DOCUMENT RESUME

ED 339 742 TM 017 674

AUTHOR Anderson, David O.
TITLE Scale Stability of the NTE Core Battery Test of Professional Knowledge.
PUB DATE Oct 91
NOTE 28p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, NY, October 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Testing; Education Majors; Elementary Secondary Education; *Equated Scores; Estimation (Mathematics); Higher Education; Knowledge Level; *Licensing Examinations (Professions); Maximum Likelihood Statistics; Professional Education; Scaling; Scores; *Scoring; *Teacher Certification; *Test Reliability
IDENTIFIERS Drift (Scoring); *NTE Test of Professional Knowledge; Parallel Test Forms; Section Pre Equating (Tests); Teacher Competency Testing; Variance (Statistical)

ABSTRACT

Various parallel forms of the NTE (National Teacher Examinations) Test of Professional Knowledge (NTETPK) have been administered to examinees aspiring to initial teacher certification since 1982. To confirm that recent reported scores are comparable to those from earlier forms, the stability of the score scale was studied using section pre-equating methodology (SPE). THE NTETPK consists of four 35-item sections with parallel content. Three sections are operational, and one section is non-operational. Three sections of a 1983 test form (Form F-4) were separately administered (non-operationally) in the fall of 1988 to randomly selected thirds of the testing population along with another entire operational form (Form K2I). Each of the three sections was given to approximately 7,500 examinees. Using SPE, maximum likelihood mean and standard deviation and variance/covariance estimates were generated for the "newly constituted" F4 form, which was then equated to Form K2I. Comparison of the "new" and old conversion tables revealed differences ranging from -0.04 to -0.21 points. For all practical purposes, these differences are not significant. Within the range scored by 90% of the examinees, differences of one point occurred at only 8 out of 56 possible scores. Based on these results, it appears that there has been virtually no drift in the NTETPK during the period 1983 through 1988. A 12-item list of references is included. There are five tables of statistical data and three figures illustrating the study. (SLD)

Scale Stability of the NTE Core Battery Test of Professional Knowledge

David O. Anderson

Educational Testing Service, Princeton, New Jersey

Running head: SCALE STABILITY

Abstract

The NTE Test of Professional Knowledge is administered annually to over 85,000 examinees who aspire to receive initial teacher certification. Various parallel forms of the test have been administered since 1982 and scores are reported on a common scale. To confirm that reported scores on recent forms of the test are comparable to those from earlier forms, a study of the stability of the score scale was conducted utilizing the methodology of Section Pre-Equating (SPE).

The NTE Test of Professional Knowledge consists of four 35-item sections with parallel content. Three sections are operational; one is non-operational. The total raw score is the sum of the operational section raw scores. One or more new forms are developed every year, each form containing from 20 to 35 items in common with a previous form. For this study, three sections of a 1983 test form (constituting the entire operational form - Form F4) were separately re-administered non-operationally in the Fall of 1988 to random thirds of the testing population, along with another entire operational form (Form K2I). Using SPE, maximum likelihood mean and standard deviation and variance/covariance estimates were generated for the "newly-constituted" F4-form. The "new" F4-form was then equated to Form K2I.

Comparison of the "new" and old conversion tables revealed differences ranging from -0.04 to -0.21 points. For all practical purposes, these differences are not significant, especially across five intervening equatings and with SEMs of 4.1 and 3.7 for the F4-form and the K2I-form, respectively. The rounded integer scores, reported to examinees, differ by no more than 1 score point. These larger differences are simply an artifact of the integer rounding. In the reported score range between 630 and 661 (the minimum and maximum state qualifying scores), non-zero integer differences were found only at four scores. Within the range scored by 90% of examinees, differences of one point occurred at only eight scores. Based on these results, it appears that there

was virtually no drift in the NTE Test of Professional Knowledge scale during the five year period 1983 through 1988.

Scale Stability of the NTE Core Battery Test of Professional Knowledge[1]

Standard 4.9 of the APA/AERA/NCME Standards for Educational and Psychological Testing (1985) recommends that "continuing testing programs that attempt to maintain a common scale over time...conduct periodic checks of the stability of the scale" (page 34). Likewise, the ETS Standards for Quality and Fairness (Educational Testing Service, 1987) require ETS testing programs to periodically evaluate the stability of the score scale being used.

The 1986 and 1989 Audit Reports of NTE Programs indicate that a decision was made to study score scale stability when the Core Battery had been administered for five years. This study concerns itself with only the Test of Professional Knowledge and one particular study methodology, [Section Pre-Equating (SPE)]. A plan to study scale stability in the remaining Core Battery tests will be developed at a later date.

One method to study scale stability is to administer a given test form at widely separated times and equate the form to itself through multiple links. Comparison of the resulting two raw-to-scale conversion tables will give an indication of any scale drift or instability. For scales without any drift, the conversion tables would be virtually identical. Any differences could be attributed to non-equivalence of the testing populations (or samples), to measurement error, or to equating error (Brennan & Kolen, 1986).

The scale stability of several tests developed at ETS has been investigated over the years. Stewart (1966) studied the stability of the SAT-Verbal score scale over a fifteen year period using Tucker and Levine linear equating and equipercentile nonlinear equating, with common item

[1]Paper presented at the annual meeting of the Northeastern Educational Research Association (NERA), Ellenville, NY. October, 1991.

anchors. She found relatively small changes around the mean between 1953 and 1963 scores, and between 1957 and 1963 scores; changes of 20-35 points (on a scale of 200-800) were found between 1948 and 1963 scores. Using similar equating methods, Modu and Stern (1975) report finding average differences of 14 and 17 points for the SAT-Verbal and SAT-Mathematical scores, respectively, when comparing conversion tables for 1963 and 1966 with those of 1973. Modu and Stern (1977) confirmed differences of 8-10 points on the SAT-Verbal Scale between 1963 and 1973 score scales, using an alternative equating method.

Angell and Schneider (1989) investigated the scale stability of the Test of English for International Communication (TOEIC) between 1984 and 1988 with 4 and 5 intervening equatings. Using an internal anchor test nonequivalent groups data collection design, Tucker and Levine equating models, and the bisection of double-part equating, they found differences of over 16.3 points on the TOEIC Listening Comprehension scale which ranges from 5 to 450, and less than 3 points on the TOEIC Reading Comprehension scale (also 5 to 450).

To my knowledge, no stability studies have been published using the Section Pre-equating methodology adopted in the current study.

Description of the Test

The NTE Core Battery consists of the three separate tests: Test of Communication Skills, Test of General Knowledge, and Test of Professional Knowledge. These tests were first administered in November 1982. The Test of Professional Knowledge consists of four 35-item sections with parallel content, each separately timed for 30 minutes. Three sections are operational (count toward an examinee's score) and one is non-operational (not included in an examinee's score). Generally, the non-operational section is used for pretest purposes, albeit not for this study,

as described below. The total raw score is the sum of the operational section raw scores, the number of right answers on the items.

One or more new forms are developed every year; each form containing from 20 to 35 items in common with a previous form. Common item linear equating (using internal anchors, nonequivalent groups, and Tucker and Levine equating models) is typically used to place the raw total scores on the reported score scale. Scores are reported on a scale of 600 to 695, in one point intervals.

Insert Figure 1 about here

Figure 1 shows the linkage between form 3ENT and subsequent forms important to this study. Form 3ENT was first administered in November 1982 and contained 3 operational sections (E_{1-3}) with 104 scored items, and 1 non-operational section. Subsequent forms 3FNT1 and 3FNT2 had 35 items and 34 items, respectively, in common with 3ENT, and 33 items in common with each other. Forms 3FNT3 and 3FNT4 were first administered in late 1983; subsequently, forms 3GNT, 3HNT and 3INT were introduced at yearly intervals thereafter. Forms K-3INT (form 3INT with several revised items) and K2-3INT (form K-3INT slightly revised) were introduced in 1988.

Design of the Scale Drift Study

Scale stability can be investigated by administering a given test form at widely separated times and then equating the newly administered form to a recent form that is on scale, resulting in a "new" conversion table for that old form. Comparison of the resulting two raw-to-scale conversion tables (original and "new") will give an indication of any scale drift or instability. For scales without any drift, the conversion tables would be virtually equivalent.

To facilitate this scale stability study, Test Development staff assembled three versions of form K2-3INT, each of which contained one of the three different operational sections from form 3FNT4 as a non-operational section 4:

$K2\text{-}3INT_1$ contained section 1 of 3FNT4,
$K2\text{-}3INT_2$ contained section 2 of 3FNT4, and
$K2\text{-}3INT_3$ contained section 3 of 3FNT4.

That is, sections 1, 2, and 3 of 3FNT4 were operational when that form was administered as an intact form in 1983, but were considered non-operational when administered as section 4 in $K2\text{-}3INT_{1\text{-}3}$.

Form 3FNT4 (the F4-form) was first administered in October 1983 to 4,516 examinees. Two items were PINned[2] at that time. Each of the three K2-3INT forms (I-forms) was given in October 1988 to approximately 7,500 examinees. The forms were distributed in a spiralled arrangement at each test center. Section Pre-Equating methodology was then used to statistically combine these F4 sections and equate the combination ("new" 3FNT4 form) to the K2-3INT form. The "new" conversion table that results from this procedure was then compared to the original 3FNT4 conversion table.

The original conversion parameters for form 3FNT4 were obtained by averaging the results of the following three equatings:

1) total raw scores on 3FNT4 were equated, using Tucker equating, through scores on 33 common items to scores on 3FNT1;

2) total raw scores on 3FNT4 were equated, using Tucker equating, through scores on 68 common items to scores on 3FTN2;

[2] Items 10 and 16 in section 1. Form $K2\text{-}3INT_1$ also contained the two PINned items from 3FNT4. These items were designated DNS items on both forms.

3) total raw scores on 3FNT4 were equated by setting means and standard deviations equal to scores on form 3FNT3 (which was spiralled with 3FNT4) and averaging the result with the bisector of the two Tucker equatings.

Methodology

The technique of Section Pre-Equating (SPE) is ideally suited for the situation in which random groups of examinees take only portions of a complete test form, the whole of which must be equated to a form all examinees have taken (Holland & Wightman, 1982; Petersen, Kolen & Hoover, 1989; Wightman & Wightman, 1988). Typically, SPE is used when a section (or sections) of non-operational items is included within a complete (operational) test form; several versions of the form are developed, differing only in their non-operational section(s). These versions are then spiralled among examinees. Following the administration, it is possible to use the resulting combination of complete information on the operational items and incomplete information on the non-operational items to estimate the parameters of the desired equating function for the new test form. In this study, examinees took all three I-form sections and only one of three possible F4-form sections, as illustrated below.

	Sections					
Subform	I_1	I_2	I_3	$F4_1$	$F4_2$	$F4_3$
K2-3INT$_1$	X	X	X	X		
K2-3INT$_2$	X	X	X		X	
K2-3INT$_3$	X	X	X			X

That is, form K2-3INT$_1$ contained sections I_1, I_2, I_3, and $F4_1$. This is the one-variable section pre-equating design as described in detail by Holland and Wightman (1982) and by Wightman and Wightman (1988).

The goal is to equate the raw scores from the three non-operational F4-form sections (combined across sections) to the raw scores based on all the I-form items by the use of SPE. Linear equating can proceed once we have the observed means and standard deviations for the I-form and the maximum likelihood estimated mean and standard deviation and variance/covariance estimates for the F4-form sections on that form.

Results

Conversions

The means and standard deviations for the three 3FNT4 sections at both administrations are reported in Table 1. Means and standard deviations of the I-form sections are also reported. The 1988 group had higher means (about .9 points) and smaller standard deviations for all three F4 sections compared to the 1983 group of examinees. Note that the I-form sections appear to be slightly easier than the F4-form sections. The standard error of measurement (SEM) for the original F4-form was 4.1; for the I-form, 3.7.

. . . . . .

Insert Table 1 about here

. . . . . .

The covariances and correlations between sections and the section means are listed in Table 2. Because a different one/third of the examinees took each of the F4 sections, values shown in the table involving the F-form sections are maximum likelihood estimates calculated through the SPE process.

. . . . . .

Insert Tables 2 and Table 3 about here

. . . . . .

The parameters for converting the original F4-form raw scores to scale were $A_{original}$ = .9379 and $B_{original}$ = 594.4079. The SPE analysis produced new estimates of the conversion parameters, A_{new} = .9363 and B_{new} = 594.3650.

Table 3 details the differences between the original and new conversion lines. Column 1 shows the possible total raw scores from 0 to 103[3]. Column 2 contains the percentage of examinees scoring below a particular raw score (based on the original F4-form administered in 1983); this provides an indication of the distribution of scores. Columns 3 through 5 show the original converted scaled scores, the newly converted scaled scores and their differences, respectively. Differences in unrounded scaled scores ranged from -0.04 to -0.21. The last columns show the scaled scores rounded to integers (as they are reported to examinees[4]) and the 12 score points out of 91 for which the new score values differed by one point from the original scale value. Plots of the two conversion lines are shown in Figures 2 and 3.

Almost ninety percent of the examinees score between 620 and 675 (using the original conversions); within this range, the rounded differences are not zero at 8 points. Currently, 22 states stipulate a qualifying score on the NTE Test of Professional Knowledge as part of their certification requirements. The qualifying scores range from a minimum of 630 to a maximum of 661. Within this critical range, the original rounded converted scores were identical to the "new" rounded converted scores except for four score points, for which the original values exceeded the new values by just one point.

---

[3] The total raw score is the sum of raw scores from the three sections of 35 items each minus two PINned items.

[4] Converted scores below 600 are reported as 600.

Insert Table 4 about here

Scaled score means, standard deviations, and conversion parameters for the various forms are shown in Table 4. When the original and new conversion parameters for 3FNT4 are used with both examinee groups, the difference in scaled score means between conversions is 0.26 for 1988 examinees and -0.15 for 1983 examinees.

Standard Error of Equating

The standard error of equating (SEE) expresses the amount of error in test form equating that is due to using examinees from a particular administration date. For a given score on one form of the test, the error in estimating its equated score on another form is indexed by a standard error. In order to better evaluate the difference between the two conversions, SEEs for an illustrative equating were calculated. The SEE formula for SPE has not yet been developed; however, one SEE formula for linear equating that may be appropriate for our situation is from Lord [see Angoff (1984), pg. 106]:

$$SEE = \sqrt{2\sigma_y^2(1-r^2)[\frac{(1+r^2)Z_x^2+2}{N_t}]}$$

This particular formula assumes that (1) the distributions of examinee scores are normal, (2) both examinee groups are random samples from the same population, and (3) a set of common items are administered to each examinee group. Most equatings for the Test of Professional

Knowledge follow these assumptions, so that calculating this set of SEEs will provide preliminary information to help evaluate the conversions. For these calculations, the value of r used was .79[5].

Insert Table 5 about here

The standard errors of equating shown in Table 5 range from .72 to .15 for the mean scaled score of 658.61, when N= 10,000; they range from .51 to .11 for N=20,000.

Discussion

Comparison of the new converted scores with the original scores, in Table 3, shows scaled score differences of -0.04 to -0.21. For all practical purposes, these differences are not significant, especially across five intervening equatings and with SEMs of 4.1 and 3.7 for the F4-form and the I-form, respectively.

The rounded integer scores, reported to examinees, differ by no more than 1 score point. These larger differences are simply an artifact of the integer rounding: scores with fractions just below .5 rounded down, while those with fractions of .5 or just above, rounded up. In the reported score range between 630 and 661 (the minimum and maximum state qualifying scores), non-zero differences were found only at scores 637, 638, 652, and 653. Ninety percent of examinees score between 620 and 675 (using the original conversions); within this range, the reported score differences of one point occurred only at eight out of 56 possible scores.

If the two sets of conversion parameters are indeed equivalent, then they should obtain similar results when applied to a particular set of examinee data. As shown in Table 4, when using the original and "new" sets of conversion parameters with the 1983 examinee data, the difference

[5] The correlation between scores on the common item set and the non-common item set was .7904 for Form 3HNT and .7934 for Form 3INT.

in scaled score means is 0.26; for 1988 examinee data, the difference is -0.15. That is, if the "new" conversion parameters were used with the 1983 data, the unrounded scores would be .26 points higher, on average, than originally determined.

Conclusion

Based on these results, it appears that there was virtually no drift in the Test of Professional Knowledge scale during the five year period 1983 through 1988.

References

Angell, A. & Schneider, L. (1989). TOEIC Scale Stability Study. ETS Statistical Report SR-89-105.

Angoff, W.H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.

Braun, H.I., & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland & D.B. Rubin (Eds.), Test Equating. New York: Academic Press.

Brennan R.L., & Kolen, M.J. (1986). Practical issues in Linear Equating using the common item nonequivalent populations design (ACT Technical Bulletin No. 53). Iowa City, IA: The American College Testing Program.

Educational Testing Service. (1987). ETS standards for quality and fairness. Princeton, NJ: Author.

Holland, P.W., & Wightman, L.E. (1982). Section pre-equating. A preliminary investigation. In Holland, P.W. & D.B. Rubin (Eds.), Test Equating. New York: Academic Press.

Modu, C.C., & Stern, J. (1975). The stability of the SAT score scale (College Entrance Examination Board Research and Development Report RDR-74-75, No. 3. Princeton, NJ: Educational Testing Service.

Modu, C.C., & Stern, J. (1977). The stability of the SAT-Verbal score scale. Unpublished report. Princeton, NJ: Educational Testing Service.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In Linn, R.L. (Ed.), Educational Measurement (3rd ed.). New York: ACE, Macmillan.

Standards for educational and psychological testing. (1985). Washington, DC: American Psychological Association.

Stewart, E.E. (1966). The stability of the SAT-Verbal score scale (College Entrance Examination Board Research and Development Report RDR-66-7, No. 3. Princeton, NJ: Educational Testing Service.

Wightman, L.F., & Wightman, L.E. (1988). An empirical investigation of one variable section pre-equating. ETS Research Report No. 88-37.

Table 1. Means and Standard Deviations for the Operational Sections from Original and Latest Administrations NTE Test of Professional Knowledge

	Original F4-Form 1983 Admin.	I-Form 1988 Admin.	New F4-Form 1988 Admin.
1st Operational Section	21.75 4.65	23.93 4.74	22.62 4.62 N= 7,487
2nd Operational Section	22.69 5.28	23.98 5.66	23.43 5.18 N= 7,442
3rd Operational Section	21.59 5.29	23.23 5.15	22.59 5.22 N= 7,462
Total Operational Score	66.04 13.88 N= 4,516	71.15 14.25 N=22,391	68.62[1] 13.21[1] N=22,391

[1] Maximum likelihood estimates from SPE analysis.

E:\DXA\DRIFT\DRIFT2.NEW

Table 2. Maximum Likelihood Estimates of Covariance, Correlations, and Means[1] for NTE Test of Professional Knowledge, forms K2-3INT$_{1,2,3}$[2]

	I-form			F4-form		
	Section 1	Section 2	Section 3	Section 1	Section 2	Section 3
I-1	22.502	0.756	0.742	0.697	0.727	0.744
I-2	20.283	32.014	0.777	0.733	0.758	0.775
I-3	18.126	22.648	26.530	0.711	0.739	0.776
F4-1	15.179	19.049	16.821	21.071	0.636	0.656
F4-2	17.823	22.166	19.697	15.092	26.741	0.680
F4-3	18.611	23.129	21.074	15.866	18.540	27.803
Means	23.929	23.984	23.234	22.645	23.409	22.563

[1] Covariances are below the diagonal; correlations are above the diagonal; means are the bottom row.

[2] Statistics involving only operational sections on the I-form are actual values; those involving sections on the F-form are maximum likelihood estimates from the SPE analysis.

E:\DXA\DRIFT\DRIFT2.NEW

Table 3. Differences between Scaled Scores
Based on the Original and New Conversion Parameters
NTE Test of Professional Knowledge, form 3FNT4

		UNROUNDED			REPORTED		
Raw Score	% Below	Original Scale	New Scale	Residual	Original Scale	New Scale	Diff.
0	0.0	594.41	594.37	-0.04	600	600	0
1	0.0	595.35	595.30	-0.04	600	600	0
2	0.0	596.28	596.24	-0.05	600	600	0
3	0.0	597.22	597.17	-0.05	600	600	0
4	0.0	598.16	598.11	-0.05	600	600	0
5	0.0	599.10	599.05	-0.05	600	600	0
6	0.0	600.04	599.98	-0.05	600	600	0
7	0.0	600.97	600.92	-0.05	601	601	0
8	0.0	601.91	601.86	-0.06	602	602	0
9	0.0	602.85	602.79	-0.06	603	603	0
10	0.0	603.79	603.73	-0.06	604	604	0
11	0.0	604.72	604.66	-0.06	605	605	0
12	0.0	605.66	605.60	-0.06	606	606	0
13	0.0	606.60	606.54	-0.06	607	607	0
14	0.0	607.54	607.47	-0.07	**608**	**607**	-1
15	0.0	608.48	608.41	-0.07	608	608	0
16	0.0	609.41	609.35	-0.07	609	609	0
17	0.0	610.35	610.28	-0.07	610	610	0
18	0.0	611.29	611.22	-0.07	611	611	0
19	0.0	612.23	612.15	-0.07	612	612	0
20	0.1	613.17	613.09	-0.07	613	613	0
21	0.1	614.10	614.03	-0.08	614	614	0
22	0.1	615.04	614.96	-0.08	615	615	0
23	0.2	615.98	615.90	-0.08	616	616	0
24	0.2	616.92	616.84	-0.08	617	617	0
25	0.3	617.86	617.77	-0.08	618	618	0
26	0.3	618.79	618.71	-0.08	619	619	0
27	0.5	619.73	619.65	-0.09	620	620	0
28	0.7	620.67	620.58	-0.09	621	621	0
29	0.8	621.61	621.52	-0.09	622	622	0
30	0.8	622.54	622.45	-0.09	**623**	**622**	-1
31	1.0	623.48	623.39	-0.09	623	623	0
32	1.4	624.42	624.33	-0.09	624	624	0

E:\DXA\DRIFT\DRIFT2.NEW

		UNROUNDED			REPORTED		
Raw Score	% Below	Original Scale	New Scale	Residual	Original Scale	New Scale	Diff.
33	1.5	625.36	625.26	-0.10	625	625	0
34	1.8	626.30	626.20	-0.10	626	626	0
35	2.1	627.23	627.14	-0.10	627	627	0
36	2.3	628.17	628.07	-0.10	628	628	0
37	2.7	629.11	629.01	-0.10	629	629	0
38	3.2	630.05	629.94	-0.10	630	630	0
39	3.8	630.99	630.88	-0.11	631	631	0
40	4.3	631.92	631.82	-0.11	632	632	0
41	4.9	632.86	632.75	-0.11	633	633	0
42	5.5	633.80	633.69	-0.11	634	634	0
43	6.2	634.74	634.63	-0.11	635	635	0
44	7.0	635.68	635.56	-0.11	636	636	0
45	7.6	636.61	636.50	-0.11	**637**	**636**	-1
46	8.6	637.55	637.43	-0.12	**638**	**637**	-1
47	9.4	638.49	638.37	-0.12	638	638	0
48	10.4	639.43	639.31	-0.12	639	639	0
49	11.8	640.37	640.24	-0.12	640	640	0
50	13.1	641.30	641.18	-0.12	641	641	0
51	14.4	642.24	642.12	-0.12	642	642	0
52	15.8	643.18	643.05	-0.13	643	643	0
53	17.2	644.12	643.99	-0.13	644	644	0
54	18.8	645.05	644.93	-0.13	645	645	0
55	20.7	645.99	645.86	-0.13	646	646	0
56	22.4	646.93	646.80	-0.13	647	647	0
57	24.0	647.87	647.73	-0.13	648	648	0
58	26.1	648.81	648.67	-0.14	649	649	0
59	28.1	649.74	649.61	-0.14	650	650	0
60	29.9	650.68	650.54	-0.14	651	651	0
61	32.5	651.62	651.48	-0.14	**652**	**651**	-1
62	35.2	652.56	652.42	-0.14	**653**	**652**	-1
63	37.5	653.50	653.35	-0.14	653	653	0
64	40.0	654.43	654.29	-0.15	654	654	0
65	42.0	655.37	655.22	-0.15	655	655	0
66	44.8	656.31	656.16	-0.15	656	656	0
67	47.7	657.25	657.10	-0.15	657	657	0
68	50.2	658.19	658.03	-0.15	658	658	0

		UNROUNDED			REPORTED		
Raw Score	% Below	Original Scale	New Scale	Residual	Original Scale	New Scale	Diff.
69	52.6	659.12	658.97	-0.15	659	659	0
70	55.2	660.06	659.91	-0.15	660	660	0
71	58.1	661.00	660.84	-0.16	661	661	0
72	60.9	661.94	661.78	-0.16	662	662	0
73	63.5	662.87	662.71	-0.16	663	663	0
74	66.0	663.81	663.65	-0.16	664	664	0
75	69.3	664.75	664.59	-0.16	665	665	0
76	72.2	665.69	665.52	-0.16	666	666	0
77	74.8	666.63	666.46	-0.17	**667**	**666**	-1
78	77.6	667.56	667.40	-0.17	**668**	**667**	-1
79	80.3	668.50	668.33	-0.17	**669**	**668**	-1
80	82.7	669.44	669.27	-0.17	669	669	0
81	85.2	670.38	670.21	-0.17	670	670	0
82	87.3	671.32	671.14	-0.17	671	671	0
83	89.5	672.25	672.08	-0.18	672	672	0
84	91.3	673.19	673.01	-0.18	673	673	0
85	92.9	674.13	673.95	-0.18	674	674	0
86	94.2	675.07	674.89	-0.18	675	675	0
87	95.5	676.01	675.82	-0.18	676	676	0
88	96.1	676.94	676.76	-0.18	677	677	0
89	96.9	677.88	677.70	-0.19	678	678	0
90	97.7	678.82	678.63	-0.19	679	679	0
91	98.6	679.76	679.57	-0.19	680	680	0
92	99.2	680.69	680.50	-0.19	681	681	0
93	99.4	681.63	681.44	-0.19	**682**	**681**	-1
94	99.7	682.57	682.38	-0.19	**683**	**682**	-1
95	99.8	683.51	683.31	-0.19	**684**	**683**	-1
96	99.9	684.45	684.25	-0.20	684	684	0
97	100.0	685.38	685.19	-0.20	685	685	0
98	100.0	686.32	686.12	-0.20	686	686	0
99	100.0	687.26	687.06	-0.20	687	687	0
100	100.0	688.20	688.00	-0.20	688	688	0
101	100.0	689.14	688.93	-0.20	689	689	0
102	100.0	690.07	689.87	-0.21	690	690	0
103	100.0	691.01	690.80	-0.21	691	691	0

Table 4. Scaled Score Statistics
NTE Test of Professional Knowledge

	1988 Data			1983 Data	
	K2-3INT	3FNT4		3FNT4	
		Original Scale[1]	New Scale[2]	Original Scale[1]	New Scale[2]
Mean	658.61	658.35	658.61	656.88	656.73
Standard Deviation	12.37	12.39	12.37	12.78	12.76
Mean Difference		0.26		-0.15	

[1] Original conversion parameters (A=0.9363, B=594.3650) are used with specified data.

[2] New conversion parameters (A=0.9379, B=594.4079) are used with specified data.

E:\DXA\DRIFT\DRIFT2.NEW

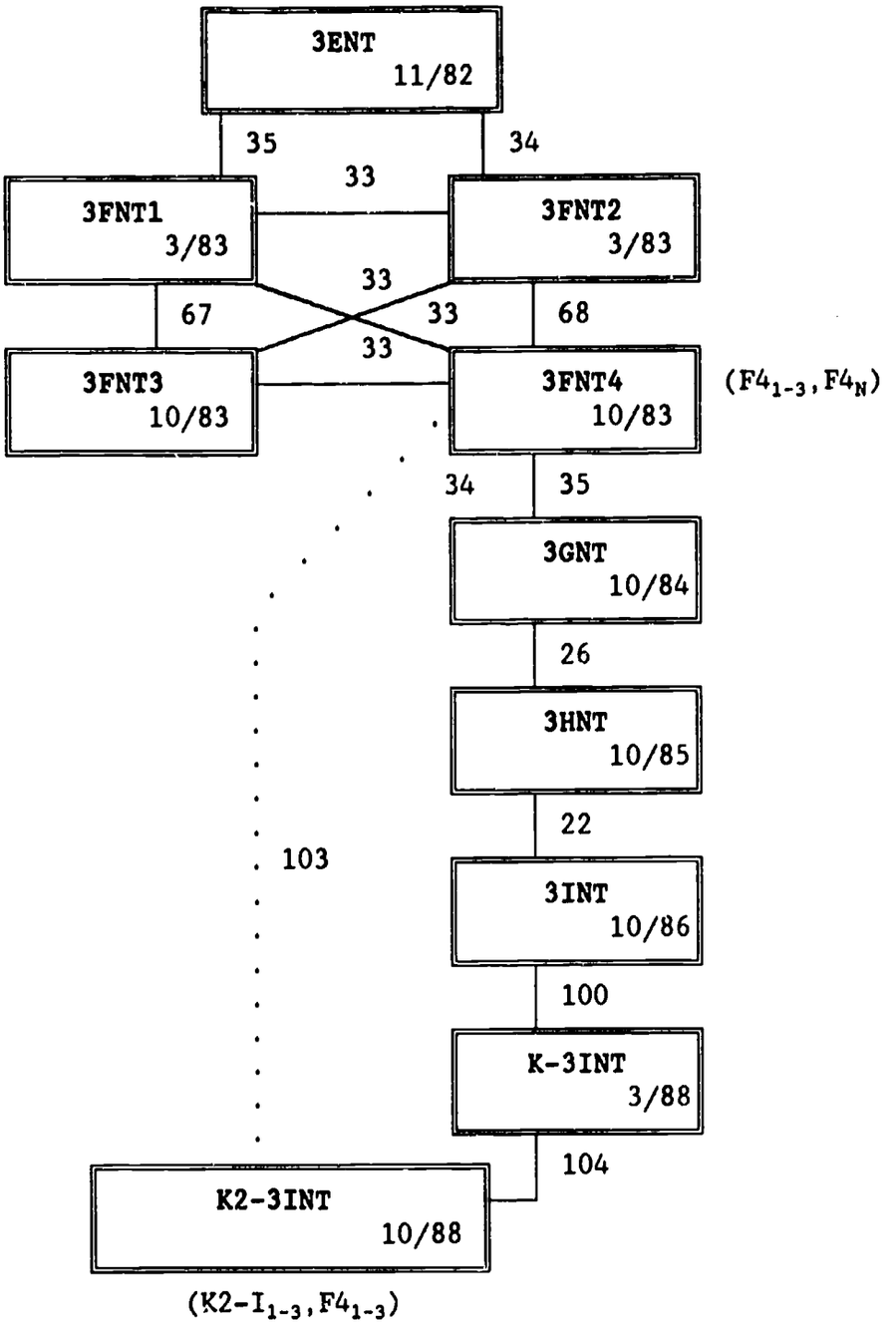
Table 5. Approximate[1] Standard Errors of Equating
Test of Professional Knowledge

Scaled Score	Standard Error of Equating N=10,000	Standard Error of Equating N=20,000
595	0.7191	0.5085
600	0.6652	0.4703
610	0.5581	0.3947
620	0.4528	0.3202
630	0.3506	0.2479
640	0.2555	0.1807
650	0.1791	0.1266
mean	0.1517	0.1073
660	0.1525	0.1078
670	0.1971	0.1394
680	0.2808	0.1986
690	0.3786	0.2677
695	0.4298	0.3039

[1] Formula is appropriate for the typical linear equating done for this test; it is not appropriate for Section Pre-Equating. The N-sizes used for the tabled values approximate those found in the data.

E:\DXA\DRIFT\DRIFT2.NEW

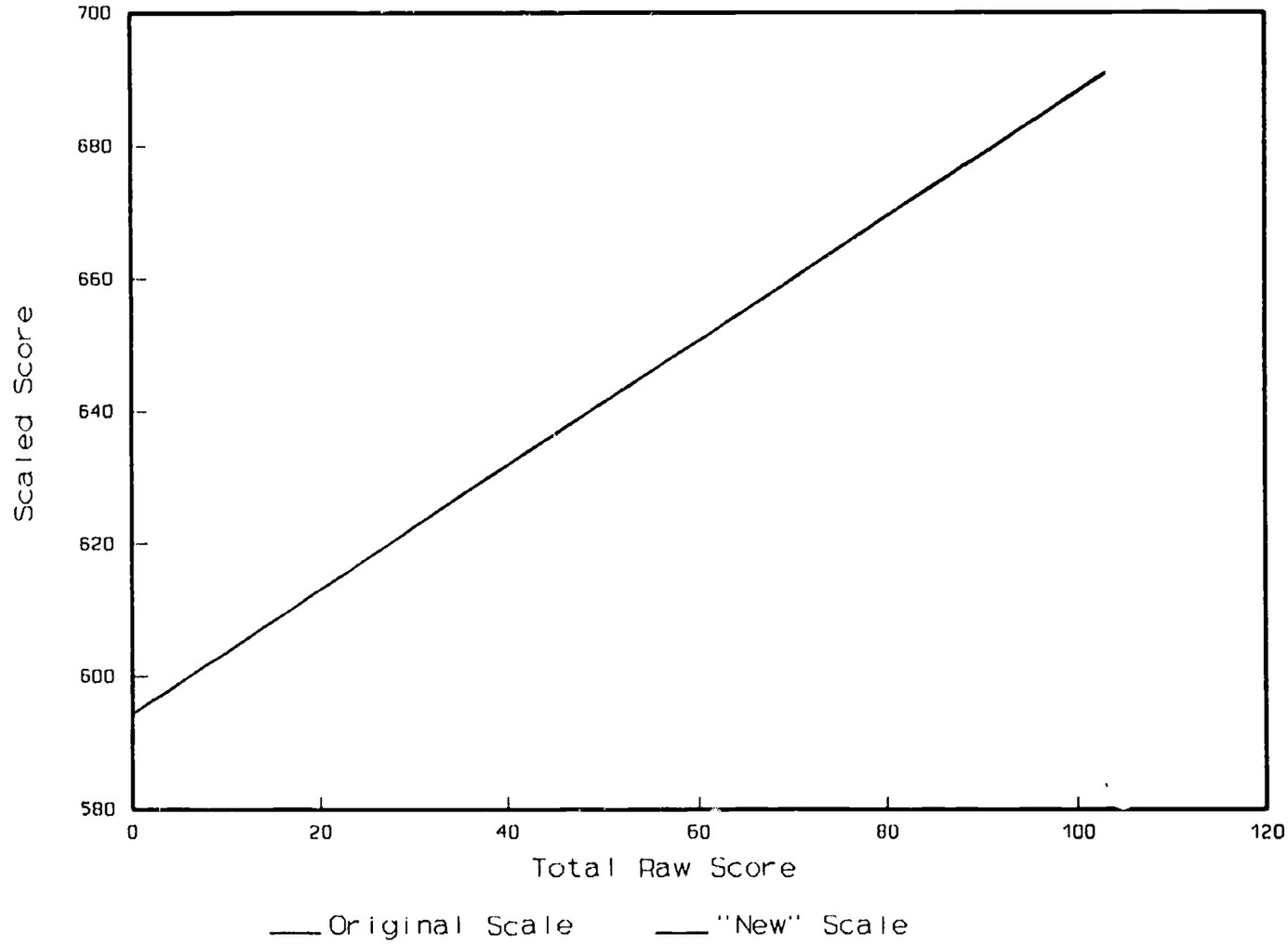
Figure 1. Equating Chain[1] for selected forms between 3ENT and K2-3INT



[1] The number of items common between forms is listed between the boxes.

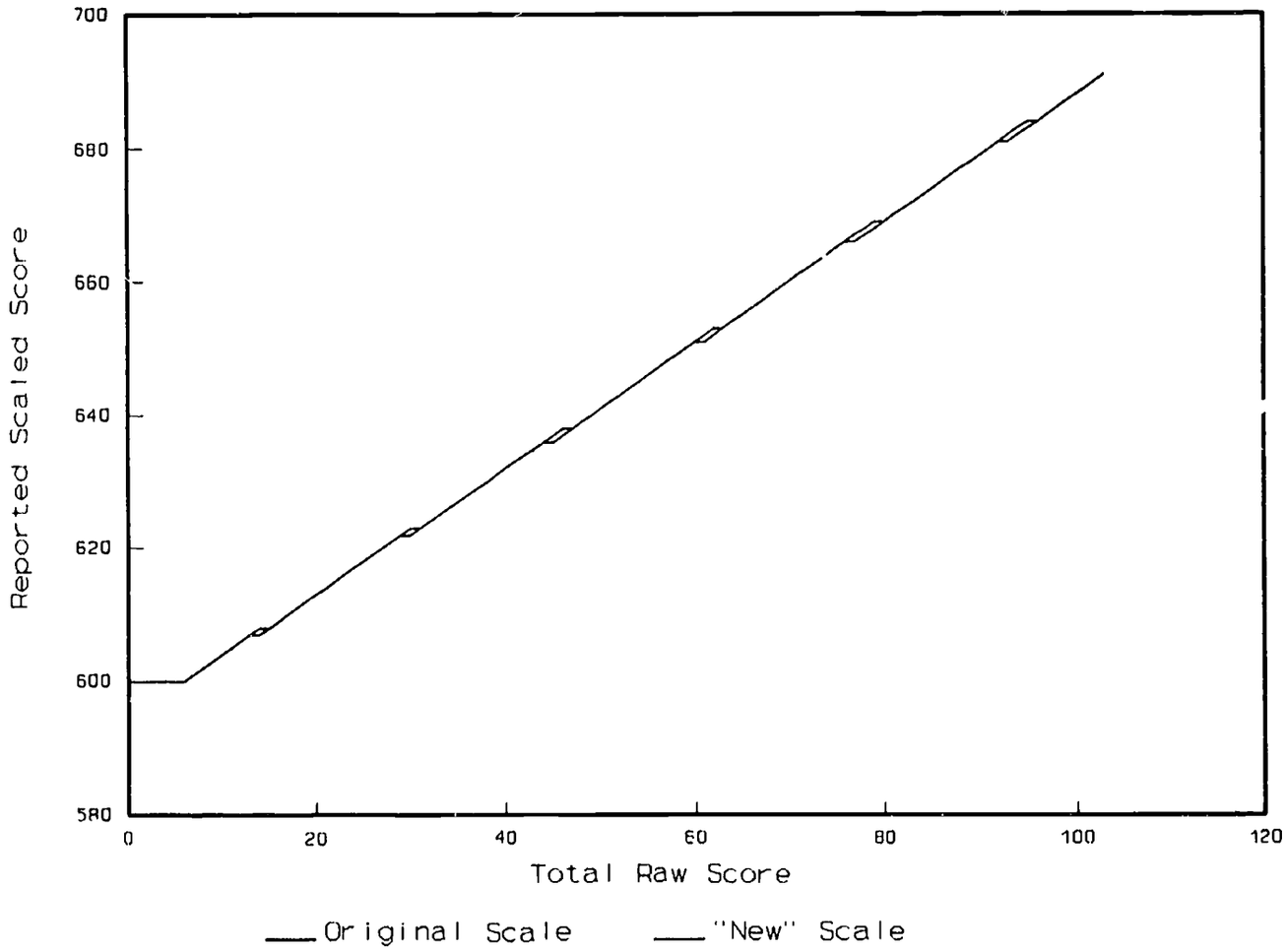
E:\DXA\DRIFT\DRIFT2.NEW

Figure 2. Scaled Score Conversion Lines for
Test of Professional Knowledge, form 3FNT4



E:\DXA\DRIFT\DRIFT2.NEW

Figure 3. Reported Scaled Score Conversion Lines for
Test of Professional Knowledge, form 3FNT4



E:\DXA\DRIFT\DRIFT2.NEW