Analysis of Rating Scales for the Measurement of Attitudes and Perceptions

Lee Ong Kim
National Institute of Education,
Nanyang Technological University, Singapore
Email: oklee@nie.edu.sg

ABSTRACT

It is still common today to see questionnaires with Likert Scale items concerning very different variables being used to capture data on aspects as varied as possible that are to be investigated by the research work. This is perfectly alright if each of the questions is to be treated as standing on its own and is not intended to add up to a measure of a single variable. This, however, has the problem of inadequate sampling of items to come to any meaningful measure of persons on that set of multiple variables, with as small a standard error of measurement (SEM) as possible. Each variable to be measured is best put on a single rating scale, with items being replicated a sufficient number of times to reduce the SEM. There can be more than one rating scale in one questionnaire, but they should obviously be placed in separate sections, and their analyses done separately. This paper discusses a specific example of the measurement of attitude towards teaching and perceptions of subjects' own teaching knowledge and skills, and how to measure their changes over time, through the anchoring of item calibrations, using a Rasch model.

Introduction

Data used in this paper to show the determination of attitudes towards teaching and their perceptions of their own knowledge about teaching and their teaching skills, are from longitudinal research on Teacher

Preparation and Professional Development conducted by a team from the National Institute of Education (NIE), Nanyang Technological University (NTU), Singapore.[1] One objective of the research is to determine how well the program at NIE has prepared student teachers to face the world of teaching, by tracking changes in their attitude towards teaching, their perceptions of their own knowledge about teaching and their skills in teaching. An extension to this is the determination of reasons why they have decided to take up teaching as a career, and later, after a few years into their teaching in schools, why some of them may decide to leave teaching or why they stay. For the measurement of attitudes and perceptions, rating scales are used while the push-pull factors towards being a teacher will be obtained through intensive interviews. It is useful to note that these measurements will give a good indication of the efficacy of the training programs provided by NIE. The three programs investigated by this study are the Postgraduate Diploma in Education (PGDE) for both the Primary and Secondary school student teachers' programs, the Bachelor of Art and Bachelor of Science (B.A./B.Sc.) with Education program, and the Diploma in Education (Dip.Ed.) program. For this paper however, we use only the PGDE Primary and Secondary July 2004 Cohorts to demonstrate how attitudes towards teaching and perceptions of knowledge about teaching and skills in teaching can be measured and hence the pre-post measures compared. The PGDE is is used because it is a one-year program and both the entry and exit data are available while the Dip. Ed. and B.A./B.Sc. are two-year and three-year programs respectively for which the exit data are not yet available.

Methodology

This study uses the survey design and two intakes of student teacher, namely the 2004 and 2005 cohorts were involved as respondents. Each variable will be measured at several time points, starting from the time the student teachers are enrolled up until they are already two years into their teaching in schools. These have different time frames for the different programs. The PGDE programs are one-year programs and the Dip. Ed. program is over two years while the B.A./B.Sc. programs run over four years. The measures taken at the different time points will indicate how attitudes towards teaching, perceptions of knowledge about teaching and perceptions of teaching skills change over time, as student teachers go through their professional development programs and through

their actual teaching in schools. The study intends to examine what school factors can explain these changes.

i. Instrumentation

There are three questionnaires and they are labeled as Part A, Part B and Part C. Part A is for the purpose of collecting demographic data and includes respondents' reasons for joining teaching. Part B is a rating scale with 44 items for the measurement of attitudes towards teaching while Part C has two components, namely a rating scale for the measurement of respondents' perceptions of their knowledge about teaching and their perceptions of their teaching skills. The same set of 50 items were used in Part C. In one component respondents rate their perceptions of their knowledge about teaching on those stated aspects of teaching and in the other component, they rate their perceptions of their skills in carrying out those aspects of teaching.

The 44 items for part B covers statements that express feelings about teaching to which respondents indicate their level of agreement. Examples of statements are:

- Being a teacher, I would enjoy good recognition by the public
- Teaching is a respectable profession
- I plan to teach until I retire
- Teaching is a boring job
- I would encourage my friends to take up teaching

Some items in this instrument were reversed, such as "Teaching is a boring job" and the scores of such items are reversed during the analyses.

Part C contains 50 statements about teaching activities and strategies to which respondents are required to indicate two aspects, namely, (i) how much they think they have knowledge of that aspect and (ii) to what extent they think they have the skill to carry it out. Examples of statements are:

- Choosing appropriate teaching strategies for teaching particular topics.
- Choosing teaching strategies for students' ability level.
- Asking students the right questions to facilitate their learning.
- Production of appropriate teaching materials.
- Incorporates use of IT appropriately in the lessons

The face-to-face interview is a structured one for the purpose of collecting qualitative data on the same variables so that the data can be triangulated with the quantitative set. This portion, however, is not included in this paper.

ii. The Respondents

In calibrating instruments, the larger the number of respondents, the smaller the standard error of measurement (SEM). This study takes advantage of the briefing session given to new cohorts for each program, at the start of each of those programs. In this way we can capture all new students on that program to be respondents for the purpose of item calibration. The study will track these students at several time points, including the time they graduate, and at several other time points two years into their actual teaching in schools. For the 2004 cohorts, a total of 202 participants from the Postgraduate Diploma in Education (PGDE) Primary and 607 participants from PGDE Secondary, responded. In addition there were 170 respondents from the Diploma in Education (Dip. Ed.) program and 95 from the Bachelor of Arts and Bachelor of Science programs. The total number of participants who responded at entry point was 1074. These responses were "visually" cleaned and four respondents were removed from the list as it was found that two of them did not respond to the items at all and two more responded to less than 75% of the items. This leaves a total of 1070 respondents. It will be shown below that a further cleaning of data through removal of misfitting persons leaves 1065 respondents in the final analysis of entry point data.

Analyses for Data Cleaning

The discussion of the analysis here is for the scale on attitude towards teaching, but the same procedure is used for the two other variables, "perception of knowledge about teaching" and "perception of teaching skills".

A good calibration can be obtained by using as large a number of respondents as possible, and if the variance across the respondents is also large. Hence the respondents from all the programs (Postgraduate Diploma in Education Primary and Secondary, Diploma in Education, and the Degree Programs of B.A./B.Sc.), totaling 1070, were put into a single matrix as shown in Figure 1. Rasch analysis using Winsteps (Linacre and Wright, 2000) was done on this single matrix for the entry point data. In Figure 1, the codes to identify the programs are "DG" for degree programs, "DE" for diploma in education, "PP" for PGDE primary, and "PS" for PGDE secondary.

```
DG Respondent001 001S 443423444243442334514444144441444415555145 44
DG Respondent002 002S 333412224343333423332322234442434425444213 34
DG Respondent003 003S 453413334353441334414444154451545515455135 45
.. ..... ..... .....
.. ..... ..... .....
DE Respondent068 016S 333323334343334343343433244432444424444244 44
DE Respondent069 017S 455544555455451455515555155551555515555115 55
DE Respondent070 018S 233433223443342434524444244442444434444234 44
.. ..... ..... .....
.. ..... ..... .....
PP Respondent092 027S 432414344444342434424424244435445245451354 5
PP Respondent093 028S 443423334343333333333333343433433425445234 5
PP Respondent094 029S 343334334343443333324334343442434434444234 4
.. ..... ..... .....
.. ..... ..... .....
PS Respondent196 032S 342322334343333334433444244442444423444324 44
PS Respondent197 033S 453422335454442434424454254442444423445224 45
PS Respondent198 034S 444413435344442433424424243444444414444234 44
```

Figure 1: Single Data Matrix for Responses of All Respondents

What Rasch analysis does is to calibrate all items in this rating scale, on a single linear scale. Putting all the respondents from the different programs into the single matrix gets them measured on this variable, on the same linear scale and hence makes them comparable across individuals, as well as across groups. This is akin to getting the instrument "equated" in a single step, for all the different respondents since all of them were used simultaneously in the calibration of the instrument (see Lee, 2003). This calibration is done using participants at the entry point into their programs. Subsequent measures of these participants on this variable will be made in the same way, but with the item calibrations anchored on to the values obtained at this first time point. This now makes the subsequent measures comparable across the different points as they are measured on the same linear scale. An alternative way of getting respondents from different groups to be measured on the same scale would be to run a Rasch analysis on one group at a time, but with analyses of subsequent groups having item calibrations anchored on values obtained from the first. However, this method of continuous anchoring has the disadvantage of cumulative standard errors.

i. The Separation Reliabilities

The first run of the analysis of the 44 items on 1070 persons shows a large person separation reliability of 0.90 under real root mean square error (RMSE) as shown in Table 1. Separation is defined as the number

of RMSE units within the adjusted standard deviation of the measures. The adjusted standard deviation is obtained after correcting the observed variance of the measures for the error variance, estimated by taking the square of the RMSE. The separation reliability is then obtained using the definition:

$$\text{Separation reliability} = \frac{\text{Separation}}{1 + (\text{Separation})^2}$$

The separation reliability is an indicator of how well the instrument has separated the respondents on this variable if they are indeed different on this measure. The item separation reliability shows how well items are separated to represent the different aspects of feelings towards teaching, that constitute what may be defined as attitude towards teaching. It indicates how large the variance is of the item measures, in terms of how "difficult" it is for respondents to agree or disagree to. As shown in Table 1, the item separation reliability is 1.00.

ii. Removal of Misfitting Persons

At this point of the analysis, we will not take the person measures as final just yet and neither will we take the item calibrations as final. We need to look at the misfitting items and misfitting persons if there are any.

Table 2 is an extract from the larger table of poorly fitting persons from the output file. It shows the infit and outfit mean squares for the most misfitting persons along with the residuals for each item. Person DE 162 has an infit mean-square of 3.8 and outfit mean-square of 6.2, the contribution to which comes from her responses to items 10, 16 and 38, each with residuals of -3, -3 and -9 respectively.

In general, we remove the misfitting persons and run the analysis again. In this particular case, we will remove persons who have both infit and outfit mean squares of greater than 3.0. Values below 3.0 can be tolerated because this instrument has quite a varied number of aspects of teaching included in the instrument and it is quite normal for persons to have better feelings towards some aspects but not others and persons may differ considerably across these aspects resulting is more of the unexpected responses. Persons dropped are therefore DE162, PS745, PP298, PP430, PP355. The data are now analysed again without these 5 persons, that is a total of 1065 respondents with "clean" data.

Table 1: Summary Statistics for Time 1 Measures of Feelings About Teaching on Data Before Cleaning through Removal of Misfitting Persons

INPUT: 1070 PERSONS, 44 ITEMS MEASURED: 1070 PERSONS, 44 ITEMS, 5 CATS

SUMMARY OF 1070 MEASURED PERSONS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	169.5	43.9	1.85	.25	1.05	.0	1.02	-.2
S.D.	13.8	.5	.85	.02	.49	2.1	.50	2.1
MAX.	210.0	44.0	5.15	.40	3.81	7.7	6.25	7.8
MIN.	123.0	33.0	-.48	.22	.25	-5.0	.24	-5.2
REAL RMSE	.27	ADJ.SD	.80	SEPARATION	2.92	PERSON RELIABILITY		.90
MODEL RMSE	.25	ADJ.SD	.81	SEPARATION	3.27	PERSON RELIABILITY		.91
S.E. OF PERSON MEAN = .03								

SUMMARY OF 44 MEASURED ITEMS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	4121.1	1067.4	.00	.05	1.00	-1.1	1.02	-.8
S.D.	579.4	1.8	1.33	.01	.36	7.1	.37	7.0
MAX.	4949.0	1070.0	3.03	.06	1.85	9.9	1.89	9.9
MIN.	2651.0	1063.0	-2.27	.04	.52	-9.9	.53	-9.9
REAL RMSE	.05	ADJ.SD	1.32	SEPARATION	24.50	ITEM RELIABILITY		1.00
MODEL RMSE	.05	ADJ.SD	1.32	SEPARATION	26.17	ITEM RELIABILITY		1.00
S.E. OF ITEM MEAN = .20								

Table 2: Table of Poorly Fitting Persons for Time 1 Measures of Feeling About Teaching

```
INPUT: 1070 PERSONS, 44 ITEMS  MEASURED: 1070 PERSONS, 44 ITEMS,
-------------------------------------------------------------------
NUMBER - NAME -- POSITION ------ MEASURE - INFIT (MNSQ) OUTFIT

   162  DE Respondent 069S      5.00     3.8   A    6.2
  RESPONSE:      1:   5   5   5   5   5    5   5   5   5   2
Z-RESIDUAL:                                                -3

  RESPONSE:     11:   5   5   5   5   5    3   5   5   5   5
Z-RESIDUAL:                               -3

  RESPONSE:     21:   5   5   4   5   5    5   5   5   5   5
Z-RESIDUAL:

  RESPONSE:     31:   5   5   5   5   5    5   5   1   5   5
Z-RESIDUAL:                                    -9

  RESPONSE:     41:   4   5   5   5
Z-RESIDUAL:

   745  PS Respondent 284S      1.13     3.8   B    3.8
  RESPONSE:      1:   4   3   4   2   2    2   2   1   4   1
Z-RESIDUAL:                     -2  -3        -2

  RESPONSE:     11:   1   1   2   2   5    5   1   5   5   5
Z-RESIDUAL:          -4  -2      -2   2    2  -2   2   2

  RESPONSE:     21:   5   4   2   5   5    4   5   5   4   5
Z-RESIDUAL:           2           2         2

  RESPONSE:     31:   4   5   4   5   2    5   5   4   5   4
Z-RESIDUAL:                         -3

  RESPONSE:     41:   2   2   5   5
Z-RESIDUAL:             -3
```

In a similar way, the table of poorly fitting items shown in Table 3, is examined. Table 3 shows only a portion of the respnses and residuals for Item 23 and the rest of the reported items with mean squares below 1.9 are also not shown. The item reported to be most misfitting is Item 23. From the infit and outfit mean squares of 1.9 each, and from the distribution of residuals, it cannot be considered unusual for persons feelings to vary in the manner shown, for those respective items. The values of the mean squares are certainly tolerable. All items are therefore retained in the instrument.

The 1065 persons are now reanalyzed to give the final item calibrations and person measures. Both calibrations and measures are on the same

scale and in log odds units (logits). Table 4 show part of the table of person measures and Table 5 shows the item calibrations. It was found that in this second run, the item separation reliability deid not change from 1.00 and the person separation reliability reduced to 0.89 from 0.90. Although the effect may not be much, it is proper procedure to remove misfitting persons and items when calibrating items in a rating scale.

Table 3: Table of Poorly Fitting Items

```
TABLE OF POORLY FITTING ITEMS   (PERSONS IN ENTRY ORDER)
NUMBER - NAME -- POSITION ------ MEASURE - INFIT (MNSQ) OUTFIT
-------------------------------------------------------------------
    23  Q23                          .89     1.9   A    1.9
  RESPONSE:      1:   4   2   4   4   4    5   2   4   3   5
Z-RESIDUAL:                              2

  RESPONSE:     11:   4   5   4   3   5    2   4   5   4   4
Z-RESIDUAL:                          2   -2

  RESPONSE:     21:   3   2   4   4   4    2   3   3   4   4
Z-RESIDUAL:                              -2

  RESPONSE:     31:   4   2   5   5   4    5   3   5   4   1
Z-RESIDUAL:                  2                   2

  RESPONSE:     41:   4   2   5   2   4    4   4   5   5   3
Z-RESIDUAL:                      -2

  RESPONSE:     51:   4   4   5   4   4    5   4   4   4   3
Z-RESIDUAL:

  RESPONSE:     61:   4   5   2   2   3    4   4   4   5   4
Z-RESIDUAL:                  -2

  RESPONSE:     71:   5   3   5   5   3    4   4   4   4   2
Z-RESIDUAL:
```

Results of Analysis of Person Measures

The analysis results in all persons being measured and the output table of person measures (Table 4) reports these measures and their standard errors. The table reports the raw scores as well.

Why are results spuriously high or spuriously low when we use raw scores instead of measures? The main reason is the non-linearity of raw scores (Wright, 1992, 1993). As an illustration, consider the following pairs of persons in Table 4. Person with ID DE133S scored 210 raw score points and measures 5.20 logits. Compare this with Person PS311S

whose raw score is 204 and whose measure is 5.04 logits. Their difference in raw scores is six points and in measures is 0.16 logits. Another pair of persons with a raw score difference of six points would be PS271S with a raw score of 137 and a measure of 0.06 logits, with person PP176S with a raw score of 131 and a measure of -0.23. While the difference in raw scores is six (137 - 131), the difference in measures is 0.06-(-0.23) = 0.29 which is almost double 0.16.

Table 4: Person Statistics for Time 1 Measures of Attitude Towards Teaching (*Showing top and bottom 12 persons only*)

INPUT: 1065 PERSONS, 44 ITEMS MEASURED: 1065 PERSONS, 44 ITEMS, 5 CATS

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PERSON ID
158	210	44	5.20	.40	1.76	2.2	1.36	.8	DE 065S
225	210	44	5.20	.40	1.58	1.8	.79	-.3	DE 133S
767	204	43	5.04	.39	1.43	1.4	.79	-.3	PS 311S
111	204	43	5.04	.39	2.43	3.7	1.17	.5	DE 017S
210	207	44	4.77	.36	1.29	1.1	1.16	.5	DE 118S
99	205	44	4.52	.34	.88	-.4	.57	-1.4	DE 005S
827	205	44	4.52	.34	1.43	1.6	.97	.0	PS 371S
239	200	43	4.37	.34	1.53	1.9	1.05	.3	DE 147S
71	203	44	4.29	.33	2.38	4.1	1.65	1.9	DG 071S
222	203	44	4.29	.33	.70	-1.2	.83	-.5	DE 130S
212	202	44	4.18	.32	.96	-.1	.64	-1.3	DE 120S
202	201	44	4.08	.32	2.17	3.8	1.50	1.7	DE 110S
. . . . .									
. . . . .									
508	138	44	.11	.22	1.87	3.4	1.98	3.7	PS 049S
514	138	44	.11	.22	1.01	.1	1.02	.2	PS 055S
728	137	44	.06	.22	.94	-.2	.91	-.3	PS 271S
80	136	44	.02	.22	1.19	.9	1.17	.9	DG 080S
490	128	43	-.19	.22	1.23	1.1	1.14	.7	PS 031S
435	131	44	-.23	.22	1.63	2.6	1.67	2.8	PP 176S
392	130	44	-.27	.22	.95	-.2	.95	-.2	PP 132S
779	130	44	-.27	.22	1.07	.4	1.08	.5	PS 323S
40	129	44	-.32	.22	2.29	4.7	2.38	4.9	DG 040S
512	123	43	-.48	.22	2.05	4.0	1.99	3.8	PS 053S
MEAN	169.5	43.9	1.86	.25	1.05	.0	1.01	-.2	
S.D.	13.7	.5	.85	.02	.47	2.0	.45	2.0	

The item calibrations in Table 5 are those used to create an anchor file for the subsequent measures of attitude at future time points.

How do the cohorts for the different programmes compare in their feelings about teaching? In this single run of Rasch analysis, the attitude towards teaching for all the respondents from all the programs are simultaneously measured. To make this comparison, the complete file of person measures for all the 1065 persons was sorted out into the different programmes, namely Diploma in Education (DE), PGDE Primary (PP),

PGDE Secondary (PS), and the Degree Programme (DG). The respective mean measures for these groups are then calculated. Table 6 shows the mean measures, the standard deviations, the maximum and minimum measures for each of the four programme groups. It can be seen that the group with the highest mean measure of positive feelings about teaching is the Diploma in Education group, followed by the the other groups whose means are comparable.

Table 5: Item Calibrations for the Attitude Towards Teaching Rating Scale

ITEM	CALIBRATIONS	COUNT	SCORE	ERROR	IN.MSQ	IN.ZSTD	OUT.MS	OUT.ZSTD
1	1.51	1063	3481	0.04	1.46	9.53	1.48	9.87
2	0.49	1063	3959	0.05	1.07	1.62	1.08	1.82
3	2.23	1061	3094	0.04	1.69	9.90	1.73	9.90
4	0.15	1060	4091	0.05	1.49	9.48	1.51	9.82
5	-1.21	1058	4579	0.06	1.64	9.90	1.70	9.90
6	2.21	1059	3098	0.04	1.25	5.56	1.27	5.93
7	1.64	1063	3415	0.04	1.07	1.69	1.11	2.51
8	2.20	1061	3112	0.04	0.90	-2.52	0.91	-2.18
9	-0.86	1063	4482	0.05	0.81	-4.58	0.82	-4.54
10	3.03	1062	2646	0.04	1.04	1.00	1.07	1.59
11	-0.81	1061	4454	0.05	0.98	-0.42	0.98	-0.51
12	1.29	1063	3592	0.05	1.11	2.38	1.15	3.32
13	1.33	1061	3566	0.05	1.13	2.87	1.15	3.31
14	0.42	1058	3970	0.05	0.69	-7.68	0.70	-7.41
15	0.63	1062	3897	0.05	1.13	2.86	1.17	3.55
16	1.04	1061	3704	0.05	0.77	-5.61	0.82	-4.39
17	1.68	1064	3399	0.04	0.83	-4.30	0.86	-3.54
18	0.64	1061	3887	0.05	0.95	-1.17	0.96	-0.94
19	0.43	1062	3980	0.05	1.36	7.20	1.37	7.37
20	-0.84	1062	4471	0.05	0.85	-3.58	0.84	-3.91
21	0.01	1062	4155	0.05	0.66	-8.52	0.67	-8.34
22	-0.31	1062	4276	0.05	0.64	-9.02	0.65	-9.08
23	0.90	1065	3784	0.05	1.87	9.90	1.90	9.90
24	-0.16	1062	4219	0.05	0.67	-8.34	0.67	-8.34
25	-0.81	1063	4462	0.05	0.93	-1.64	0.92	-1.89
26	-1.25	1062	4610	0.06	0.70	-8.09	0.70	-8.00
27	-0.15	1065	4227	0.05	0.70	-7.47	0.70	-7.51
28	-1.03	1065	4550	0.05	0.59	-9.90	0.60	-9.90
29	-0.38	1065	4317	0.05	0.60	-9.90	0.60	-9.90
30	-1.03	1065	4547	0.05	1.36	7.69	1.33	7.05
31	-1.17	1064	4591	0.05	0.52	-9.90	0.53	-9.90
32	-0.44	1064	4334	0.05	0.79	-5.04	0.79	-5.06
33	-1.51	1063	4695	0.06	0.65	-9.90	0.64	-9.80
34	-1.63	1063	4734	0.06	0.67	-9.53	0.66	-9.19
35	-0.97	1064	4525	0.05	1.78	9.90	1.77	9.90
36	-1.18	1060	4577	0.06	1.54	9.90	1.55	9.90
37	-1.55	1064	4712	0.06	0.79	-5.49	0.81	-4.88
38	-1.03	1061	4531	0.05	0.66	-8.96	0.69	-8.26
39	-2.28	1065	4925	0.06	0.80	-5.43	0.75	-5.48
40	-1.27	1061	4612	0.06	1.14	3.27	1.17	3.91
41	3.06	1064	2637	0.04	1.47	9.90	1.50	9.90
42	-0.75	1061	4433	0.05	0.58	-9.90	0.59	-9.90
43	-0.50	1065	4362	0.05	0.63	-9.69	0.62	-9.90
44	-1.76	1064	4777	0.06	1.04	1.10	1.04	0.82

Table 6: Person Measure Statistics for Time 1 Cleaned Data on Feelings About Teaching

COURSE		MEASURE	COUNT	SCORE	ERROR	IN.MSQ	IN.ZSTD	OUT.MS	OUT.ZSTD
Diploma in Education (N=168)	Mean	2.41	43.82	177.58	0.26	1.13	0.31	1.07	0.09
	S.D.	0.96	0.92	14.23	0.03	0.53	2.16	0.50	2.15
	Max	5.20	44.00	210.00	0.40	2.90	6.12	2.61	5.35
	Min	0.36	33.00	139.00	0.22	0.26	-4.99	0.24	-5.18
Degree Course (N=94)	Mean	1.90	43.94	170.26	0.25	1.09	0.21	1.07	0.13
	S.D.	0.82	0.25	13.68	0.02	0.46	1.93	0.45	1.94
	Max	4.29	44.00	203.00	0.33	2.52	5.13	2.78	5.86
	Min	-0.32	43.00	129.00	0.22	0.42	-3.51	0.46	-3.37
PGDE Primary (N=198)	Mean	1.80	43.92	168.63	0.25	1.06	0.04	1.03	-0.08
	S.D.	0.83	0.33	13.82	0.02	0.47	2.04	0.46	2.07
	Max	3.88	44.00	199.00	0.31	2.92	6.23	2.69	5.75
	Min	-0.27	41.00	130.00	0.22	0.29	-4.58	0.29	-4.72
PGDE Secondary (N=605)	Mean	1.73	43.90	167.34	0.24	1.01	-0.15	0.98	-0.29
	S.D.	0.76	0.38	12.66	0.02	0.45	1.97	0.43	1.97
	Max	5.04	44.00	205.00	0.39	2.86	6.06	2.87	6.38
	Min	-0.48	40.00	123.00	0.22	0.27	-4.85	0.27	-5.01

To see if the differences in these mean measures between these four groups are significant, a one-way ANOVA is run on the data and the output is shown in Table 7.

Table 7 also shows that the variances of both measures and scores for all the groups are equal. The test also shows that there is significant difference between the measures of the four groups. Using the Tukey test, the significant difference is contributed by the Diploma in Education groups while the other three groups are not different from each other. Hence the beginning attitude towards teaching is highest amongst the Diploma in Education group of students. In the case of this study, this observation can probably be explained by the fact that the Diploma in Education student teachers are those who have done contract teaching in schools and hence have teaching experience before joining the program. They must have liked their experience and have high positive attitude towards teaching so that they decided to register for the Dip. Ed. program and make teaching their career.

Table 7: The ANOVA Output Table for Comparison of Means for Feeling About Teaching Between the Four Course Groups

Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
MEASURE	6.039	3	1061	.000
SCORE	2.724	3	1061	.043

ANOVA

Variable		Sum of Squares	df	Mean Square	F	Sig.
MEASURE	Between Groups	62.342	3	20.781	31.535	.000
	Within Groups	699.180	1061	.659		
	Total	761.523	1064			
SCORE	Between Groups	13992.503	3	4664.168	26.650	.000
	Within Groups	185691.538	1061	175.016		
	Total	199684.041	1064			

Multiple Comparisons (Tukey HSD)

Dependent Variable	(I) COURSE	(J) COURSE	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
MEASURE	Dip Ed	Degree	.5139(*)	.10456	.000	.2449	.7830
		PGDE(P)	.6072(*)	.08515	.000	.3881	.8263
		PGDE(S)	.6834(*)	.07079	.000	.5012	.8656
	Degree	Dip Ed	-.5139(*)	.10456	.000	-.7830	-.2449
		PGDE(P)	.0933	.10168	.796	-.1684	.3549
		PGDE(S)	.1695	.09000	.236	-.0621	.4010
	PGDE(P)	Dip Ed	-.6072(*)	.08515	.000	-.8263	-.3881
		Degree	-.0933	.10168	.796	-.3549	.1684
		PGDE(S)	.0762	.06646	.661	-.0948	.2472
	PGDE(S)	Dip Ed	-.6834(*)	.07079	.000	-.8656	-.5012
		Degree	-.1695	.09000	.236	-.4010	.0621
		PGDE(P)	-.0762	.06646	.661	-.2472	.0948
SCORE	Dip Ed	Degree	7.33(*)	1.704	.000	2.94	11.71
		PGDE(P)	8.96(*)	1.388	.000	5.39	12.53
		PGDE(S)	10.24(*)	1.154	.000	7.27	13.21
	Degree	Dip Ed	-7.33(*)	1.704	.000	-11.71	-2.94
		PGDE(P)	1.63	1.657	.759	-2.63	5.89
		PGDE(S)	2.91	1.467	.194	-.86	6.69
	PGDE(P)	Dip Ed	-8.96(*)	1.388	.000	-12.53	-5.39
		Degree	-1.63	1.657	.759	-5.89	2.63
		PGDE(S)	1.28	1.083	.637	-1.50	4.07
	PGDE(S)	Dip Ed	-10.24(*)	1.154	.000	-13.21	-7.27
		Degree	-2.91	1.467	.194	-6.69	.86
		PGDE(P)	-1.28	1.083	.637	-4.07	1.50

* The mean difference is significant at the .05 level

Attitudinal Change

Now that we have the anchor file, we are ready to measure change in attitude between the entry point measures and the exit point measures. For PGDE Primary, 198 student teachers responded to the attitudinal rating scale at their entry point into the program. At the exit point, 258 of them responded. Running the analyses anchored on item calibrations obtained earlier in the single step analysis, resulted in the person and item distributions shown in Figure 2. Figure 3 shows the corresponding person and item distributions for the PGDE Secondary student teachers, in which 605 of them responded at entry point into the program and 425 responded at the exit point from the program. Notice that the item distributions are the same since they are anchored on to the same scale.

It was observed that an appreciable number of student teachers who responded to the rating scale at entry point, did not respond to the rating scale at exit point after one year. Likewise, there were those who did not respond at the entry point but responded to the rating scale at their exit point. For the purpose of comparing the change in the mean measures between entry and exit points, it will have to be for the same group of persons. Hence respondents who appear in only one of the two time points were dropped from the determination of the mean measures. The same was done for the other two variables, namely "perception of knowledge about teaching" and "perception of teaching skills". The number of respondents left within each analysis is shown in the respective result tables below. Table 8 shows the mean attitude measure at Time 1 and Time 2 for both PGDE Primary and PGDE Secondary.

We notice that for both primary and secondary student teachers, the means of the post-measures are below their respective means of the pre-measures of attitude towards teaching. This means that after their respective programs, the attitude of the student teachers towards teaching has deteriorated. The mean attitudes, however, remained in the positive territory as can be seen by the fact that they remain above the mean item calibration of zero (item calibrations are centered at zero by Winsteps). Hence while the attitudes remain positive, they have diminished.

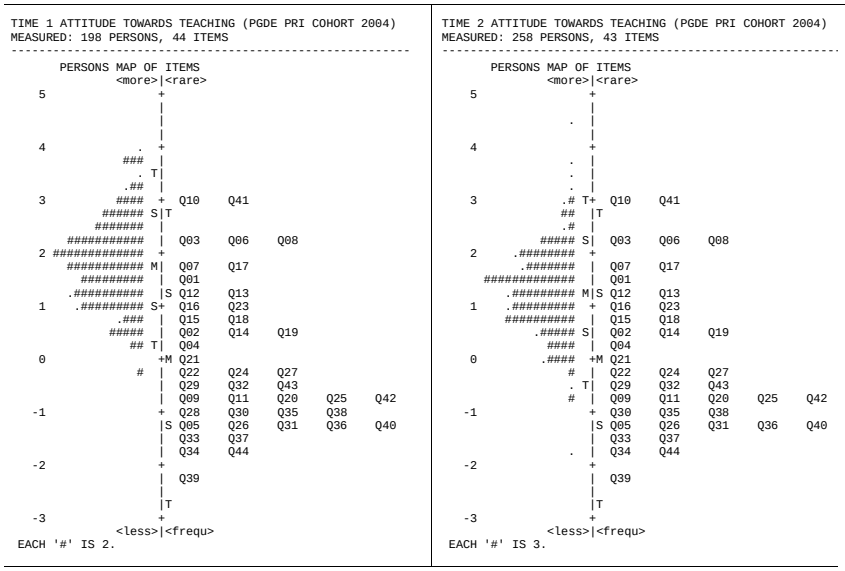

Figure 2: Distribution of Attitude Towards Teaching at Times 1 and 2 for PGDE Primary Cohort 2004

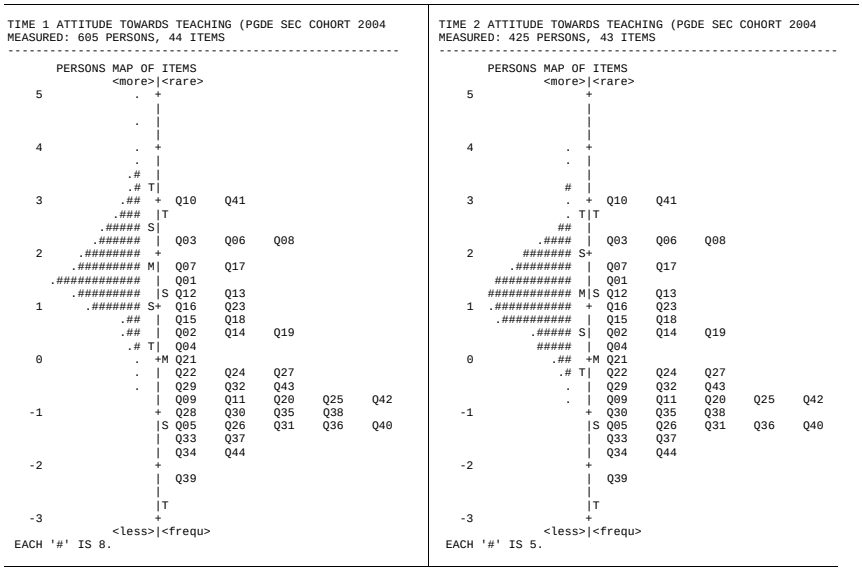

Figure 3: Distribution of Attitude Towards Teaching at Times 1 and 2 for PGDE Secondary Cohort 2004

Table 8: Mean Measures of Attitude for Time 1 and Time 2 for PGDE Cohort 2004

Programme	N	Time	Attitude Measures		Standard Errors	
			Mean	S.D.	Mean	S.D.
PGDE Primary	139	Time 1	1.84	0.88	0.25	0.02
		Time 2	1.37	0.88	0.24	0.02
PGDE Secondary	197	Time 1	1.77	0.76	0.25	0.02
		Time 2	1.32	0.71	0.24	0.01

To check if this drop is significant, the dependent sample t-test was used, and Table 9 shows that the change is significant at the level of significance of 0.05.

Table 9: Dependent t-test for T1-T2 Measures of Attitude Towards Teaching for PGDE Primary

Paired Samples Statistics

Pair	Mean	N	Std. Deviation	Std. Error Mean
TIME 1	1.8371	139	.87797	.07447
TIME 2	1.3691	139	.87662	.07435

Paired Samples Correlations

Pair	N	Correlation	Sig.
TIME 1 & TIME 2	139	.656	.000

Paired Samples Test

Pair	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
TIME 1 - TIME 2	.4680	.72723	.06168	.3460	.5900	7.587	138	.000

Similarly, the dependent sample t-test was used to check for the significance of the drop in mean attitude for the PGDE Secondary student teachers. Again this drop was found to be significant at the 0.05 level of significance as shown in Table 10.

Table 10: Dependent t-test for T1-T2 Mean Measures of Attitude Towards Teaching for PGDE Secondary

Paired Samples Statistics

Pair	Mean	N	Std. Deviation	Std. Error Mean
TIME 1	1.7662	197	.75561	.05384
TIME 2	1.323	197	.70500	.05020

Paired Samples Correlations

Pair	N	Correlation	Sig.
TIME 1 & TIME 2	197	.542	.000

Paired Samples Test

Pair	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
TIME 1 - TIME 2	.4429	.70054	.04991	.3445	.5413	8.874	196	.000

Change in Perception of Knowledge About Teaching

The analysis for attitude measure was repeated in exactly the same way for the other two variables, "perception of knowledge about teaching" and "perception of skills in teaching".

Table 11 summarises the change in mean perceptions of their knowledge about teaching which increases from 0.49 logits at entry point for PGDE Primary to 1.25 logits at exit point from the program a year later, and an increase from 0.57 logits to 1.04 logits in the case of PGDE Secondary.

Table 11: Mean Measures of Perceptions of Knowledge for Time 1 and Time 2 for PGDE Cohort 2004

Programme	N	Time	Knowledge Perception Measures		Standard Errors	
			Mean	S.D.	Mean	S.D.
PGDE Primary	139	Time 1	0.49	1.06	0.22	0.02
		Time 2	1.25	1.14	0.25	0.03
PGDE Secondary	326	Time 1	0.57	1.20	0.23	0.03
		Time 2	1.04	1.30	0.25	0.05

Tables 12 and 13 show that both these increases in the perception of knowledge about teaching for Primary and Secondary student teachers, were significant.

Table 12: Dependent t-test for T1-T2 Mean Measures of Perceptions of Knowledge of Teaching for PGDE Primary

Paired Samples Statistics

Pair	Mean	N	Std. Deviation	Std. Error Mean
TIME 1	.4866	139	1.05690	.08964
TIME 2	1.2527	139	1.14013	.09670

Paired Samples Correlations

Pair	N	Correlation	Sig.
TIME 1 & TIME 2	139	.140	.101

Paired Samples Test

Pair	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
TIME 1 - TIME 2	-.7660	1.44223	.12233	-1.0079	-.5242	-6.262	138	.000

Table 13: Dependent t-test for T1-T2 Mean Measures of Perceptions of Knowledge of Teaching for PGDE Secondary

Paired Samples Statistics

Pair	Mean	N	Std. Deviation	Std. Error Mean
TIME 1	.5749	329	1.24626	.06871
TIME 2	1.0469	329	1.34884	.07436

Paired Samples Correlations

Pair	N	Correlation	Sig.
TIME 1 & TIME 2	329	.028	.609

Paired Samples Test

Pair	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
TIME 1 - TIME 2	-.4720	1.81033	.09981	-.6683	-.2756	-4.729	328	.000

Change in Perception of Skills in Teaching

For the student teachers' change in perceptions of their teaching skills, the same analysis as for perceptions of knowledge about teaching was carried out and the outcome of the mean measures at entry and exit points are shown in Table 14.

Table 14: Mean Measures of Perceptions of Teaching Skills for Time 1 and Time 2 for PGDE Cohort 2004

Programme	N	Time	Skills Perception Measures		Standard Errors	
			Mean	S.D.	Mean	S.D.
PGDE Primary	137	Time 1	0.83	1.12	0.23	0.02
		Time 2	1.05	1.22	0.24	0.04
PGDE Secondary	314	Time 1	0.62	1.30	0.23	0.05
		Time 2	0.86	1.33	0.24	0.10

For PGDE Primary, the measure of perception of skills increased from 0.83 logits at entry point to 1.05 logits when they exit the program after one year. The corresponding values for PGDE Secondary are 0.62 logits and 0.86 logits.

The significance of these increases were tested using the dependent t-test at $\alpha = 0.05$ and both were found to be significant. The t-values are shown in Tables 8 and 9 for the Primary and Secondary groups respectively.

Table 8: Dependent t-test for T1-T2 Mean Measures of Perceptions of Skills for PGDE Primary

Paired Samples Statistics

Pair	Mean	N	Std. Deviation	Std. Error Mean
TIME 1	.8274	137	1.11615	.09536
TIME 2	1.0536	137	1.21735	.10400

Paired Samples Correlations

Pair	N	Correlation	Sig.
TIME 1 & TIME 2	137	.414	.000

Paired Samples Test

Pair	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
TIME 1 - TIME 2	-.2262	1.26612	.10817	-.4401	-.0123	-2.091	136	.038

Table 9: Dependent t-test for T1-T2 Mean Measures of Perceptions of Skills for PGDE Secondary

Paired Samples Statistics

Pair	Mean	N	Std. Deviation	Std. Error Mean
TIME 1	.6170	314	1.29984	.07335
TIME 2	.8622	314	1.32998	.07506

Paired Samples Correlations

Pair	N	Correlation	Sig.
TIME 1 & TIME 2	314	.110	.052

Paired Samples Test

Pair	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
TIME 1 - TIME 2	-.2452	1.75450	.09901	-.4400	-.0504	-2.476	313	.014

Conclusion

It is clearly important that comparisons of measures of variables across different groups over time, such as in the measurement of growth, must be on a linear scale. Raw scores will not do as their non-linearity will lead to either spuriously high or spuriously low values on the variable. Differences and changes that are not significant may be seen as significant and vice versa, if raw scores are used. It is clearly demonstrated above that two pairs of people may differ by the same amount of raw scores, but differ by different amounts in logit measures. There are two main reasons for the non-linearity of raw scores. One is that it is artificially subjected to the "floor" and "ceiling" effects. For example, if a seven-point (1 to 7) likert scale has 40 items, then the minimum any one person can score is 40 and the maximum is 280. What this means is that when a person is at either the low or high ends, no matter how much more a a second person is below him or above him, that person will not be scoring much lower or higher. Measures on the other hand, runs from minus infinity to plus infinity, thereby giving the appropriate linear distances between persons, if they indeed differ by those quantities on the variable. The second reason is that different persons responding to a rating scale,

will have different "zero" points. One person responding with a "4" on an agreement scale (from "1" for strongly disagree to "5" for strongly agree) does not necessarily agree by the same "amount" with another person also responding with a "4".

It is very common to see in research that two means are compared or the correlation between two variables are determined. What can happen in the comparison of means is that differences that are significant may be shown to be insignificant and vice-versa if raw scores are used. Even if the differences are shown to be significant using either the raw scores or measures, the magnitude of the difference will be erroneously high or low. Likewise, correlations between two variables may be determined to be spuriously high or low.

Note

[1] The author would like to express his gratitude to the research team of the Teacher Preparation and Professional Development longitudinal study headed by Professor Goh Kim Chuan for the use of data. This study is financed by a research grant EP 2/04 GKC, from the Ministry of Education, Singapore.

References

Lee, O. K. (2003). Rasch simultaneous vertical equating for measuring reading growth *Journal of Applied Measurement*. Vol. 4 No. 1.

Linacre, J. M. & Wright, B. D. (2000). *Winsteps: A Rasch model computer program*. Chicago: MESA Psychometric Laboratory.

Thurstone, L. L. (1959). Attitudes and be measured. In Thurstone, L. L. (ed.) *The Measurement of Values (pp. 215-233)*. Chicago: Chicago University Press.

Wright, B. D. (1992). Scores are not measures. *Rasch measurement: Transactions of the Rasch Measurement SIG, American Educational Research Association. 6*(1), p. 208.

Wright, B. D. (1993). *Thinking raw scores. Rasch measurement: Transactions of the Rasch Measurement SIG, American Educational Research Association. 7*(2), p. 299.