Creating Objective Measures of Examinee Speed and Item Length for the NCLEX-RN® Examination

December 2, 1997

Brian D. Bontempo, M. A.
The National Council of State Boards of Nursing

In the world of test design, the length of time allowed for test administration is often determined by financial and administrative concerns rather than psychometric rationale. In a historical sense, this was logical since data were not available to provide the psychometric rationale to guide this policy decision. However, with the advent of computerized testing, accurate item response time data are now available which provide the potential to determine the length of an item (in time rather than number of words), the speed of an examinee, or the speededness of a given examination. Once these pieces of information are determined, the psychometric rationale for enacting a given time limit follows naturally. In order to realize this potential, objective measures of examinee speed and item length must be devised.

The process of creating objective measures of examinee speed and item length was applied to the licensure examination for registered nurses. This examination, called the NCLEX-RN® examination, is a high-stakes, variable-length, computerized adaptive test which is administered to over 100,000 examinees each year. Eventually, the objective measures of examinee speed and item length will help to determine the speededness of the examination, the impact of speededness on examinee performance, and eventually the maximum length of time allowed for administration.

Data collection

In order to sample from the areas of the matrix that contain the least amount of missing data, item response times were collected from a random sample of examinees that had taken at least 120 items. To keep the sample as homogenous as possible, only first-time, US-educated examinees were used. Other research on this examination (Bontempo and Julian, 1997) indicated that examinees behave differently on the first 60 items than they do on the other items. Therefore, only data from the first sixty items answered by each examinee was used. Due to the nature of computerized adaptive testing, many of the items were not administered to any of the people in the sample, while many others were administered to a very small number of examinees. In order to achieve stabile item lengths while also analyzing as many items as possible, those items which did not contain at least 50 responses were eliminated from the analyses. In the end, data on 1051 people and 530 items were analyzed. All data were collected between April 1, 1996 and September 30, 1996. Although the constraints placed on the sample are numerous, the diversity in item response times across both people and items was still large (See Figure 1). For the items, the minimum of the mean item response time was 34 seconds while the maximum was 311 seconds. For the people, the minimum speed was 25 seconds/item while the maximum was 170 seconds/item.

Developing a common rating scale

With such disparity in the amount of time that people spend on items, one might think that a separate rating scale would be necessary for each item in the sample. However, if a common rating scale can yield useful measures, then this simpler, more parsimonious model should be used for all items.

Linearizing the item response times

Quick inspection of Figure 1 reveals that the distribution of item response times is extremely skewed. This is typical of distributions using time and for that matter typical

of distributions based on ratio scales. One could codify the counts of time by simply breaking the raw response time distribution into equal intervals. Take, for example, the ten equal intervals carved in Figure 2. However, time is counted on a ratio scale. Therefore, it will not fit as well as time counted on a linear, equal interval, scale. In order to linearize a ratio scale, one must take the natural logarithm of the counts, in this case the natural logarithm of item response time. Once the counts of time have been linearized, then the counts can be codified by dividing the scale into equal intervals (See Figure 4). This is the same as codifying the original raw response times using logarithmic intervals as opposed to equal intervals (See Figure 3). The data in the NCLEX-RN® examination sample were codified using logarithmic intervals.

Determining the precision of the rating scale

Exactly how many intervals upon which the data should be divided, is still unknown. One can conceive of a rating scale that would be as precise as the second. If one were to codify the NCLEX-RN® examination data using a scale such as this, then there would be well over 500 different data points for each item. This precise of a rating scale might be too precise to discern meaning in examining item response time. After all what's the difference between an examinee who spent 61 seconds as opposed to 60 seconds on an item. On the other hand, a rating scale with very few points might not be precise enough. Therefore, an investigation using the Rasch model (Rasch, 1980, Wright & Masters, 1982) must be conducted, in order to determine the number of rating scale points that yields the most meaning for examining item response time.

Since Big Steps, can handle a one hundred point scale, this is where the investigation began. The natural log of item response time data were calibrated first using a one hundred point rating scale and then a fifty point rating scale. Both of these analyses showed that these rating scales were not useful in describing the 530 items (See Figure 5 & Table 1). Twenty and fifteen point rating scales were then employed (See Figures 6 & 7 & Tables 2 & 3). The data fit these rating scales better than the previous two, but still had problems. The most salient problem was that the data did not fit the ends of the rating scale; the step measures at both ends were out of order. Finally, a ten point rating scale was analyzed (See Figure 8). The data fit this model (See Table 1). When this rating scale was employed there were no extreme items. Therefore, the rating scale was useful across all items in the sample although it handled the shorter items better than the longer items. This is evident in figures 9 and 10 which show the frequency distribution of rating scale categories for the longest and shortest item in the sample using the ten point rating scale.

In conclusion, using the Rasch model, a universal ten point rating scale was successfully developed to objectively measure examinee speed and item length for the NCLEX-RN examination. In the future, this 'ruler' will be used to investigate examinee speed in different settings. Those settings will be: on items with difficulties that vary in distance from the person's ability, on correct vs. incorrect items, and in timed vs. virtually untimed settings.

References:

Bontempo, B. D. & Julian, E. J. (1997). *Assessing Speededness in Variable-Length Computer Adaptive Testing*. A paper presented at the National Council on Measurement in Education, Chicago.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: MESA Press.

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago, MESA Press.

Figure 1. Distribution of Item Response Time

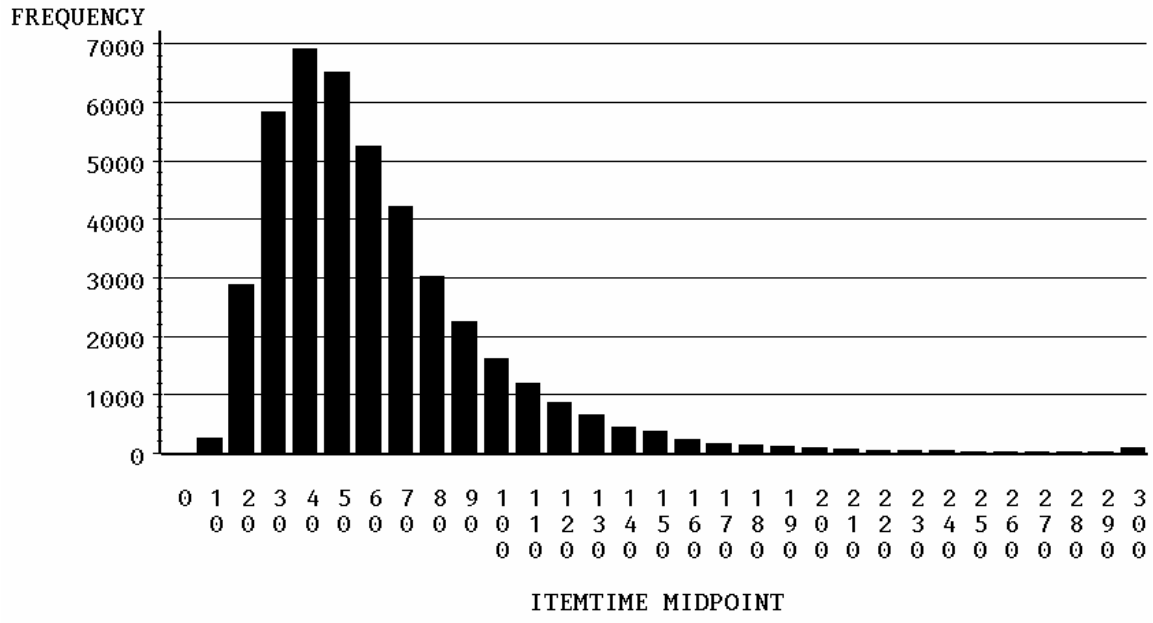

Figure 2. Distribution of the item response time split into 10 equal intervals

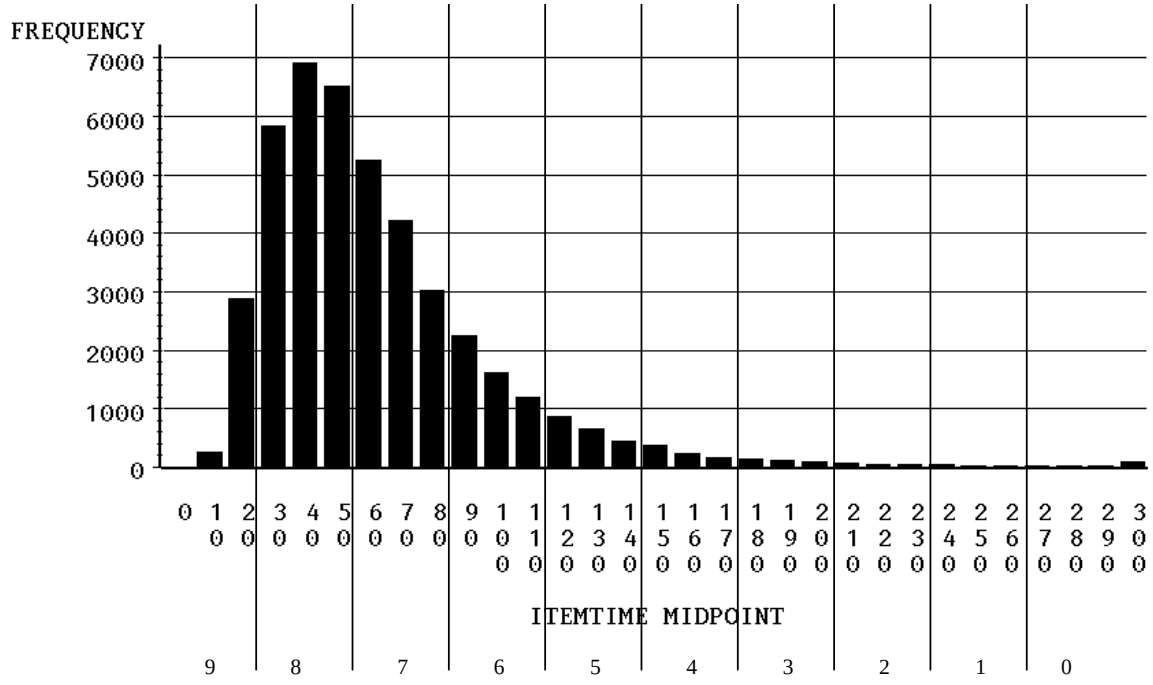

Figure 3. Distribution of the item response time split into 10 logarithmic intervals

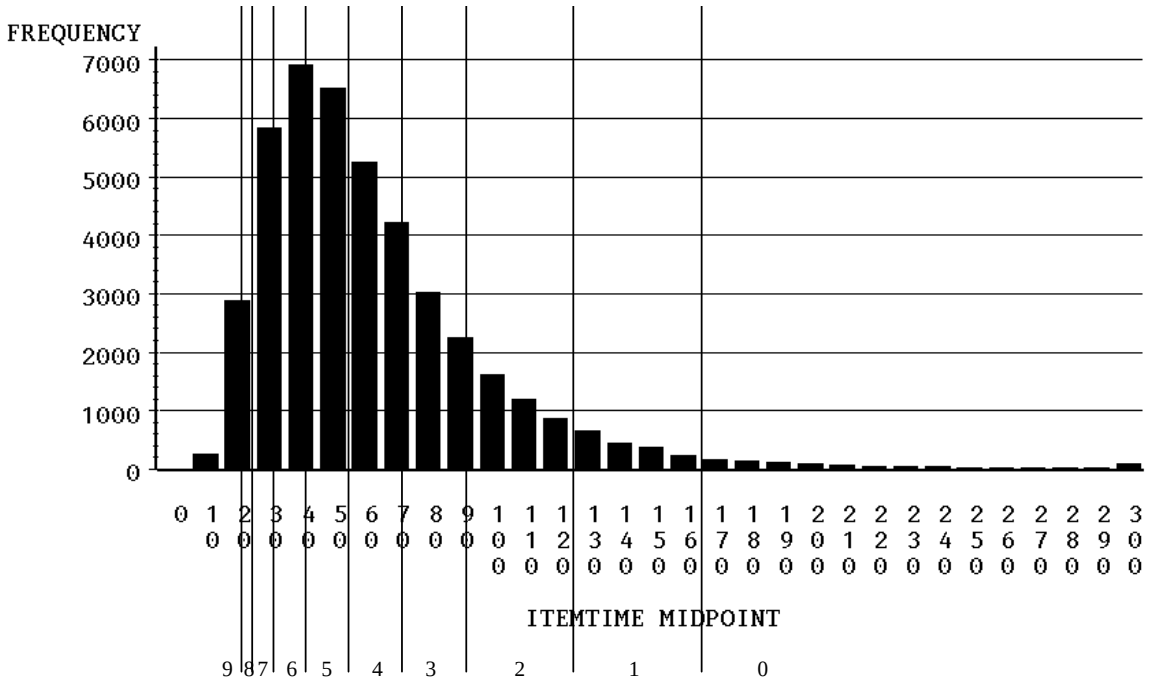

Figure 4. Distribution of ln of item response time split into 10 equal intervals

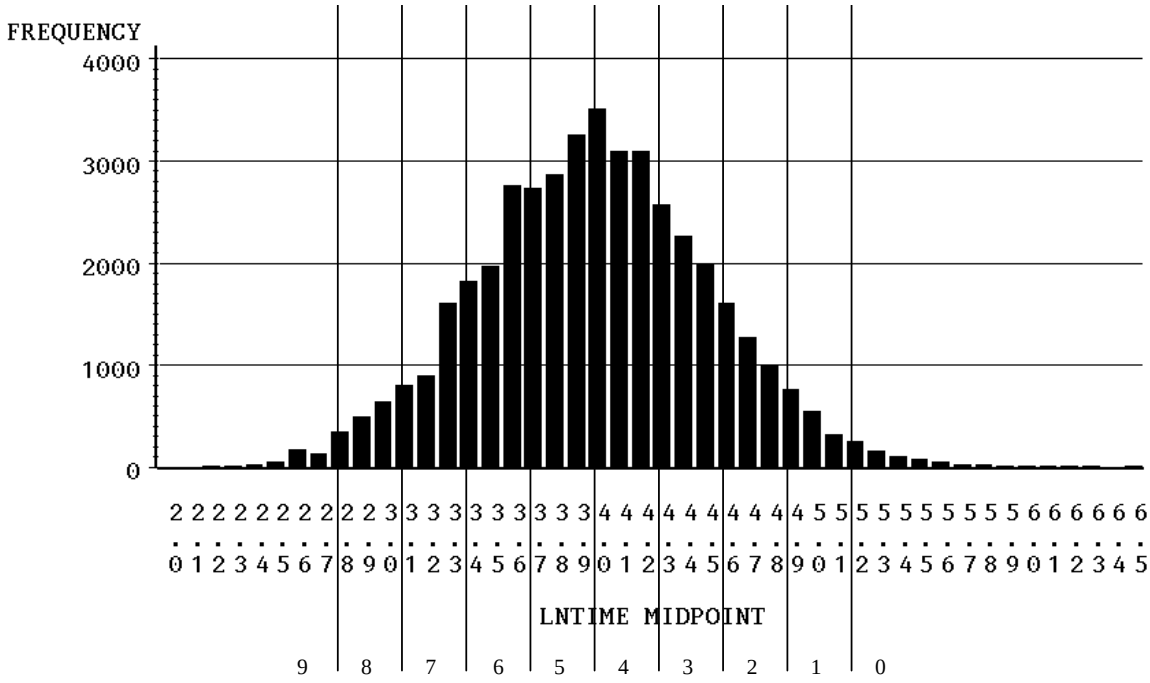

Figure 5. Frequency Distribution of the 50 point rating scale

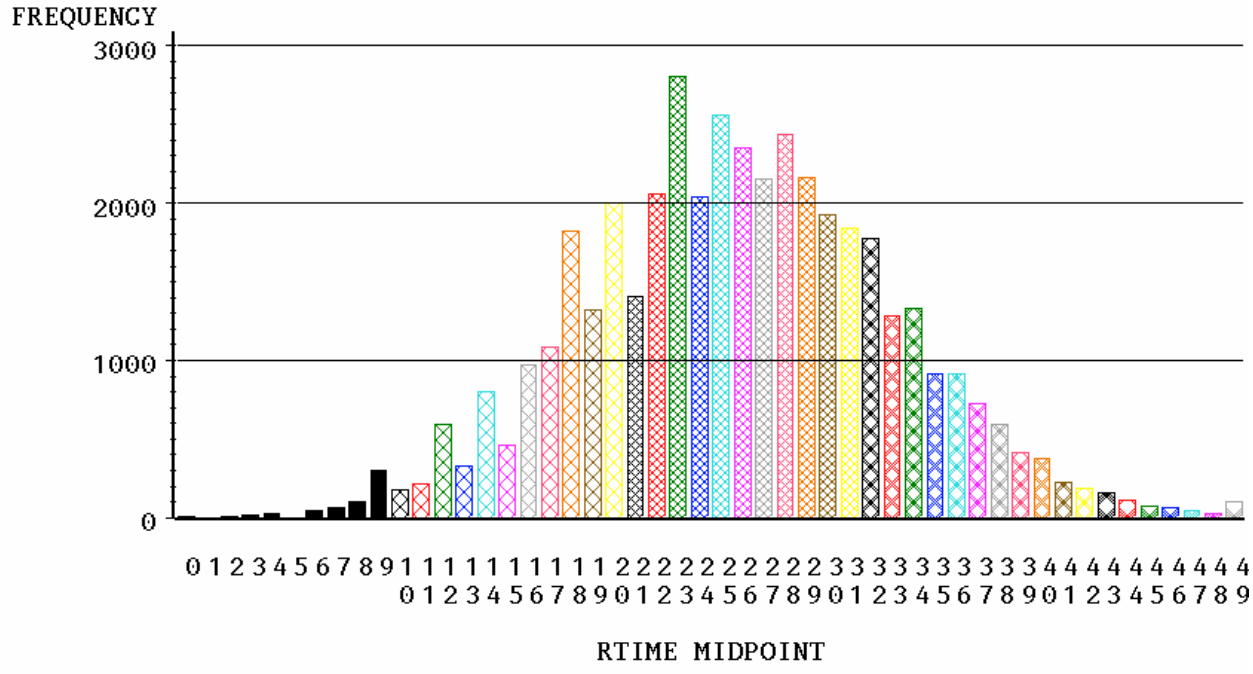

Table 1. Big Steps output of the 50 point rating scale

```
1051 PERSONS  750 ITEMS   ANALYZED:  1051 PERSONS  328 ITEMS    47 CATEGS v2.71
--------------------------------------------------------------------------------

     SUMMARY OF  1051 MEASURED (NON-EXTREME) PERSONS
+------------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|------------------------------------------------------------------------------|
| MEAN     531.8      21.5         .04     .04       .99    -.2    .99    -.2  |
| S.D.     149.9       4.9         .13     .01       .44    1.4    .44    1.4  |
|------------------------------------------------------------------------------|
| REAL RMSE    .04  ADJ.SD     .12  SEPARATION  2.68  PERSON RELIABILITY  .88  |
|MODEL RMSE    .04  ADJ.SD     .12  SEPARATION  2.91  PERSON RELIABILITY  .89  |
| S.E. OF PERSON MEAN    .00                                                   |
+------------------------------------------------------------------------------+
        VALID RESPONSES:   6.6%

     SUMMARY OF   328 MEASURED (NON-EXTREME) ITEMS
+------------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|------------------------------------------------------------------------------|
| MEAN    1704.1      69.0         .00     .02       .99    -.1    .99    -.1  |
| S.D.     723.5      28.1         .10     .01       .27    1.5    .27    1.5  |
|------------------------------------------------------------------------------|
| REAL RMSE    .03  ADJ.SD     .10  SEPARATION  3.73   ITEM  RELIABILITY  .93  |
|MODEL RMSE    .02  ADJ.SD     .10  SEPARATION  3.93   ITEM  RELIABILITY  .94  |
| S.E. OF  ITEM  MEAN    .01                                                   |
+------------------------------------------------------------------------------+
      LACKING RESPONSES:  422 ITEMS
```

SUMMARY OF MEASURED STEPS

CATEGORY LABEL	OBSERVED COUNT	AVGE MEASURE	INFIT MNSQ	OUTFIT MNSQ	STEP MEASURE
2	4	-.24	1.31	1.28	NONE
3	8	-.41	.52	.60	-.98
4	21	-.25	1.12	1.11	-1.24
6	30	-.25	1.01	1.00	-.62
7	27	-.22	1.11	1.10	-.14
8	56	-.25	.86	.87	-.96
9	147	-.24	.85	.85	-1.18
10	97	-.19	1.00	1.00	.21
11	123	-.18	.95	.95	-.42
12	330	-.17	.98	.98	-1.16
13	167	-.15	.93	.93	.53
14	434	-.13	1.01	1.01	-1.09
15	240	-.13	.88	.88	.47
16	493	-.11	.88	.88	-.83
17	552	-.09	.93	.93	-.21
18	980	-.07	1.01	1.01	-.65
19	702	-.06	1.00	.99	.27
20	1044	-.03	1.08	1.07	-.45
21	713	-.03	.95	.95	.35
22	1122	-.01	.98	.98	-.47
23	1450	.00	.95	.95	-.26
24	1087	.03	1.01	1.00	.30
25	1329	.04	1.01	1.01	-.17
26	1202	.05	.94	.95	.14
27	1152	.06	.98	.98	.10
28	1287	.08	.98	.98	-.04
29	1137	.09	.95	.95	.21
30	973	.11	1.05	1.05	.26
31	953	.12	1.02	1.02	.13
32	916	.14	.97	.97	.17
33	679	.16	.94	.94	.44
34	721	.16	1.07	1.07	.10
35	459	.18	1.04	1.04	.62
36	499	.19	1.08	1.08	.10
37	357	.21	1.00	1.00	.54
38	270	.22	1.01	1.01	.50
39	207	.23	1.06	1.06	.50
40	199	.25	1.10	1.10	.29
41	120	.28	1.00	1.01	.77
42	100	.30	.93	.93	.46
43	73	.28	1.15	1.15	.61
44	47	.26	1.33	1.34	.75
45	28	.33	1.05	1.05	.85
46	26	.33	1.13	1.13	.42
47	16	.32	1.36	1.37	.85
48	10	.43	.82	.83	.85
49	37	.33	1.38	1.38	-.91

Figure 6. Frequency Distribution of the 20 point rating scale

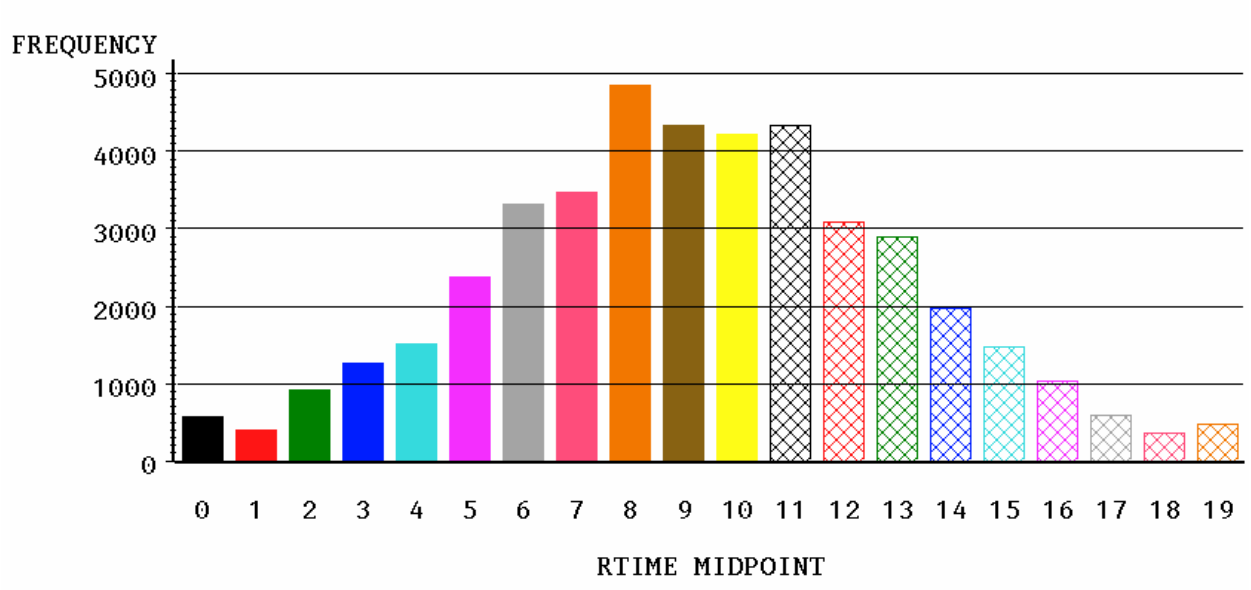

Table 2. Big Steps output of the 20 point rating scale

```
1051 PERSONS 1536 ITEMS   ANALYZED:  1051 PERSONS  530 ITEMS    20 CATEGS v2.71
-------------------------------------------------------------------------------

     SUMMARY OF  1051 MEASURED (NON-EXTREME) PERSONS
+-----------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     382.6      41.3         -.03     .06      1.00    -.2   1.00    -.2 |
| S.D.     121.0       9.6          .23     .01       .37    1.6    .37    1.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .06  ADJ.SD     .22  SEPARATION  3.71  PERSON RELIABILITY  .93 |
|MODEL RMSE    .06  ADJ.SD     .23  SEPARATION  3.99  PERSON RELIABILITY  .94 |
| S.E. OF PERSON MEAN    .01                                                  |
+-----------------------------------------------------------------------------+
       VALID RESPONSES:   7.8%

     SUMMARY OF   530 MEASURED (NON-EXTREME) ITEMS
+-----------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     758.7      81.9          .00     .04      1.01    -.1   1.01    -.1 |
| S.D.     288.3      26.4          .21     .01       .26    1.6    .25    1.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .04  ADJ.SD     .21  SEPARATION  4.87   ITEM  RELIABILITY  .96 |
|MODEL RMSE    .04  ADJ.SD     .21  SEPARATION  5.18   ITEM  RELIABILITY  .96 |
| S.E. OF  ITEM  MEAN    .01                                                  |
+-----------------------------------------------------------------------------+
     LACKING RESPONSES: 1006 ITEMS
```

SUMMARY OF MEASURED STEPS

CATEGORY LABEL	OBSERVED COUNT	AVGE MEASURE	INFIT MNSQ	OUTFIT MNSQ	STEP MEASURE
0	564	-.54	.93	.94	NONE
1	390	-.46	.97	.98	-.11
2	924	-.41	.98	.98	-1.29
3	1261	-.35	1.02	1.02	-.69
4	1505	-.30	.93	.93	-.50
5	2373	-.25	.96	.96	-.73
6	3321	-.19	1.03	1.03	-.56
7	3466	-.15	.94	.94	-.21
8	4847	-.09	.99	.99	-.46
9	4339	-.04	1.01	1.00	.04
10	4224	.01	.98	.99	.01
11	4335	.06	.98	.98	.00
12	3084	.11	.99	.98	.42
13	2897	.16	1.00	1.02	.19
14	1969	.21	1.04	1.04	.57
15	1470	.25	1.05	1.04	.53
16	1026	.30	1.06	1.05	.65
17	597	.37	1.08	1.08	.89
18	359	.43	1.12	1.11	.94
19	471	.67	1.05	1.04	.29

CATEGORY LABEL	STEP MEASURE	STEP ERROR	SCORE-TO-MEASURE AT CAT	--INTERVAL--		THURSTONE THRESHOLD
0	NONE		(-2.13)	-INF	-1.72	
1	-.11	.04	-1.35	-1.72	-1.15	-1.26
2	-1.29	.03	-1.01	-1.15	-.90	-1.04
3	-.69	.03	-.81	-.90	-.73	-.84
4	-.50	.02	-.66	-.73	-.60	-.70
5	-.73	.02	-.53	-.60	-.47	-.59
6	-.56	.01	-.41	-.47	-.35	-.48
7	-.21	.01	-.30	-.35	-.24	-.35
8	-.46	.01	-.18	-.24	-.12	-.25
9	.04	.01	-.06	-.12	.00	-.11
10	.01	.01	.06	.00	.12	.00
11	.00	.01	.18	.12	.23	.12
12	.42	.01	.29	.23	.35	.24
13	.19	.01	.41	.35	.47	.35
14	.57	.02	.53	.47	.59	.47
15	.53	.02	.66	.59	.73	.58
16	.65	.02	.81	.73	.90	.70
17	.89	.03	1.01	.90	1.15	.85
18	.94	.04	1.35	1.15	1.72	1.02
19	.29	.05	(2.15)	1.72	+INF	1.28

Figure 7. Frequency Distribution of the 15 point rating scale

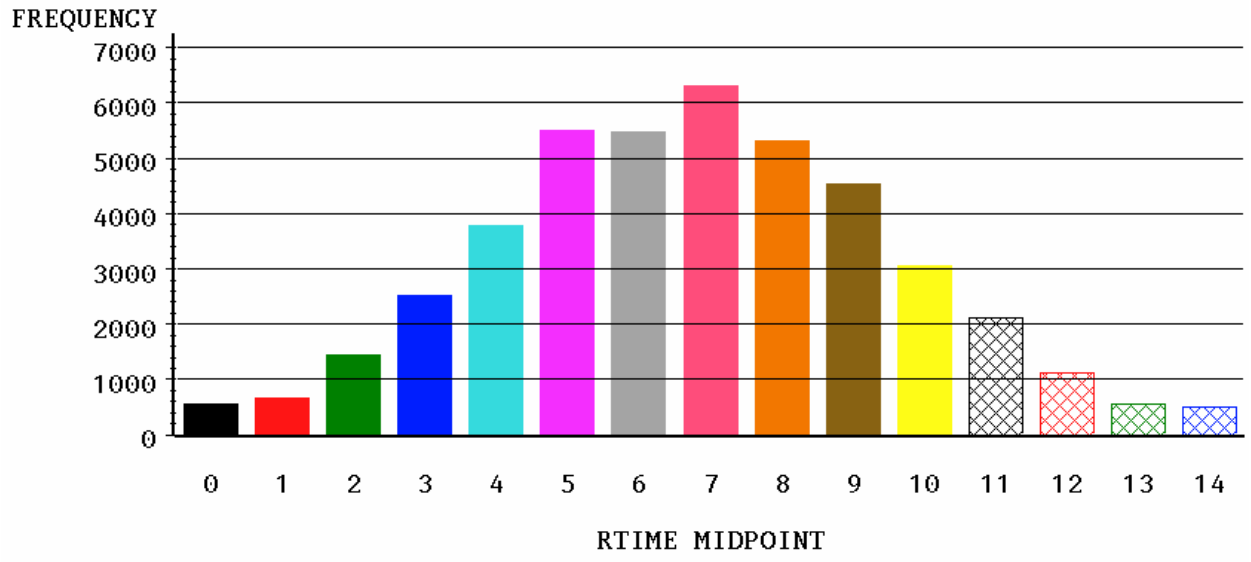

Table 3. Big Steps output of the 15 point rating scale

```
TABLE 3.1 Speededness Pilot                    PILOTB15.OUT Nov  6 17:35 1997
 1051 PERSONS 1536 ITEMS   ANALYZED:  1051 PERSONS  530 ITEMS    15 CATEGS v2.71
-------------------------------------------------------------------------------

     SUMMARY OF  1051 MEASURED (NON-EXTREME) PERSONS
+-----------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT   |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     280.9      41.3         -.04     .08      1.00    -.2   1.00    -.2 |
| S.D.      88.5       9.6          .32     .01       .37    1.6    .37    1.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .08  ADJ.SD     .31  SEPARATION  3.70  PERSON RELIABILITY  .93 |
|MODEL RMSE    .08  ADJ.SD     .31  SEPARATION  3.99  PERSON RELIABILITY  .94 |
| S.E. OF PERSON MEAN    .01                                                  |
+-----------------------------------------------------------------------------+
        VALID RESPONSES:   7.8%

     SUMMARY OF   530 MEASURED (NON-EXTREME) ITEMS
+-----------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT   |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     557.1      81.9          .00     .05      1.01    -.1   1.01    -.1 |
| S.D.     211.0      26.4          .29     .01       .25    1.6    .25    1.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .06  ADJ.SD     .28  SEPARATION  4.84   ITEM  RELIABILITY  .96 |
|MODEL RMSE    .06  ADJ.SD     .28  SEPARATION  5.14   ITEM  RELIABILITY  .96 |
| S.E. OF  ITEM  MEAN    .01                                                  |
+-----------------------------------------------------------------------------+
      LACKING RESPONSES: 1006 ITEMS
```

SUMMARY OF MEASURED STEPS

CATEGORY LABEL	OBSERVED COUNT	AVGE MEASURE	INFIT MNSQ	OUTFIT MNSQ	STEP MEASURE
0	564	-.73	.95	.95	NONE
1	670	-.61	.98	.98	-.82
2	1447	-.51	.99	.99	-1.32
3	2512	-.41	.93	.93	-1.00
4	3787	-.30	1.01	1.01	-.76
5	5496	-.21	.98	.98	-.63
6	5467	-.10	1.00	1.00	-.16
7	6299	-.02	.99	.99	-.21
8	5307	.08	.98	.98	.20
9	4530	.17	1.00	1.01	.28
10	3063	.26	1.02	1.03	.60
11	2122	.35	1.03	1.03	.68
12	1124	.43	1.10	1.09	1.05
13	544	.58	1.09	1.08	1.26
14	490	.90	1.07	1.06	.84

CATEGORY LABEL	STEP MEASURE	STEP ERROR	SCORE-TO-MEASURE AT CAT	--INTERVAL--		THURSTONE THRESHOLD
0	NONE		(-2.58)	-INF	-2.09	
1	-.82	.04	-1.64	-2.09	-1.39	-1.66
2	-1.32	.03	-1.20	-1.39	-1.04	-1.29
3	-1.00	.02	-.91	-1.04	-.78	-1.01
4	-.76	.02	-.66	-.78	-.55	-.78
5	-.63	.01	-.43	-.55	-.32	-.55
6	-.16	.01	-.21	-.32	-.10	-.31
7	-.21	.01	.01	-.10	.12	-.10
8	.20	.01	.22	.12	.33	.12
9	.28	.01	.44	.33	.55	.33
10	.60	.01	.66	.55	.78	.55
11	.68	.02	.90	.78	1.03	.76
12	1.05	.02	1.19	1.03	1.37	1.01
13	1.26	.03	1.63	1.37	2.08	1.27
14	.84	.05	(2.58)	2.08	+INF	1.65

Figure 8. Frequency Distribution of the 10 point rating scale

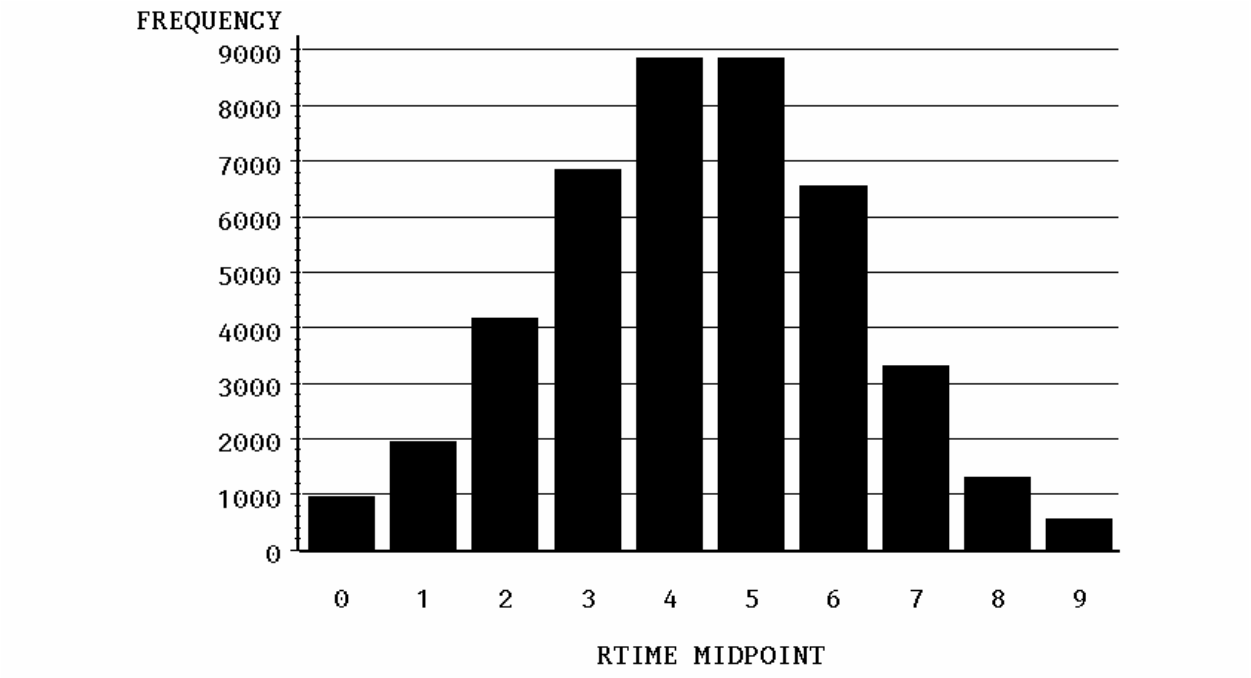

Table 4. Big Steps out of the 10 point rating scale

```
1051 PERSONS 1536 ITEMS   ANALYZED:  1051 PERSONS  530 ITEMS    10 CATEGS v2.71
-------------------------------------------------------------------------------

     SUMMARY OF  1051 MEASURED (NON-EXTREME) PERSONS
+-----------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT   |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     179.6      41.3         -.09     .11      1.00    -.2   1.00    -.2 |
| S.D.      57.4       9.6          .46     .02       .36    1.6    .36    1.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .12  ADJ.SD     .45  SEPARATION  3.63  PERSON RELIABILITY  .93 |
|MODEL RMSE    .12  ADJ.SD     .45  SEPARATION  3.91  PERSON RELIABILITY  .94 |
| S.E. OF PERSON MEAN    .01                                                  |
+-----------------------------------------------------------------------------+
        VALID RESPONSES:   7.8%

     SUMMARY OF   530 MEASURED (NON-EXTREME) ITEMS
+-----------------------------------------------------------------------------+
|           RAW                          MODEL          INFIT        OUTFIT   |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     356.1      81.9          .00     .08      1.00    -.1   1.00    -.1 |
| S.D.     130.3      26.4          .43     .02       .25    1.6    .24    1.5 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .09  ADJ.SD     .42  SEPARATION  4.78   ITEM  RELIABILITY  .96 |
|MODEL RMSE    .08  ADJ.SD     .42  SEPARATION  5.07   ITEM  RELIABILITY  .96 |
| S.E. OF  ITEM  MEAN    .02                                                  |
+-----------------------------------------------------------------------------+
      LACKING RESPONSES: 1006 ITEMS
```

SUMMARY OF MEASURED STEPS

CATEGORY LABEL	OBSERVED COUNT	AVGE MEASURE	INFIT MNSQ	OUTFIT MNSQ	STEP MEASURE
0	961	-1.25	1.04	1.03	NONE
1	1955	-.76	1.06	1.05	-1.69
2	4183	-.56	1.03	1.04	-1.44
3	6855	-.36	.99	1.00	-.95
4	8864	-.16	.99	.99	-.51
5	8861	.04	1.00	.99	-.05
6	6550	.26	.99	.99	.46
7	3315	.49	.94	.94	1.04
8	1314	.71	.96	.96	1.50
9	564	.93	.96	.96	1.63

CATEGORY LABEL	STEP MEASURE	STEP ERROR	SCORE-TO-MEASURE AT CAT	--INTERVAL--		THURSTONE THRESHOLD
0	NONE		(-3.18)	-INF	-2.56	
1	-1.69	.04	-1.93	-2.56	-1.55	-2.19
2	-1.44	.02	-1.24	-1.55	-.98	-1.48
3	-.95	.01	-.73	-.98	-.49	-.97
4	-.51	.01	-.26	-.49	-.01	-.50
5	-.05	.01	.23	-.01	.48	-.02
6	.46	.01	.74	.48	.99	.48
7	1.04	.02	1.26	.99	1.57	1.00
8	1.50	.03	1.95	1.57	2.55	1.51
9	1.63	.04	(3.16)	2.55	+INF	2.17

Figure 9. Distribution of rating scale counts for the longest item in the sample

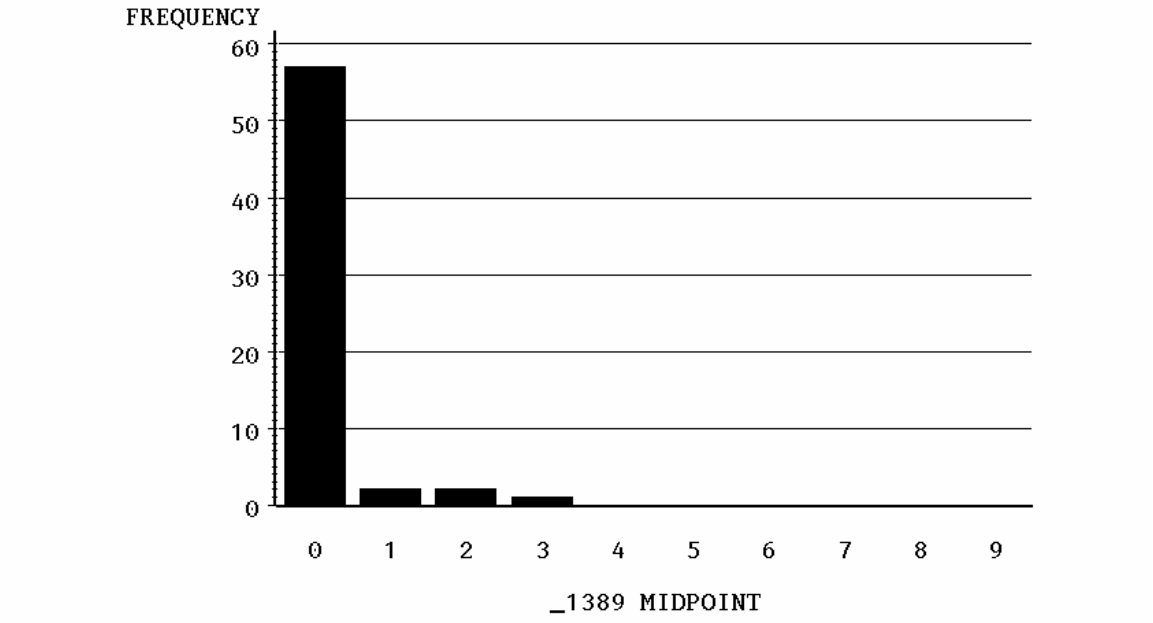

Figure 10. Distribution of rating scale counts for the shortest item in the sample