Evaluación Inicial de los Procesos de Calibración y Equiparación de las Pruebas del Proyecto de *Estándares Nacionales*

Antonio Magriñá

CUADERNO No. 6

Instituto Nacional para la Evaluación de la Educación

Colección Cuadernos de Investigación

ISSN 1665-9457

Evaluación Inicial de los Procesos de Calibración y Equiparación de las Pruebas del Proyecto de *Estándares Nacionales*

Antonio Magriñá*

CUADERNO No. 6

*Director de *Test Development Psichometrics & Research-The College Board*, San Juan, Puerto Rico

Este es un resumen ejecutivo de su estudio, el texto completo puede consultarse en: www.inee.edu.mx

MÉXICO, OCTUBRE, 2003

Contenido

Introducción

Con el propósito de examinar los resultados de las equiparaciones de las Pruebas de *Estándares Nacionales*, se realizaron diversos análisis que evalúan la pertinencia de los parámetros de calificación desarrollados para las pruebas. Los resultados examinados en este documento corresponden exclusivamente a la administración de las pruebas del año 1999 y evalúa las equiparaciones verticales realizadas para los grados 3°, 5° y 6°. Los datos de administraciones más recientes están en proceso de análisis y sus resultados se informarán posteriormente.

Para estimar puntuaciones a escala, es recomendable realizar equiparaciones utilizando diversos modelos. Obtener resultados de modelos distintos de equiparación permite examinar la validez del modelo de equiparación seleccionado para propósitos operacionales y seleccionar el método (o métodos) y modelo (s) que produzcan menos error.

El proyecto de *Estándares Nacionales* utilizó el modelo de un parámetro de Rasch en sus procesos de calibración y equiparación. A continuación se delinean los procesos y resultados de las comparaciones de estas equiparaciones para el desarrollo de las puntuaciones a escala, aplicadas, con dos métodos de equiparación lineal (Tucher y Levine, 1955; Angoff, 1971; Kolen y Brennan, 1995). El propósito es comparar los resultados de las escalas operacionales desarrollados y aplicados para las pruebas de *Estándares Nacionales* con los resultados que se obtienen mediante la aplicación de otros modelos de equiparación, para determinar si se producen o no diferencias significativas al aplicar otros modelos y métodos de equiparación.

Análisis 1. Desarrollo de parámetros lineales para el grupo base (5° grado)

Inicialmente, se desarrollaron mediante el método de ejes principales *main axis*, parámetros lineales, utilizando como origen de la escala las medidas de tendencia central de la población examinada en quinto grado (definida como grupo base de equiparación). Se igualó el promedio y desviación de las puntuaciones crudas a la puntuación estimada por el modelo Rasch. Este procedimiento arrojó las siguientes equivalencias lineales con el modelo Rasch (tabla 1).

TABLA 1

PUNTUACIONES CRUDAS	RASCH	LINEAR 1	DIFERENCIAS RAS-LIN
0	0	25	-24.6
1	3	25	-22.3
2	10	26	-15.8
3	14	26	-12.1
4	18	27	-9.6
5	20	28	-7.7
6	22	28	-6.2
7	24	29	-5.1
8	25	30	-4.1
9	27	30	-3.3
10	28	31	-2.7
11	29	31	-2.1
12	30	32	-1.6
13	31	33	-1.2
14	32	33	-0.8
15	33	34	-0.5
16	34	34	-0.3
17	35	35	-0.1
18	36	36	0.1
19	37	36	0.3
20	37	37	0.4
21	38	38	0.5
22	39	38	0.6
23	39	39	0.7
24	40	39	0.7
25	41	40	0.8
26	41	41	0.8
27	42	41	0.8
28	43	42	0.8
29	43	42	0.8
30	44	43	0.7
31	44	44	0.7
32	45	44	0.7
33	45	45	0.6
34	46	45	0.6
35	47	46	0.5
36	47	47	0.4
37	48	47	0.4
38	48	48	0.3
39	49	49	0.2
40	49	49	0.2
41	50	50	0.1
42	50	50	0.0
43	51	51	-0.1
44	51	52	-0.2
45	52	52	-.02
46	53	53	-0.3

47	53	53	-0.4
48	54	54	-0.4
49	54	55	-0.5
50	55	55	-0.6
51	55	56	-0.6
52	56	57	-0.7
53	56	57	-0.7
54	57	58	-0.7
55	58	58	-0.8
56	58	59	-0.8
57	59	60	-0.8
58	59	60	-0.8
59	60	61	-0.8
60	61	61	-.07
61	61	62	-0.7
62	62	63	-0.6
63	63	63	-0.5
64	63	64	-0.4
65	64	64	-.03
66	65	65	-0.1
67	66	66	0.1
68	67	66	0.3
69	68	67	0.6
70	69	68	1.0
71	70	68	1.3
72	71	69	1.8
73	72	69	2.3
74	73	70	3.0
75	74	71	3.7
76	76	71	4.7
77	78	72	5.8
78	80	72	7.2
79	82	73	9.0
80	85	74	11.5
81	89	74	15.1
82	97	75	21.7
83	104	76	28.1

Parámetro lineal: a= .6136 b= 24.6157

La tabla 1 muestra los estimados a escala (20 a 80) de cada puntuación cruda (0 a 83).

Los límites operacionales de la escala son 20 y 80, respectivamente. Se observa que el estimado lineal se mantiene dentro de los límites de la escala (20-80), y el estimado de Rasch sobrepasa los límites (mínimos y máximos) de la escala. No obstante, las diferencias mayores entre ambos resultados son para las puntuaciones extremas, observándose diferencias menores a un punto (en la escala operacional, de 20 a 80 puntos) entre los estimados del modelo Rasch y los estimados lineales.

La ecuación lineal siguiente: Punt Escala=0.6136(Pt.cruda)+24.6157, representa adecuadamente ($p < .01$), los estimados del modelo Rasch para más del 68 por ciento de las puntuaciones (es decir, para dos desviaciones estándar de la escala). Se observan diferencias significativas entre el estimado Rasch, y el estimado lineal para las puntuaciones a escala menores de 32 puntos y mayores de 70 puntos, en la escala de 20 a 80 puntos, es decir, para la tercera desviación estándar de la escala.

Análisis 2. Equiparación de escalas de 5° y 6° grados

Los parámetros lineales de Análisis 1 , correspondientes al 5° grado, se tomaron como parámetros de origen para estimar las puntuaciones a escala de los grados 3°. y 6°. respectivamente, utilizando los métodos de equiparación lineal de Tucker y Levine. Entre los supuestos de estos métodos de equiparación, están los siguientes (Angoff, 1971):

Supuestos del método Tucker:

- las poblaciones equiparadas son similares en habilidad
- las versiones de prueba son paralelas
- las pruebas no difieren significativamente en confiabilidad

Supuestos del método Levine:

- las poblaciones difieren en habilidad
- las pruebas difieren en confiabilidad

La tabla 2 presenta las medidas de tendencia central, confiabilidades y correlaciones de las puntuaciones operacionales y de equiparación para 5° y 6° grados, de la prueba operacional y los reactivos comunes (aplicación 1999):

TABLA 2

	PRUEBA	**OPERACIONAL:**
	5° grado	6° grado
PROMEDIO=	34.98	32.569
D. S. =	12.076	10.301
N =	57187	34566
CONFIABILIDAD=	0.884	0.854
REACTIVOS DE EQUIPARACIÓN		
	5° grado	6° grado
PROMEDIO =	20.087	21.899
D. S. =	7.366	7.353
N =	57187	34566
CONFIABILIDAD=	0.827	0.827
CORR OP/EQ=	0.942	0.949

En la tabla 2 se observa que el promedio de la prueba operacional de sexto grado, es menor que el promedio de quinto, pero el grupo de 6° obtiene un promedio mayor que el de 5°, en los reactivos de equiparación. Se observa también que las desviaciones estándar y confiabilidades no difieren significativamente.

La tabla 3 provee la reestimación de las medidas de tendencia central y los nuevos parámetros de conversión lineal equiparados usando los métodos de Tucker y Levine, respectivamente:

TABLA 3

Reestimación de medidas de tendencia central de las puntuaciones crudas y parámetros, usando métodos de Tucker y Levine

TUCKER METHOD:

TUCKER 5to=	1.54433777	TUCKER 6to=	1.32947763

Mean, variance and standard deviation estimates:		
	5to.	6to
Mean=	36.0342154	31.0675314
Variance=	145.657758	106.321326
S.D. =	12.0688756	10.3112233

PARAMETERS: A PARAM	B PARAM:
6 TO 5 =	1.17046011 -0.3290909
5 TO SCALED =	0.6136 24.6157
6 TO SCALED=	0.71819432 24.4137698

LEVINE METHOD	
LEVINE 5TO= 1.69498074	LEVINE 6TO. 1.42360991

Mean, variance and standard deviation estimates:		
	5to.	6to.
MEAN =	36.1370492	30.9612214
VARIANCE=	145.622562	106.352223
S.D. =	12.0674174	10.3127214

PARAMETERS: A PARAM:	B PARAM:
6 TO 5=	1.17014868 -0.0921833
5 TO SCALED=	0.6136 24.6157
6 TO SCALED=	0.71800323 24.5591363

La tabla 4 provee los parámetros y valores lineales derivados de los mismos, para los métodos de equiparación Tucker y Levine:

TABLA 4
Parámetros lineales para sexto grado, usando métodos de equiparación Tucker y Levine

CONVERSIÓN DE PUNTUACIONES CRUDAS A ESCALA
PARÁMETROS :

	TUCKER	LEVINE
A=	0.71819	0.71800
B =	24.41377	24.55914

PUNTUACIONES A ESCALA

PUNTUACIÓN CRUDA	TUCKER	LEVINE
79	81.1511	81.2814
78	80.4329	80.5634
77	79.7147	79.8454
76	78.9965	79.1274
75	78.2783	78.4094
74	77.5601	77.6914
73	76.8420	76.9734
72	76.1238	76.2554
71	75.4056	75.5374
70	74.6874	74.8194
69	73.9692	74.1014
68	73.2510	73.3834
67	72.5328	72.6654
66	71.8146	71.9473
65	71.0964	71.2293
64	70.3782	70.5113
63	69.6600	69.7933
62	68.9418	69.0753
61	68.2236	68.3573
60	67.5054	67.6393
59	66.7872	66.9213
58	66.0690	66.2033
57	65.3508	65.4853
56	64.6327	64.7673
55	639145	64.0493
54	63.1963	63.3313
53	62.4781	62.6133
52	61.7599	61.8953
51	61.0417	61.1773
50	60.3235	60.4593
49	59.6053	59.7413
48	58.8871	59.0233
47	58.1689	58.3053
46	57.4507	57.5873
45	56.7325	56.8693
43	55.2961	55.4333
42	54.5779	54.7153

41	53.8597	53.9973
40	53.1415	53.2793
39	52.4233	52.5613
38	51.7052	51.8433
37	50.9870	51.1253
36	50.2688	50.4073
35	49.5506	49.6892
34	48.8324	48.9712
33	48.1142	48.2532
32	47.3960	47.5352
31	46.6778	46.8172
30	45.9596	46.0992
29	45.2414	45.3812
28	44.5232	44.6632
27	43.8050	43.9452
26	43.0868	43.2272
25	42.3686	42.5092
24	41.6504	41.7912
23	40.9322	41.0732
22	40.2140	40.3552
21	39.4959	39.6372
20	38.7777	38.9192
19	38.0595	38.2012
18	37.3413	37.4832
17	36.6231	36.7652
16	35.9049	36.0472
15	35.1867	35.3292
14	34.4685	34.6112
13	33.7503	33.8932
12	33.0321	33.1752
11	32.3139	32.4572
10	31.5957	31.7392
9	30.8775	31.0212
8	30.1593	30.3032
7	29.4411	29.5852
6	28.7229	28.8672
5	28.0047	28.1492
4	27.2865	27.4311
3	26.5684	26.7131
2	25.8502	25.9951
1	25.1320	25.2771
0	24.4138	24.5591

En la tabla 4 se observa que no hay diferencias significativas entre ambos métodos de equiparación. Análisis de homoscedasticidad, análisis de errores de mínimos cuadrados *mean square error* y otros análisis, sugieren que se utilicen los parámetros obtenidos con el método de Tucker (aunque los resultados sugieren que puede utilizarse indistintamente cualquiera de los parámetros adecuados de los dos métodos). Los análisis plantean que los supuestos del método Tucker son válidos para esta equiparación.

Al aplicar los parámetros derivados de la equiparación del método Tucker a los datos crudos de 6° grado, se obtuvo un promedio a escala (20-80) de 47.8 y una desviación estándar de 7.4. El modelo de Rasch obtuvo un promedio de 47.7 y 7.3 en 6° grado.

Análisis 3. Equiparación de escalas de grado 3^0 con grado 5^0

Al igual que en la equiparación de 5° y 6° grados, el desarrollo de la escala de grado tres, mediante su equiparación con la escala de grado 5, se realizó usando los métodos lineales de Tucker y Levine. Los parámetros lineales de 5° se tomaron como parámetros de origen, para calibrar las puntuaciones a escala del grado 3°, como se hizo en la equiparación de 6° con 5°.

Los datos crudos iniciales arrojaron las siguientes medidas de tendencia central, confiabilidades y correlaciones de las puntuaciones operacionales y de equiparación para 5° y 3er. grados de la prueba operacional y los reactivos comunes (aplicación 1999):

TABLA 5

PRUEBA OPERACIONAL:

	5 ° grado	3er. grado
PROMEDIO=	34.98	33.879
D. S. =	12.076	12.014
N =	57187	60679
CONFIABILIDAD =	0.884	0.885

REACTIVOS DE EQUIPARACIÓN:

	5 ° grado	6 ° grado
PROMEDIO =	4.285	3.699
D. S. =	1.883	1.832
N =	57187	60679
CONFIABILIDAD=	0.512	0.488
CORR OP/EQ=	0.56	0.708

En la tabla 5 se observa que los promedios, desviaciones y confiabilidad operacionales, no difieren significativamente. No obstante se observa que la correlación de los resultados en los reactivos de equiparación es significativamente más baja en 5° que en 3er. grado (.56 y .708 respectivamente).

La tabla 6 provee la reestimación de las medidas de tendencia central y los nuevos parámetros de conversión lineal equiparados usando los métodos de Tucker y Levine respectivamente:

TABLA 6
Reestimación de medidas de tendencia central de las puntuaciones crudas y parámetros usando métodos de Tucker y Levine

Equiparación de 3º. con 5to.
TUCKER METHOD:

TUCKER 5° GRADO =3.59137546	TUCKER 3º =4.64296507

Mean, variance and standard desviation estimates .

	5° grado	3° grado
MEAN=	33.8965514	35.1990847
VARIANCE=	144.563339	146.303848
S.D. =	12.0234495	12.0956128
PARAMETERS:	A PARAM.	B PARAM:
3° GRADO TO 5° GRADO =	0.99403393	-1.0925332
5° GRADO TO SCALED=	0.6136	24.6157
3° GRADO TO SCALED=	0.60993922	23.9453216

LEVINE METHOD:

LEVINE 5° = 8.42682425	LEVINE 3°= 8.83128464

Mean, variance and standard deviation estimates:

	5° GRADO	3° GRADO
MEAN=	32.43779	36.3899049
VARIANCE=	138.857255	151.454971
S.D.=	11.7837708	12.3067043

PARAMETERS:	A PARAM:	B PARAM:
3° GRADO TO 5° GRADO =	0.95750824	-2.4058439
5° GRADO TO SCALED=	0.6136	24.6157
3° GRADO TO SCALED=	0.58752706	23.1394742

La tabla 7 provee los parámetros y valores lineales derivados de los mismos, para los métodos de equiparación de Tucker y Levine aplicados al grado 3.

TABLA 7
Parámetros lineales para tercer grado, usando métodos de equiparación Tucker y Levine

CONVERSIÓN DE PUNTUACIONES CRUDAS A ESCALA		
Parámetros:		
	TUCKER	LEVINE
A =	0.60994	0.58753
B=	23.94532	23.13947

PUNTUACIONES A ESCALA

PUNTUACIÓN CRUDA	TUCKER	LEVINE
80	72.7405	70.1416
79	72.1305	69.5541
78	71.5206	68.9666
77	70.9106	68.3791
76	70.3007	67.7915
75	69.6908	67.2040
74	69.0808	66.6165
73	68.4709	66.0289
72	67.8609	65.4414
71	67.2510	64.8539
70	66.6411	64.2664
69	66.0311	63.6788
68	65.4212	63.0913
67	64.8112	62.5038
66	64.2013	61.9163
65	63.5914	61.3287
64	62.9814	60.7412
63	62.3715	60.1537
62	617616	59.5662
61	61.1516	58.9786
60	60.5417	58.3911
59	59.9317	57.8036
58	59.3218	57.2160
57	58.7119	56.6285
56	58.1019	56.0410
55	57.4920	55.4535
54	56.8820	54.8659
53	56.2721	54.2784
52	55.6622	53.6909
51	55.0522	53.1034
50	54.4423	52.5158
49	53.8323	51.9283
48	53.2224	51.3408
47	52.6125	50.7532
46	52.0025	50.1657
45	51.3926	49.5782
44	50.7826	48.9907
43	50.1727	48.4031
42	49.5628	47.8156
41	48.9528	47.2281

40	48.3429	46.6406
39	47.7330	46.0530
38	47.1230	45.4655
37	46.5131	44.8780
36	45.9031	44.2904
35	45.2932	43.7029
34	44.6833	43.1154
33	44.0733	42.5279
32	43.4634	41.9403
31	42.8534	41.3528
30	42.2435	40.7653
29	41.6336	40.1778
28	41.0236	39.5902
27	40.4137	39.0027
26	39.8037	38.4152
25	39.1938	37.8277
24	38.5839	37.2401
23	37.9739	36.6526
22	37.3640	36.0651
21	36.7540	35.4775
20	36.1441	34.8900
19	35.5342	34.3025
18	34.9242	33.7150
17	34.3143	33.1274
16	33.7043	32.5399
15	330944	31.9524
14	32.4845	31.3649
13	31.8745	30.7773
12	31.2646	30.1898
11	30.6547	29.6023
10	30.0447	29.0147
9	29.4348	28.4272
8	28.8248	27.8397
7	28.2149	27.2522
6	27.6050	26.6646
5	26.9950	26.0771
4	26.3851	25.4896
3	25.7751	24.9021
2	25.1652	24.3145
1	24.5553	23.7270
0	23.9453	23.1395

En los resultados de equiparación de tercer grado, en la tabla 7, se observan mayores diferencias entre los parámetros de los métodos Tucker y Levine que en sexto grado. Pruebas de homoscedasticidad, mínimos cuadrados y otras pruebas, sugieren que el método Levine produce menos error, se ajusta mejorar a los datos, y tiene mayor probabilidad de representar una mejor equiparación que el método Tucker. Los resultados sugieren que los datos confirman los supuestos del método Levine.

Al aplicar los parámetros derivados de la equiparación del método Levine a los datos crudos de 3er grado, se obtuvo un promedio a escala (20-80) de 43 y una desviación estándar de 7.1. El modelo de Rasch obtuvo un promedio de 42.6 y 7.7 en 3er grado.

Análisis 4. Comparaciones del modelo Rasch con métodos Tucker y Levine

La siguiente tabla resume los resultados de promedios y desviaciones estándar para el modelo Rasch, y los métodos de Tucker y Levine.

TABLA 8
Comparación de medidas de tendencia central para los distintos métodos de equiparación

	Pomedios			Desviaciones Estándar		
	Modelo	Método	Método	Modelo	Método	Método
GRADO	Rasch	Tucker	Levine	Rasch	Tucker	Levine
3ro.	**42.56**	44.61	**43.04**	7.77	7.33	**7.06**
6°	**47.73**	**47.8**	47.9	7.41	**7.398**	7.396

El grupo ancla (grupo base) de quinto grado, obtuvo un promedio a escala de 46.08 y desviación estándar de 7.41.

La tabla 9 compara los resultados a escala de tercer grado, para los métodos de Tucker, Levine y Rasch.

TABLA 9
Tercer grado. Comparación métodos Tucker y Levine, 3° 1999, con modelo Rasch

PARÁMETROS

	TUCKER	LEVINE
A =	0.60994	**0.58753**
B =	23.94532	**23.13947**

PUNTUACIONES A ESCALA					
Punt. Cruda	TUCKER	**LEVINE**	**RASCH**	ERROR	Punt. Cruda
80	72.7	**70.1**			80
79	72.1	**69.6**			79
78	71.5	**69.0**			78
77	70.9	**68.4**	**80.0**	5.9	77
76	70.3	**67.8**	**77.6**	5.2	76
75	69.7	**67.2**	**75.2**	4.7	75
74	69.1	**66.6**	**73.2**	4.3	74
73	68.5	**66.0**	**71.5**	4.0	73
72	67.9	**65.4**	**69.9**	3.8	72
71	67.3	**64.9**	**68.6**	3.6	71
70	66.6	**64.3**	**67.3**	3.5	70
69	66.0	**63.7**	**66.2**	3.3	69
68	65.4	**63.1**	**65.1**	3.2	68
67	64.8	**62.5**	**64.1**	3.1	67
66	64.2	**61.9**	**63.2**	3.0	66
65	63.6	**61.3**	**62.3**	3.0	65
64	63.0	**60.7**	**61.4**	2.9	64
63	62.4	**60.2**	**60.6**	2.8	63
62	61.8	**59.6**	**59.8**	2.8	62

61	61.2	**59.0**	**59.1**	2.7	61
60	60.5	**58.4**	**58.4**	2.7	60
59	59.9	**57.8**	**57.7**	2.6	59
58	59.3	**57.2**	**57.0**	2.6	58
57	58.7	**56.6**	**56.3**	2.6	57
56	58.1	**56.0**	**55.7**	2.5	56
55	57.5	**55.5**	**55.0**	2.5	55
54	56.9	**54.9**	**54.4**	2.5	54
53	56.3	**54.3**	**53.8**	2.5	53
52	55.7	**53.7**	**53.2**	2.5	52
51	55.1	**53.1**	**52.6**	2.4	51
50	54.4	**52.5**	**52.0**	2.4	50
49	53.8	**51.9**	**51.4**	2.4	49
48	53.2	**51.3**	**50.8**	2.4	48
47	52.6	**50.8**	**50.3**	2.4	47
46	52.0	**50.2**	**49.7**	2.4	46
45	51.4	**49.6**	**49.2**	2.4	45
44	50.8	**49.0**	**48.6**	2.4	44
43	50.2	**48.4**	**48.0**	2.4	43
42	49.6	**47.8**	**47.5**	2.4	42
41	49.0	**47.2**	**46.9**	2.4	41
40	48.3	**46.6**	**46.4**	2.4	40
39	47.7	**46.1**	**45.8**	2.4	39
38	47.1	**45.5**	**45.3**	2.4	38
37	46.5	**44.9**	**44.7**	2.4	37
36	45.9	**44.3**	**44.2**	2.4	36
35	45.3	**43.7**	**43.6**	2.4	35
34	44.7	**43.1**	**43.1**	2.4	34
33	44.1	**42.5**	**42.5**	2.4	33
32	43.5	**41.9**	**41.9**	2.4	32
31	42.9	**41.4**	**41.4**	2.4	31
30	42.2	**40.8**	**40.8**	2.4	30
29	41.6	**40.2**	**40.2**	2.4	29
28	41.0	**39.6**	**39.6**	2.5	28
27	40.4	**39.0**	**39.0**	2.5	27
26	39.8	**38.4**	**38.3**	2.5	26
25	39.2	**37.8**	**37.7**	2.5	25
24	38.6	**37.2**	**37.1**	2.6	24
23	38.0	**36.7**	**36.4**	2.6	23
22	37.4	**36.1**	**35.7**	2.6	22
21	36.8	**35.5**	**35.0**	2.7	21
20	36.1	**34.9**	**34.3**	2.7	20
19	35.5	**34.3**	**33.6**	2.7	19
18	34.9	**33.7**	**32.8**	2.8	18
17	34.3	**33.1**	**32.0**	2.9	17
16	33.7	**32.5**	**31.2**	2.9	16
15	33.1	**32.0**	**30.3**	3.0	15
14	32.5	**31.4**	**29.4**	3.1	14
13	31.9	**30.8**	**28.4**	3.1	13
12	31.3	**30.2**	**27.4**	3.2	12
11	30.7	**29.6**	**26.3**	3.4	11
10	30.0	**29.0**	**25.2**	3.5	10
9	29.4	**28.4**	**23.9**	3.7	9
8	28.8	**27.8**	**22.5**	3.8	8
7	28.2	**27.3**	**20.9**	4.1	7
6	27.6	**26.7**	**20.0**	4.3	6
5	27.0	**26.1**			5
4	26.4	**25.5**			4
3	25.8	**24.9**			3
2	25.2	**24.3**			2
1	24.6	**23.7**			1
0	23.9	**23.1**			0
Puntuación cruda	TUCKER	**LEVINE**	**RASCH**	ERROR	Puntuación cruda

La siguiente tabla compara los resultados de 6° grado para los métodos Tucker, Levine y Rasch:

TABLA 10
Sexto grado. Comparación métodos Tucker y Levine, 6° 1999, con modelo Rasch, *converted scores for new form*

Parámetros:

	TUCKER	LEVINE
A =	0.71819	0.718
B =	24.41377	24.55914

Puntuaciones a Escala :					
Punt Cruda	**TUCKER**	LEVINE	**RASCH**	ERROR	Punt. Cruda
79	**81.2**	81.3	**80**	14.2	79
78	**80.4**	80.6	**80**	10.1	78
77	**79.7**	79.8	**80**	7.2	77
76	**79.0**	79.1	**80**	6.0	76
75	**78.3**	78.4	**80**	5.2	75
74	**77.6**	77.7	**80**	4.7	74
73	**76.8**	77.0	**80.0**	4.4	73
72	**76.1**	76.3	**79.4**	4.1	72
71	**75.4**	75.5	**77.8**	3.9	71
70	**74.7**	74.8	**76.4**	3.7	70
69	**74.0**	74.1	**75.1**	3.5	69
68	**73.3**	73.4	**73.9**	3.4	68
67	**72.5**	72.7	**72.8**	3.3	67
66	**71.8**	71.9	**71.7**	3.2	66
65	**71.1**	71.2	**70.7**	3.1	65
64	**7.04**	70.5	**69.8**	3.0	64
63	**69.7**	69.8	**68.9**	3.0	63
62	**68.9**	69.1	**68.0**	2.9	62
61	**68.2**	68.4	**67.2**	2.9	61
60	**67.5**	67.6	**66.4**	2.8	60
59	**66.8**	66.9	**65.7**	2.8	59
58	**66.1**	66.2	**64.9**	2.7	58
57	**65.4**	65.5	**64.2**	2.7	57
56	**64.6**	64.8	**63.5**	2.7	56
55	**63.9**	64.0	**62.8**	2.6	55
54	**63.2**	63.3	**62.1**	2.6	54
53	**62.5**	62.6	**61.4**	2.6	53
52	**61.8**	61.9	**60.8**	2.6	52
51	**61.0**	61.2	**60.1**	2.5	51
50	**60.3**	60.5	**59.5**	2.5	50
49	**59.6**	59.7	**58.9**	2.5	49
48	**58.9**	59.0	**58.2**	2.5	48
47	**58.2**	58.3	**57.6**	2.5	47
46	**57.5**	57.6	**57.0**	2.5	46
45	**56.7**	56.9	**56.4**	2.5	45
44	**56.0**	56.2	**55.8**	2.5	44
43	**55.3**	55.4	**55.2**	2.5	43
42	**54.6**	54.7	**54.6**	2.4	42
41	**53.9**	54.0	**54.0**	2.4	41
40	**53.1**	53.3	**53.4**	2.4	40
39	**52.4**	52.6	**52.8**	2.4	39
38	**51.7**	51.8	**52.2**	2.4	38

37	**51.0**	51.1	**51.6**	2.5	37
36	**50.3**	50.4	**51.0**	2.5	36
35	**49.6**	49.7	**50.4**	2.5	35
34	**48.8**	49.0	**49.8**	2.5	34
33	**48.1**	48.3	**49.2**	2.5	33
32	**47.4**	47.5	**48.6**	2.5	32
31	**46.7**	46.8	**48.0**	2.5	31
30	**46.0**	46.1	**47.3**	2.5	30
29	**45.2**	45.4	**46.7**	2.5	29
28	**44.5**	44.7	**46.1**	2.6	28
27	**43.8**	43.9	**45.4**	2.6	27
26	**43.1**	43.2	**44.7**	2.6	26
25	**42.4**	42.5	**44.1**	2.6	25
24	**41.7**	41.8	**43.4**	2.6	24
23	**40.9**	41.1	**42.7**	2.7	23
22	**40.2**	40.4	**41.9**	2.7	22
21	**39.5**	39.6	**41.2**	2.7	21
20	**38.8**	38.9	**40.4**	2.8	20
19	**38.1**	38.2	**39.7**	2.8	19
18	**37.3**	37.5	**38.8**	2.9	18
17	**36.6**	36.8	**38.0**	2.9	17
16	**35.9**	36.0	**37.1**	3.0	16
15	**35.2**	35.3	**36.2**	3.1	15
14	**34.5**	34.6	**35.3**	3.1	14
13	**33.8**	33.9	**34.3**	3.2	13
12	**33.0**	33.2	**33.2**	3.3	12
11	**32.3**	32.5	**32.1**	3.4	11
10	**31.6**	31.7	**30.8**	3.6	10
9	**30.9**	31.0	**29.5**	3.7	9
8	**30.2**	30.3	**28.1**	3.9	8
7	**29.4**	29.6	**26.5**	4.1	7
6	**28.7**	28.9	**24.7**	4.4	6
5	**28.0**	28.1	**22.6**	4.8	5
4	**27.3**	27.4	**20.1**	5.3	4
3	**26.6**	26.7	**20**	6.0	3
2	**25.9**	26.0	**20**	7.3	2
1	**25.1**	25.3	**20**	10.1	1
0	**24.4**	24.6	**20**	14.2	0
Puntuación cruda	**TUCKER**	LEVINE	**RASCH**	ERROR	Puntuación cruda

La tabla 11 resume los resultados para las tres escalas en los tres niveles (Levine para 3º y Tucker para 6º grados).

TABLA 11
Comparación métodos lineales con Rasch

	LEVINE	TUCKER
A =	**0.5875** A=.6136	**0.71819**
B=	**23.139** B=24.6157	**24.4138**

	3º 99	5° 99	6° 99	3º 99	5º 99	6º 99
Raw scores	LEVINE	SCALING	**TUCKER**	RASCH	RASCH	**RASCH**
83		75.5				
82		74.9				
81		74.3				
80	**70.1**	73.7				
79	**69.6**	73.1	**81.2**		80	
78	**69.0**	72.5	**80.4**		79.7	
77	**68.4**	71.9	**79.7**	**80.0**	77.7	
76	**67.8**	71.2	**79.0**	**77.6**	75.9	
75	**67.2**	70.6	**78.3**	**75.2**	74.4	
74	**66.6**	70.0	**77.6**	**73.2**	73.0	
73	**66.0**	69.4	**76.8**	**71.5**	71.8	**80.0**
72	**65.4**	68.8	**76.1**	**69.9**	70.6	**79.4**
71	**64.9**	68.2	**75.4**	**68.6**	69.5	**77.8**
70	**64.3**	67.6	**74.7**	**67.3**	68.5	**76.4**
69	**63.7**	67.0	**74.0**	**66.2**	67.6	**75.1**
68	**63.1**	66.3	**73.3**	**65.1**	66.7	**73.9**
67	**65.2**	65.7	**72.5**	**64.1**	65.8	**72.8**
66	**61.9**	65.1	**71.8**	**63.2**	65.0	**71.7**
65	**61.3**	64.5	**71.1**	**62.3**	64.2	**70.7**
64	**60.7**	63.9	**70.4**	**61.4**	63.5	**69.8**
63	**60.2**	63.3	**69.7**	**60.6**	62.8	**68.9**
62	**59.6**	62.7	**68.9**	**59.8**	62.1	**68.0**
61	**59.0**	62.0	**68.2**	**59.1**	61.4	**67.2**
60	**58.4**	61.4	**67.5**	**58.4**	60.7	**66.4**
59	**57.8**	60.8	**66.8**	**57.7**	60.1	**65.7**
58	**57.2**	60.2	**66.1**	**57.0**	59.4	**64.9**
57	**56.6**	59.6	**65.4**	**56.3**	58.8	**64.2**
56	**56.0**	59.0	**64.6**	**55.7**	58.2	**63.5**
55	**55.5**	58.4	**63.9**	**55.0**	57.6	**62.8**
54	**54.9**	57.8	**63.2**	**54.4**	57.0	**62.1**
53	**54.3**	57.1	**62.5**	**53.8**	56.4	**61.4**
52	**53.7**	56.5	**61.8**	**53.2**	55.9	**60.8**
51	**53.1**	55.9	**61.0**	**52.6**	55.3	**60.1**
50	**52.5**	55.3	**60.3**	**52.0**	57.7	**59.5**
49	**51.9**	54.7	**59.6**	**51.4**	54.2	**58.9**
48	**51.3**	54.1	**58.9**	**50.8**	53.6	**58.2**
47	**50.8**	53.5	**58.2**	**50.3**	53.1	**57.6**
46	**50.2**	52.8	**57.5**	**49.7**	52.5	**57.0**
45	**49.6**	52.2	**56.7**	**49.2**	52.0	**56.4**
44	**49.0**	51.6	**56.0**	**48.6**	51.5	**55.8**
43	**48.4**	51.0	**55.3**	**48.0**	50.9	**55.2**
42	**47.8**	50.4	**54.6**	**47.5**	50.4	**54.6**
41	**47.2**	49.8	**53.9**	**46.9**	49.9	**54.0**
40	**46.6**	49.2	**53.1**	**46.4**	49.3	**53.4**
39	**46.1**	48.5	**52.4**	**45.8**	48.8	**52.8**

38	**45.5**	47.9	**51.7**	**45.3**	48.2	**52.2**
37	**44.9**	47.3	**51.0**	**44.7**	47.7	**51.6**
36	**44.3**	46.7	**50.3**	**44.2**	47.1	**51.0**
35	**43.7**	46.1	**49.6**	**43.6**	46.6	**50.4**
34	**43.1**	45.5	**48.8**	**43.1**	46.0	**49.8**
33	**42.5**	44.9	**48.1**	**42.5**	45.5	**49.2**
32	**41.9**	44.3	**47.4**	**41.9**	44.9	**48.6**
31	**41.4**	43.6	**46.7**	**41.4**	44.3	**48.0**
30	**40.8**	43.0	**46.0**	**40.8**	43.8	**47.3**
29	**40.2**	42.4	**45.2**	**40.2**	43.2	**46.7**
28	**39.6**	41.8	**44.5**	**39.6**	42.6	**46.1**
27	**39.0**	41.2	**43.8**	**39.0**	42.0	**45.4**
26	**38.4**	40.6	**43.1**	**38.3**	41.3	**44.7**
25	**37.8**	40.0	**42.4**	**37.7**	40.7	**44.1**
25	**37.2**	39.3	**41.7**	**37.1**	40.1	**43.4**
23	**36.7**	38.7	**40.9**	**36.4**	39.4	**42.7**
22	**36.1**	38.1	**40.2**	**35.7**	38.7	**41.9**
21	**35.5**	37.5	**39.5**	**35.0**	38.0	**41.2**
20	**34.9**	36.9	**38.8**	**34.3**	37.3	**40.4**
19	**343**	36.3	**38.1**	**33.6**	36.6	**39.7**
18	**33.7**	35.7	**37.3**	**32.8**	35.8	**38.8**
17	**33.1**	35.0	**36.6**	**32.0**	35.0	**38.0**
16	**32.5**	34.4	**35.9**	**31.2**	34.2	**37.1**
15	**32.0**	33.8	**35.2**	**30.3**	33.3	**36.2**
14	**31.4**	33.2	**34.5**	**29.4**	32.4	**35.3**
13	**30.8**	32.6	**33.8**	**28.4**	31.4	**34.3**
12	**30.2**	32.0	**33.0**	**27.4**	30.4	**33.2**
11	**29.6**	31.4	**32.3**	**26.3**	29.3	**32.1**
10	**29.0**	30.8	**31.6**	**25.2**	28.1	**30.8**
9	**28.4**	30.1	**30.9**	**23.9**	26.8	**29.5**
8	**27.8**	29.5	**30.2**	**22.5**	25.4	**28.1**
7	**27.3**	28.9	**29.4**	**20.9**	23.8	**26.5**
6	**26.7**	28.3	**28.7**	**20.0**	22.1	**24.7**
5	**26.1**	27.7	**28.0**		20.0	**22.6**
4	**25.5**	27.1	**27.3**			**20.1**
3	**24.9**	26.5	**26.6**			
2	**24.3**	25.8	**25.9**			
1	**23.7**	25.2	**25.1**			

En la tabla 11 se observa que para los modelos lineales, los alcances de cada escala aumentan de acuerdo con el grado, siendo el máximo 70.1 para 3°, 75.5 para 5º y 81.2 (80) para sexto. Los modelos Rasch, en cambio, ofrecen los mismos límites, lo cual sugiere que es probable que estén sobreestimados los niveles de habilidad para las puntuaciones más altas, aunque hay una alta correspondencia para las puntuaciones en los rangos de 30 a 70 puntos a escala (dos desviaciones estándar), siendo más discrepantes los resultados para la tercera desviación estándar.

Los resultados de las tablas anteriores se pueden apreciar en las gráficas 1 a 4. El eje de X (horizontal) de estas gráficas corresponde a la puntuación cruda, y el eje de Y (vertical) corresponde a la puntuación a escala (20-80).

En las gráficas 1 y 2 se observa la alta concordancia entre las puntuaciones a escala, estimadas con los métodos Tucker y Levine y el modelo Rasch. Las gráficas 3 y 4 muestran

las diferencias entre grados (3º, 5º y 6°), para los métodos lineales y el modelo Rasch, respectivamente.

Las gráficas 5 a la 7 contrastan las escalas de 20 a 80, derivadas del modelo lineal correspondiente (Levine para 3º, eje principal para 5º (grupo ancla), y Tucker para 6º),

Los ejes de X y Y de estas gráficas, corresponden a las puntuaciones a escala (20-80) para el modelo Rasch (ejes X) y el método lineal correspondiente (eje Y).

La gráfica 8 representa las distribuciones en los grados 3º, 5° y 6º obtenidas, utilizando las puntuaciones a escala (20-80) lineales y mediante el modelo Rasch. Esta gráfica de *caja y bigote* representa lo siguiente: cada *caja* representa al 50 por ciento de la población, la línea que divide a cada *caja* representa las mediana de cada distribución y los *bigotes* representan los cuarteles (25 por ciento) superior e inferior de las distribuciones, respectivamente.

En la gráfica 8 se observa un crecimiento sistemático de las distribuciones a través de los grados.

Conclusiones generales

Los análisis que se informan en este documento, permiten comparar distintos métodos y modelos para el desarrollo de las puntuaciones a escala. Los aspectos que se destacan de los mismos son:

- Alta correspondencia entre las puntuaciones a escala estimadas, utilizando distintos métodos lineales de esta investigación, y las puntuaciones a escala derivadas del modelo Rasch por el Proyecto de Estándares Nacionales.
- Errores mínimos y no significativos para la mayor parte de la distribución de puntuaciones obtenidas mediante métodos lineales, y los métodos usados por el proyecto de *Estándares Nacionales*.
- El hecho de que los resultados se ajusten mejor al método Levine para tercer grado, y a el método Tucker para quinto grado, sugiere que entre el grupo base (grupo ancla) de 5º y el grupo de 6º, hay mayor comunalidad que entre 3º y 5º. Esto último, parece quedar corroborado también por las diferencias de correlación de los reactivos ancla 3º=5º, con sus respectivas pruebas operacionales, que resultaron ser de 71. en 3º y 5.6 en 5º (ver tabla 5) y por los resultados de reestimación de los métodos Tucker y Levine. También estos resultados sugieren que las poblaciones 3º-5º difieren en habilidad, y las pruebas difieren en dificultad. Estas diferencias no son tan significativas al contrastar los resultados 5º-6º, donde las correlaciones de los reactivos ancla, con su respectiva prueba operacional, fueron de .94 y .95

Limitaciones y recomendaciones

Se observaron diferencias significativas entre las puntuaciones a escala de los métodos y modelos, para las puntuaciones en los límites de las distribuciones (en la tercera desviación estándar). Se observaron probables sobreestimaciones a escala para las puntuaciones de 70 o más (en la escala de 20 a 80 puntos). En este sector de la escala, los modelos lineales parecen proveer mejores estimados, es decir, estimados con menos error.

Los modelos lineales (Tucker y Levine) mostraron diferencias en los limites de escala (70.1 75.5 y 81.2 para 3º, 50 y 60 respectivamente). Estas diferencias no se observaron en los resultados del modelo Rasch. Es probable que el modelo Rasch esté sobreestimando las puntuaciones en los límites superiores de la escala. Éste es uno de los problemas que merece mayor análisis para, de ser necesario, realizar las correcciones correspondientes mediante el uso de modelos combinados, bisectores u otras soluciones.

Se recomienda utilizar operacionalmente más de un modelo de equiparación. Esto sirve para someter a prueba los resultados del modelo seleccionado, previo a su uso operacional, y provee opciones para estimar con mayor precisión las puntuaciones a escala, de ser necesario.

Referencias

Libros:

Angoff, W.A. "Scales. Norms and equivalent scores". En R.L. Thorndike (Ed.), *Educational Measurement,* (2da Ed., pag. 508-600), Washington, American Council on Education. 1971.

Holland, P.: Rubin. D. *Test Equating*. Academia Press, 1982.

Kolen. M, Brennan, R. *Test Equating: Methods and Practices*. Springer, 1995.

Revistas:

Han, T., Kolen, M., & Pohlmann, J. (1997). A Compararison Among Irt True – and Observed-Score Equatings and Tradicional Equipercentile Equating. *Applied Measurement in education, 10* (2), 105-121.

Harris, D.J., & Crouse, J.D. (1993). A Study of Criteria Used in Equating. *Applied Measurement in Education, 6* (3), 195-240.

Harris. D.J., & Crouse, J.D. (1993). A Study of Criteria used in Equating. *Applied Measurement in Education, 6* (3), 195-240.

Linn, R.L. (1993). Linking results of Distinet Assessments. *Applied Measumement in Education, 6 (1),* 83-102.