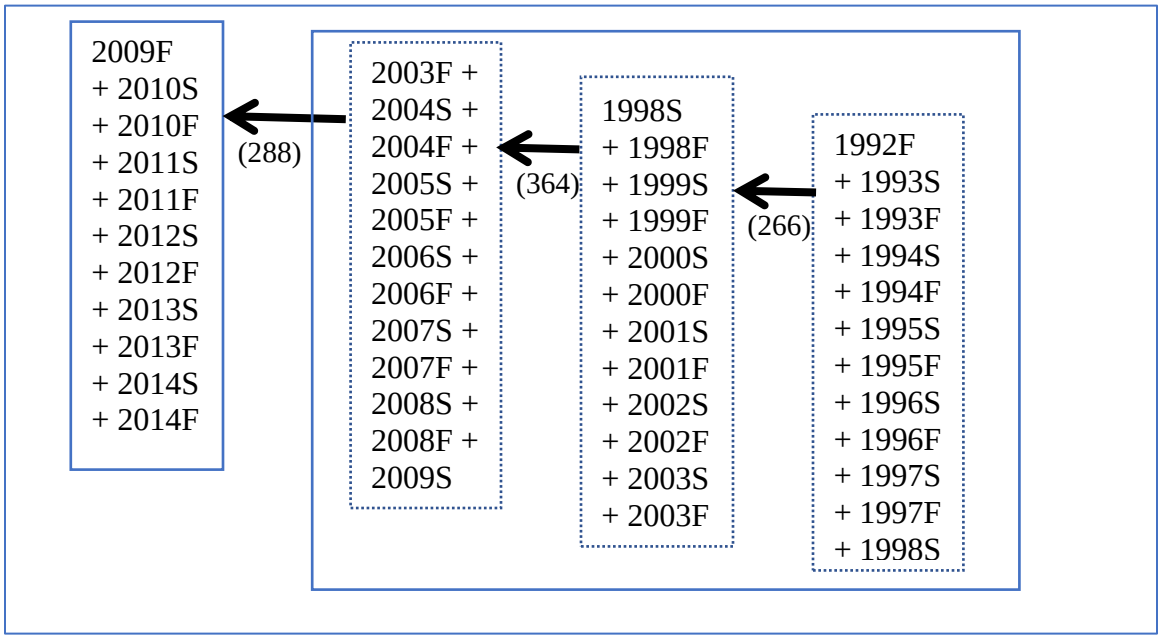
Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Measurement, 9,* 93-128.

Tables and Figures

Table 1: Blueprint change in 2009

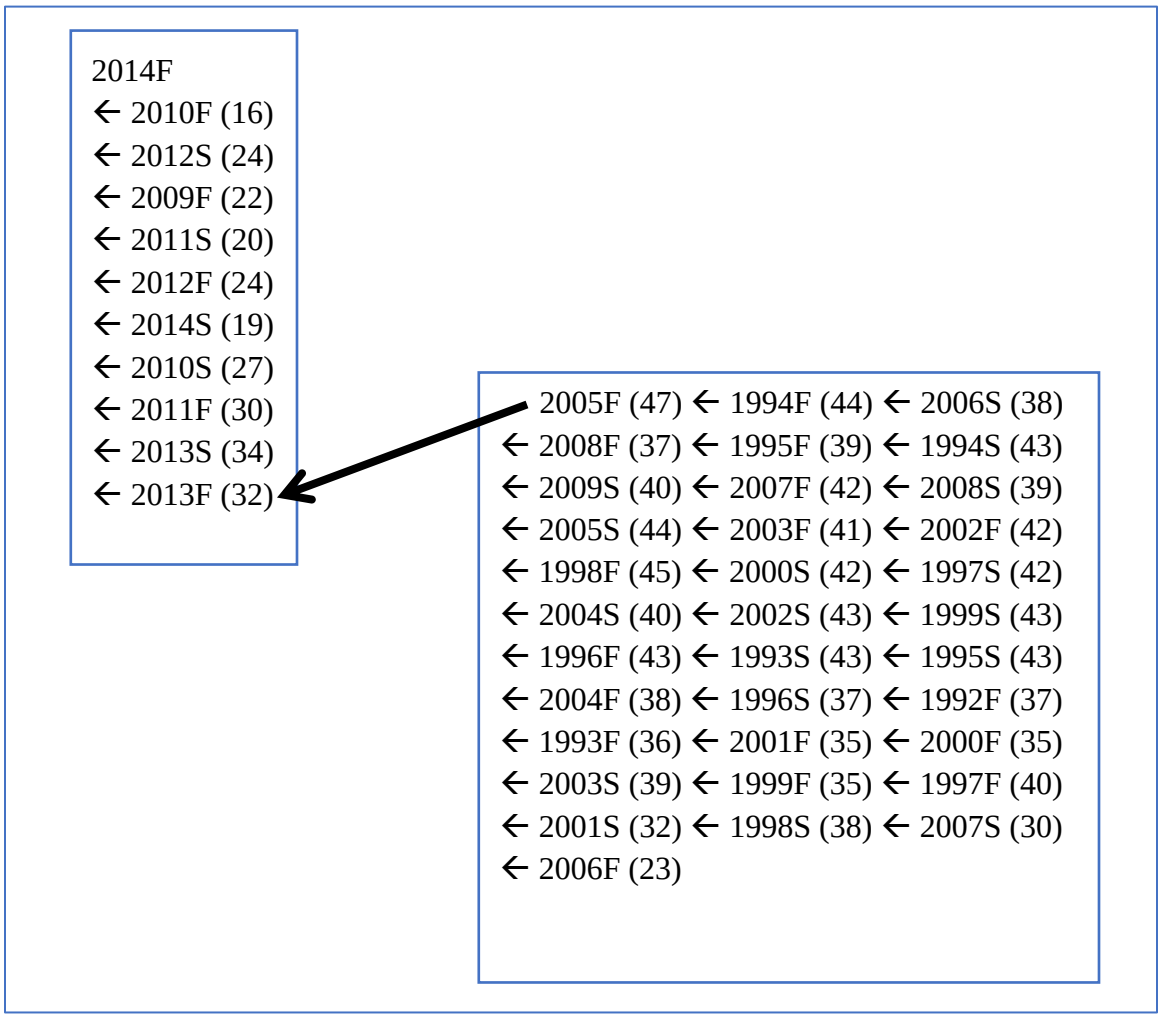
Domain	Percent in the old test	Percent in the new test	Percent Difference
1	16%	11%	-4%
2	11%	9%	-2%
3	13%	11%	-2%
4	11%	11%	0%
5	6%	7%	1%
6	12%	14%	2%
7	11%	14%	3%
8	8%	8%	0%
9	12%	15%	3%

Figure 1: Concurrent calibration (CR)



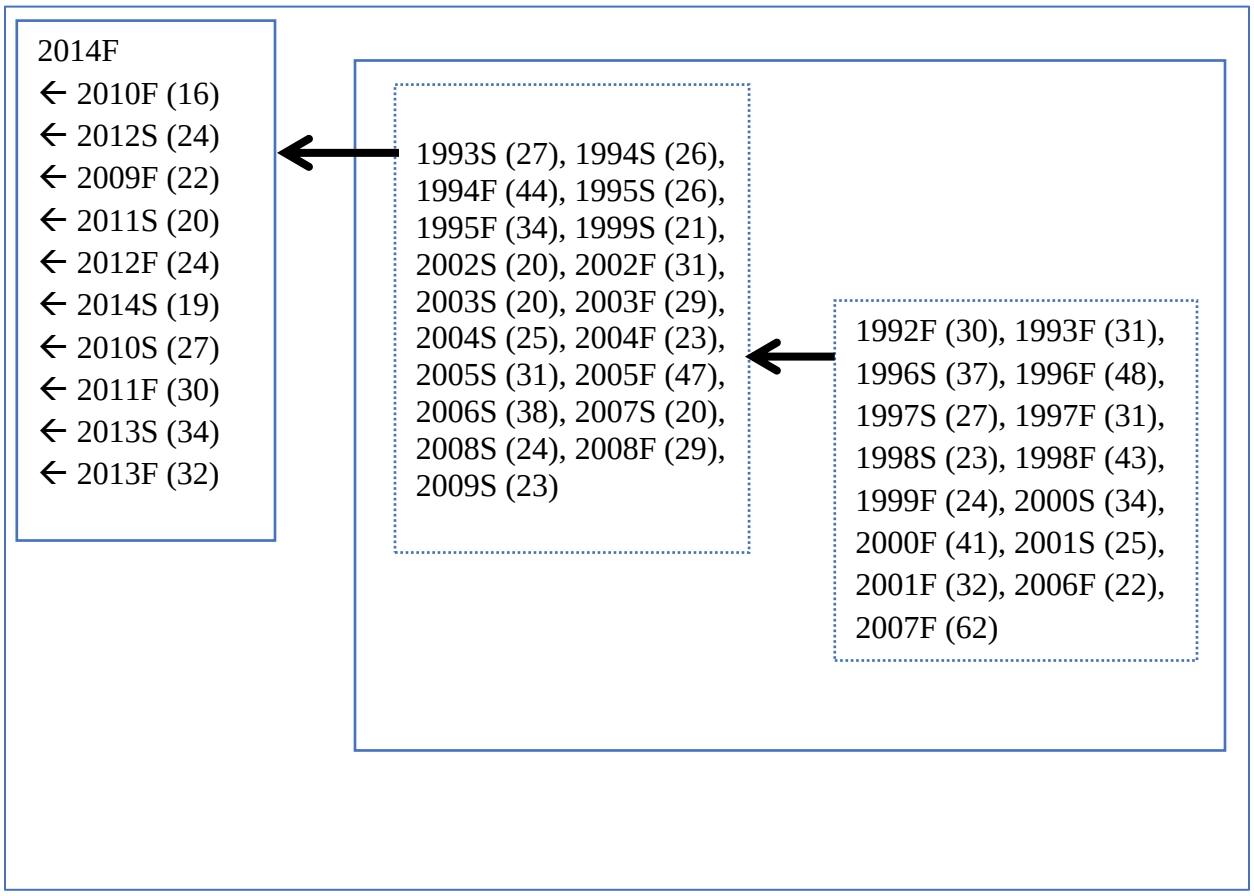
Note: In the concurrent calibration, item parameters for all tests within a sub-block were estimated in a single calibration run. First block remained a sub-block but second block was divided into three sub-blocks so that each sub-block had about 10 exams. Responses of the same items across tests within a sub-block were arranged on the same locations, and items were calibrated with all responses across years. Parameters from the four sub-block concurrent calibrations were put to the same scale via the common items between the two sub-blocks. The number in the parenthesis indicates the number of items. That is, 288, 364, and 266 items were used to equate the sub- blocks to the first block in sequence, respectively.

Figure 2: Separate calibration with chain transformation (CH)



Note: In the chain equating transformation, each test was equated one after another sequentially. For example, the last administration, 2014F, was calibrated first to establish the IRT base scale and put in the item bank. Next, from the first block, the test which had the most number of common items with the bank (which has the 2014F test), 2010F, was calibrated and equated to 2014F via the common items. Then, from the first block, the test that had the most number of common items with the bank (now has two tests), 2012S, was equated and added to the bank, and so on until all the tests in the first block were added to the bank. This process was then repeated for the each of the tests in the second block. The number of common items of each transformation was in the parenthesis in the figure above.

Figure 3: Separate calibration with simultaneous transformation (SM)



Note: In the simultaneous transformation, multiple tests were equated to the bank simultaneously, if there were at least 20 common items between each test and the bank. In the first block, there were not enough common items to apply this method. Therefore, this method was applied to the second block only. After constructing the bank for the first block with chain transformation, the common items between the bank (now has 11 tests) and 34 remaining tests were checked. A total of 19 tests had at least 20 common items with the first block. These 19 tests were equated in the first step of second block, using common items of each test. The number of common items of each equating is in the parenthesis in the figure above. Then, the common items between the bank (now has 30 tests) and 15 remaining tests were checked. As all the tests had more than 20 common items with the bank, the rest of tests were equated to the bank in the second step of the second block (e.g. the lowest number of common items was 22 with 2006F, while the highest number was 62 common items with 2007F). Thus, second block was added to the bank in two steps.

Figure 4: Scale score means for 1992F through 2014F

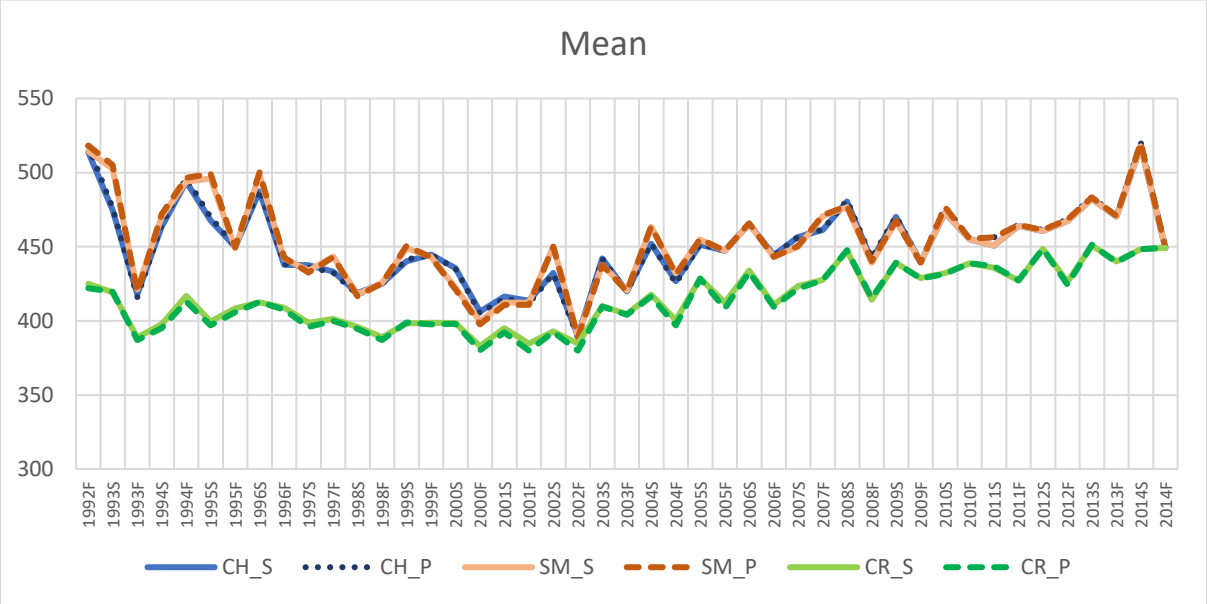


Figure 5: The percentage of failing examinees for 1992F through 2014F

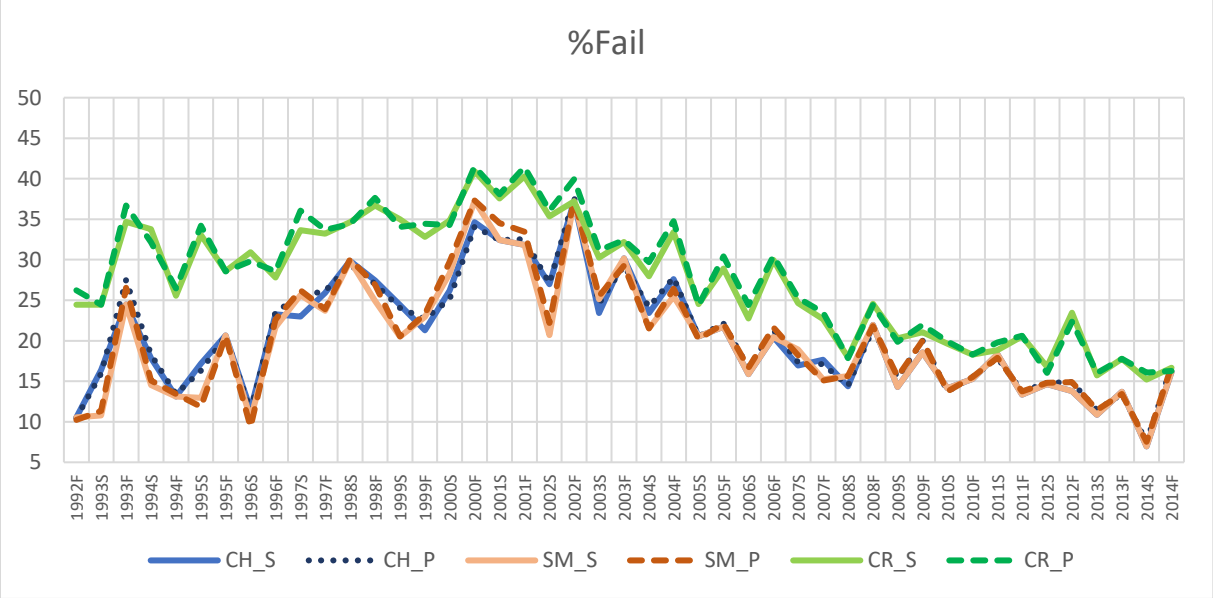


Table 2: Descriptive statistics with summed scoring

Admin	N	CH_S			SM_S			CR_S		
		M	SD	%Fail	M	SD	%Fail	M	SD	%Fail
1992F	1575	513.72	132.52	10.54	514.37	130.48	10.54	424.92	116.98	24.44
1993S	1280	474.71	131.67	16.41	502.50	132.52	10.78	419.10	111.34	24.45
1993F	1499	418.01	111.88	24.48	421.37	112.07	24.48	388.74	106.98	34.69
1994S	1306	463.49	126.20	17.84	470.16	122.61	14.55	397.70	107.62	33.77
1994F	1569	493.92	134.58	13.13	493.92	134.58	13.13	416.85	112.12	25.56
1995S	1374	467.87	129.88	17.18	496.01	133.01	12.95	399.54	115.10	33.04
1995F	1906	449.85	130.66	20.67	449.02	129.83	20.67	408.49	113.52	28.59
1996S	1723	487.37	116.03	11.55	498.42	116.28	10.39	412.46	115.97	30.93
1996F	2206	437.77	126.20	23.25	442.04	126.53	21.67	408.77	110.43	27.79
1997S	1928	437.55	119.76	22.98	433.79	118.46	25.62	398.64	114.70	33.66
1997F	1992	433.23	122.38	25.90	443.36	122.06	23.69	401.19	116.27	33.23
1998S	2136	418.79	122.51	29.87	417.84	124.52	29.87	395.80	112.73	34.64
1998F	1995	424.98	120.88	27.47	425.62	120.32	24.91	389.04	114.34	36.69
1999S	2205	440.18	126.28	24.40	448.95	123.06	20.50	398.15	116.78	35.01
1999F	1756	444.65	123.24	21.30	443.25	124.89	22.95	398.74	112.92	32.80
2000S	2135	435.60	124.66	26.18	422.42	126.29	28.29	398.56	114.14	34.85
2000F	1769	406.30	120.76	34.65	399.49	123.17	37.25	382.99	112.03	40.93
2001S	1871	416.41	127.67	32.44	412.24	130.15	32.44	395.11	117.46	37.57
2001F	1405	413.63	127.93	31.81	412.34	130.38	31.81	384.64	113.79	40.28
2002S	1695	432.45	127.38	26.96	450.05	126.93	20.71	392.96	114.82	35.34
2002F	1347	389.92	126.05	37.19	389.82	125.44	37.19	384.59	114.54	37.19
2003S	1418	442.00	125.11	23.41	437.80	125.51	25.11	409.97	118.39	30.25
2003F	1209	419.95	125.63	30.19	420.53	126.41	30.19	404.57	115.34	32.18
2004S	1435	452.09	138.63	23.41	463.46	137.18	21.53	417.72	118.31	27.94
2004F	1068	426.71	121.60	27.62	432.05	120.84	25.47	400.47	109.99	33.43
2005S	1431	451.32	119.81	20.55	454.87	121.41	20.55	428.60	111.08	24.53
2005F	1193	447.16	126.26	21.71	447.16	126.26	21.71	412.39	112.99	28.92
2006S	1410	465.34	118.08	15.89	465.34	118.08	15.89	433.88	110.48	22.77
2006F	998	444.35	111.84	20.44	443.24	111.92	20.44	410.82	108.84	29.96
2007S	1368	456.56	115.93	16.96	450.17	113.16	18.93	423.44	109.25	24.63
2007F	1094	461.21	121.06	17.64	470.51	118.86	15.27	427.55	107.25	22.67
2008S	1412	480.41	119.30	14.38	476.66	123.49	15.65	447.37	107.74	17.99
2008F	1184	440.84	126.28	21.96	439.07	125.88	21.96	414.30	109.02	24.58
2009S	1439	470.14	114.64	14.32	467.18	112.37	14.32	439.00	104.70	20.29
2009F	1088	439.70	105.72	18.57	439.70	105.72	18.57	428.89	100.24	21.05
2010S	1427	472.93	117.91	14.16	472.93	117.91	14.16	432.15	101.66	19.62
2010F	1133	455.01	105.75	15.27	455.01	105.75	15.27	438.82	100.71	18.27
2011S	1370	450.67	110.71	18.32	450.67	110.71	18.32	435.82	102.15	18.83
2011F	1018	464.07	110.75	13.36	464.07	110.75	13.36	427.29	99.79	20.53
2012S	1298	460.75	106.44	14.64	460.75	106.44	14.64	448.57	100.66	16.80
2012F	886	467.38	115.95	13.77	467.38	115.95	13.77	426.43	104.82	23.48
2013S	1317	482.19	111.72	10.86	482.19	111.72	10.86	450.88	101.25	15.72
2013F	991	470.43	107.05	13.72	470.43	107.05	13.72	440.14	98.28	17.76
2014S	1394	517.74	117.32	6.96	517.74	117.32	6.96	448.45	100.38	15.21
2014F	1021	449.09	104.22	16.16	449.09	104.22	16.16	449.19	102.96	16.65

Table 3: Descriptive statistics with pattern scoring

Admin	N	CH_S			SM_S			CR_S		
		M	SD	%Fail	M	SD	%Fail	M	SD	%Fail
1992F	1575	518.08	137.66	10.48	518.35	135.01	10.22	421.94	120.50	26.22
1993S	1280	476.20	135.05	15.94	505.13	136.13	11.33	419.62	112.70	24.45
1993F	1499	415.86	116.74	27.48	419.26	117.02	26.55	386.96	109.57	36.69
1994S	1306	464.72	128.75	18.30	471.93	124.35	15.08	395.07	107.47	32.16
1994F	1569	496.54	135.84	13.45	496.54	135.84	13.45	412.98	112.34	26.45
1995S	1374	469.84	132.14	16.30	499.38	135.39	11.86	396.92	115.73	34.21
1995F	1906	450.26	131.68	20.57	449.72	131.19	20.57	405.80	112.47	28.54
1996S	1723	488.65	118.41	11.38	500.10	118.78	9.34	412.64	116.61	29.83
1996F	2206	438.37	127.25	23.57	442.66	127.58	22.62	407.45	109.40	28.56
1997S	1928	436.28	124.42	25.52	432.54	122.63	26.24	395.91	116.31	36.05
1997F	1992	432.52	126.69	26.15	443.01	126.32	23.90	400.02	118.58	33.68
1998S	2136	417.60	125.64	29.40	416.62	127.97	30.06	394.56	114.46	34.41
1998F	1995	424.92	123.13	27.02	425.56	122.47	26.92	387.06	115.46	37.64
1999S	2205	441.49	127.59	24.04	450.61	123.79	20.54	398.90	116.32	34.06
1999F	1756	444.50	126.31	22.72	442.97	128.42	23.29	397.64	114.81	34.45
2000S	2135	435.46	127.50	25.11	421.32	129.62	29.74	398.15	114.57	34.29
2000F	1769	405.01	123.38	34.09	397.54	126.11	37.31	380.21	112.87	41.49
2001S	1871	415.18	130.77	32.44	410.75	133.71	34.53	392.31	119.03	38.05
2001F	1405	411.89	131.59	32.53	410.84	134.98	33.45	379.90	115.14	41.42
2002S	1695	431.81	130.51	27.08	449.97	129.95	22.18	392.37	114.58	36.05
2002F	1347	388.62	127.50	37.71	388.65	127.01	37.49	379.97	114.05	39.94
2003S	1418	441.86	127.40	24.61	437.78	128.20	25.53	409.71	118.92	31.17
2003F	1209	419.58	128.08	29.28	419.93	128.76	29.28	404.10	116.37	32.42
2004S	1435	452.21	141.84	24.18	463.75	140.34	21.53	416.30	119.70	29.69
2004F	1068	425.35	124.74	27.81	431.39	123.62	26.40	397.00	111.12	34.74
2005S	1431	451.77	122.21	20.55	455.21	124.03	20.20	428.51	112.29	24.39
2005F	1193	447.18	128.59	22.13	447.18	128.59	22.13	408.39	113.67	30.43
2006S	1410	465.72	120.28	16.67	465.72	120.28	16.67	432.81	111.28	24.40
2006F	998	443.85	113.44	21.34	443.10	113.63	21.64	409.42	109.23	30.46
2007S	1368	456.70	117.64	17.32	450.45	114.59	18.20	421.86	109.73	25.29
2007F	1094	461.71	123.23	17.09	471.26	120.71	15.08	427.31	107.27	23.49
2008S	1412	481.40	121.35	14.45	477.51	125.69	15.65	447.72	108.48	17.85
2008F	1184	442.21	125.81	21.45	440.36	125.49	21.79	415.41	108.21	24.41
2009S	1439	470.62	116.70	15.43	467.65	114.31	15.57	439.29	105.49	19.87
2009F	1088	439.30	106.36	20.04	439.30	106.36	20.04	428.39	100.94	21.97
2010S	1427	476.77	120.10	13.74	476.77	120.10	13.74	431.70	102.19	19.90
2010F	1133	455.36	106.64	15.53	455.36	106.64	15.53	438.84	101.48	18.27
2011S	1370	456.23	114.62	17.88	456.23	114.62	17.88	436.92	103.24	19.78
2011F	1018	464.70	112.11	13.75	464.70	112.11	13.75	427.41	100.70	20.63
2012S	1298	461.24	107.39	14.79	461.24	107.39	14.79	448.58	100.90	16.02
2012F	886	468.42	118.42	14.90	468.42	118.42	14.90	424.43	104.88	22.35
2013S	1317	483.42	112.49	11.47	483.42	112.49	11.47	451.36	102.95	16.02
2013F	991	470.89	107.40	13.42	470.89	107.40	13.42	439.43	98.43	17.76
2014S	1394	520.24	118.81	7.53	520.24	118.81	7.53	448.37	100.93	16.07
2014F	1021	449.35	104.99	16.85	449.35	104.99	16.85	449.36	104.43	16.26

Table 4: Equating order and flagged items for RMSD

Admin	Equating Order			Total No. of Items	Flagged Items					
					CH vs. SM		CH vs CR		SM vs CR	
	CH	SM	CR		No.	%	No.	%	No.	%
1992F	35	11	4	152	6	3.95%	15	9.87%	13	8.55%
1993S	31	10	4	181	1	0.55%	8	4.42%	9	4.97%
1993F	36	11	4	122	6	4.92%	11	9.02%	11	9.02%
1994S	17	10	4	149	10	6.71%	12	8.05%	15	10.07%
1994F	13	10	4	167	12	7.19%	9	5.39%	9	5.39%
1995S	32	10	4	161	6	3.73%	12	7.45%	13	8.07%
1995F	16	10	4	159	7	4.40%	13	8.18%	13	8.18%
1996S	34	11	4	160	3	1.88%	11	6.88%	10	6.25%
1996F	30	11	4	165	4	2.42%	17	10.30%	17	10.30%
1997S	26	11	4	161	7	4.35%	14	8.70%	15	9.32%
1997F	41	11	4	164	2	1.22%	13	7.93%	13	7.93%
1998S	43	11	3/4	173	7	4.05%	12	6.94%	11	6.36%
1998F	24	11	3	175	9	5.14%	18	10.29%	18	10.29%
1999S	29	10	3	174	8	4.60%	19	10.92%	17	9.77%
1999F	40	11	3	172	6	3.49%	13	7.56%	12	6.98%
2000S	25	11	3	175	3	1.71%	16	9.14%	12	6.86%
2000F	38	11	3	170	10	5.88%	10	5.88%	10	5.88%
2001S	42	11	3	179	8	4.47%	4	2.23%	4	2.23%
2001F	37	11	3	163	5	3.07%	9	5.52%	8	4.91%
2002S	28	10	3	169	1	0.59%	13	7.69%	14	8.28%
2002F	23	10	3	178	11	6.18%	13	7.30%	14	7.87%
2003S	39	10	3	177	2	1.13%	16	9.04%	16	9.04%
2003F	22	10	2/3	176	9	5.11%	13	7.39%	13	7.39%
2004S	27	10	2	181	1	0.55%	13	7.18%	12	6.63%
2004F	33	10	2	171	4	2.34%	14	8.19%	15	8.77%
2005S	21	10	2	181	15	8.29%	14	7.73%	14	7.73%
2005F	12	10	2	180	9	5.00%	13	7.22%	13	7.22%
2006S	14	10	2	188	16	8.51%	17	9.04%	17	9.04%
2006F	45	11	2	181	3	1.66%	8	4.42%	8	4.42%
2007S	44	10	2	183	9	4.92%	12	6.56%	12	6.56%
2007F	19	11	2	180		0.00%	15	8.33%	15	8.33%
2008S	20	10	2	180	5	2.78%	13	7.22%	8	4.44%
2008F	15	10	2	177	4	2.26%	12	6.78%	12	6.78%
2009S	18	10	2	175	8	4.57%	10	5.71%	10	5.71%
2009F	4		1	102			10	9.80%	10	9.80%
2010S	8		1	101			7	6.93%	7	6.93%
2010F	2		1	105			8	7.62%	8	7.62%
2011S	5		1	99			8	8.08%	8	8.08%
2011F	9		1	99			7	7.07%	7	7.07%
2012S	3		1	93			8	8.60%	8	8.60%
2012F	6		1	96			5	5.21%	5	5.21%
2013S	10		1	98			8	8.16%	8	8.16%
2013F	11		1	93			6	6.45%	6	6.45%
2014S	7		1	94			0	0.00%	0	0.00%
2014F (base)	1		1	102			8	7.84%	8	7.84%

Table 5: Equating order and flagged items for area measure

Admin	Equating Order			Total No. of Items	Flagged Items					
					CH vs. SM		CH vs CR		SM vs CR	
	CH	SM	CR		No.	%	No.	%	No.	%
1992F	35	11	4	152	0	0	86	56.58%	89	58.55%
1993S	31	10	4	181	0	0	110	60.77%	112	61.88%
1993F	36	11	4	122	0	0	64	52.46%	63	51.64%
1994S	17	10	4	149	0	0	89	59.73%	89	59.73%
1994F	13	10	4	167	0	0	103	61.68%	103	61.68%
1995S	32	10	4	161	0	0	107	66.46%	110	68.32%
1995F	16	10	4	159	0	0	92	57.86%	92	57.86%
1996S	34	11	4	160	0	0	88	55.00%	91	56.88%
1996F	30	11	4	165	0	0	104	63.03%	104	63.03%
1997S	26	11	4	161	0	0	99	61.49%	98	60.87%
1997F	41	11	4	164	0	0	77	46.95%	78	47.56%
1998S	43	11	3/4	173	0	0	88	50.87%	88	50.87%
1998F	24	11	3	175	0	0	104	59.43%	104	59.43%
1999S	29	10	3	174	0	0	106	60.92%	105	60.34%
1999F	40	11	3	172	0	0	107	62.21%	107	62.21%
2000S	25	11	3	175	0	0	102	58.29%	99	56.57%
2000F	38	11	3	170	0	0	87	51.18%	87	51.18%
2001S	42	11	3	179	0	0	108	60.34%	108	60.34%
2001F	37	11	3	163	0	0	98	60.12%	97	59.51%
2002S	28	10	3	169	0	0	91	53.85%	94	55.62%
2002F	23	10	3	178	0	0	102	57.30%	102	57.30%
2003S	39	10	3	177	0	0	83	46.89%	85	48.02%
2003F	22	10	2/3	176	0	0	87	49.43%	87	49.43%
2004S	27	10	2	181	0	0	82	45.30%	83	45.86%
2004F	33	10	2	171	0	0	103	60.23%	103	60.23%
2005S	21	10	2	181	0	0	98	54.14%	98	54.14%
2005F	12	10	2	180	0	0	103	57.22%	103	57.22%
2006S	14	10	2	188	0	0	109	57.98%	109	57.98%
2006F	45	11	2	181	0	0	101	55.80%	101	55.80%
2007S	44	10	2	183	0	0	108	59.02%	107	58.47%
2007F	19	11	2	180	0	0	100	55.56%	103	57.22%
2008S	20	10	2	180	0	0	110	61.11%	109	60.56%
2008F	15	10	2	177	0	0	111	62.71%	112	63.28%
2009S	18	10	2	175	0	0	103	58.86%	103	58.86%
2009F	4		1	102			0	0.00%	0	0.00%
2010S	8		1	101			1	0.99%	1	0.99%
2010F	2		1	105			1	0.95%	1	0.95%
2011S	5		1	99			0	0.00%	0	0.00%
2011F	9		1	99			1	1.01%	1	1.01%
2012S	3		1	93			0	0.00%	0	0.00%
2012F	6		1	96			0	0.00%	0	0.00%
2013S	10		1	98			0	0.00%	0	0.00%
2013F	11		1	93			0	0.00%	0	0.00%
2014S	7		1	94			1	1.06%	1	1.06%
2014F (base)	1		1	102			0	0.00%	0	0.00%

Table 6: Equating order and flagged items for both RMSD and area measure

Admin	Equating Order			Total No. of Items	Flagged Items					
					CH vs. SM		CH vs CR		SM vs CR	
	CH	SM	CR		ni	%	ni	%	ni	%
1992F	35	11	4	152	0	0	12	7.89%	10	6.58%
1993S	31	10	4	181	0	0	7	3.87%	7	3.87%
1993F	36	11	4	122	0	0	5	4.10%	5	4.10%
1994S	17	10	4	149	0	0	7	4.70%	10	6.71%
1994F	13	10	4	167	0	0	6	3.59%	6	3.59%
1995S	32	10	4	161	0	0	7	4.35%	8	4.97%
1995F	16	10	4	159	0	0	11	6.92%	11	6.92%
1996S	34	11	4	160	0	0	6	3.75%	6	3.75%
1996F	30	11	4	165	0	0	10	6.06%	10	6.06%
1997S	26	11	4	161	0	0	10	6.21%	11	6.83%
1997F	41	11	4	164	0	0	5	3.05%	5	3.05%
1998S	43	11	3/4	173	0	0	9	5.20%	8	4.62%
1998F	24	11	3	175	0	0	9	5.14%	8	4.57%
1999S	29	10	3	174	0	0	13	7.47%	12	6.90%
1999F	40	11	3	172	0	0	12	6.98%	11	6.40%
2000S	25	11	3	175	0	0	9	5.14%	6	3.43%
2000F	38	11	3	170	0	0	6	3.53%	6	3.53%
2001S	42	11	3	179	0	0	2	1.12%	2	1.12%
2001F	37	11	3	163	0	0	5	3.07%	5	3.07%
2002S	28	10	3	169	0	0	8	4.73%	8	4.73%
2002F	23	10	3	178	0	0	7	3.93%	8	4.49%
2003S	39	10	3	177	0	0	11	6.21%	12	6.78%
2003F	22	10	2/3	176	0	0	9	5.11%	8	4.55%
2004S	27	10	2	181	0	0	4	2.21%	3	1.66%
2004F	33	10	2	171	0	0	9	5.26%	11	6.43%
2005S	21	10	2	181	0	0	7	3.87%	6	3.31%
2005F	12	10	2	180	0	0	8	4.44%	8	4.44%
2006S	14	10	2	188	0	0	11	5.85%	11	5.85%
2006F	45	11	2	181	0	0	4	2.21%	4	2.21%
2007S	44	10	2	183	0	0	10	5.46%	9	4.92%
2007F	19	11	2	180	0	0	9	5.00%	9	5.00%
2008S	20	10	2	180	0	0	7	3.89%	5	2.78%
2008F	15	10	2	177	0	0	8	4.52%	8	4.52%
2009S	18	10	2	175	0	0	4	2.29%	4	2.29%
2009F	4		1	102			0	0.00%	0	0.00%
2010S	8		1	101			1	0.99%	1	0.99%
2010F	2		1	105			1	0.95%	1	0.95%
2011S	5		1	99			0	0.00%	0	0.00%
2011F	9		1	99			0	0.00%	0	0.00%
2012S	3		1	93			0	0.00%	0	0.00%
2012F	6		1	96			0	0.00%	0	0.00%
2013S	10		1	98			0	0.00%	0	0.00%
2013F	11		1	93			0	0.00%	0	0.00%
2014S	7		1	94			0	0.00%	0	0.00%
2014F (base)	1		1	102			0	0.00%	0	0.00%