La Confiabilidad en una Evaluación Mediante Rúbrica de Ensayos de Escritura del Español de Egresados de las Escuelas Primarias en Baja California.

Luis Ángel Contreras Nino, Manuel Jorge González-Montesinos Martínez,
Erick Urias Luzanilla

Introducción.

En Baja California se han impulsado diversas acciones orientadas a evaluar la situación que guarda la calidad de la educación en el nivel básico. En su mayor parte, dichas acciones se refieren a la aplicación en la entidad de instrumentos diseñados por la SEP o el INEE, como es el caso de exámenes como las pruebas de Estándares Nacionales, el IDANIS, los EXCALE o, más recientemente ENLACE; así como aquellas evaluaciones de carácter internacional aplicadas en nuestro país, como PISA y TIMSS. No obstante, también se han desarrollado en la entidad instrumentos para evaluar el aprendizaje como el Examen Criterial de Español al que se refiere el presente trabajo. De hecho se trata del primer examen criterial a gran escala elaborado en Baja California (Contreras, 2000).

Por su condición criterial el examen se alineó al contenido del área de español del currículum de la educación primaria. En consecuencia se orientó a evaluar el dominio de conocimientos y habilidades en dicha área de español logrado por los egresados de primaria en el Estado. Es decir, explora el dominio de:

- Lengua hablada
- Lengua escrita
- Reflexión sobre la lengua
- Recreación Literaria

Por tratarse de una prueba de gran escala, las especificaciones se orientaron a producir ítems de respuesta seleccionada, de opción múltiple. Sin embargo, el enfoque del nuevo currículum del área de español que se empezó a adoptar a partir de 1993, enfatizaba en gran medida el desarrollo de habilidades comunicativas de producción, principalmente en los ejes de lengua hablada, lengua escrita y recreación literaria, mismas que para su evaluación apropiada requerían de ítems de respuesta construida, como los de ejecución. En consecuencia, el comité que diseño la prueba tomó la decisión de incorporar al examen una especificación de ítems para producir un reactivo de ejecución orientado a capturar una muestra de la

habilidad para redactar de los niños, pues dicha habilidad se consideró demasiado importante en el contexto del currículo como para ser ignorada en el examen.

En cuanto a la estructura del examen, estuvo integrado por cuatro versiones de la prueba, cada una con 44 ítems de opción múltiple y un ítem de respuesta construida; es decir, el que solicitó a los examinados producir un ensayo que capturó una muestra de su escritura. Para ilustrar dicho ítem enseguida se presenta uno de los 4 instigadores que fueron utilizados para ello.

> "Escribe una composición donde describas a la persona de tu familia en quien más confías. Explica por qué esta persona es digna de tu confianza. Piensa que tu lector no sabe nada sobre esta persona; por lo tanto, piensa en los elementos que debes incluir en tu composición para que el lector tenga una imagen clara de ella. Utiliza oraciones completas, y no descuides la puntuación y la ortografía. Una vez que termines tu composición, léela y corrige tus errores. Escribe un mínimo de 10 renglones."

El examen se aplicó a fines de 2000 a una muestra estatal conformada por 3,151 niños que egresaron en el ciclo escolar 1999-2000 de 48 escuelas de educación primaria en Baja California, cuando ingresaron a la secundaria.

Dadas sus características, la evaluación de la ejecución en los ítems de respuesta construida utilizados en la evaluación a gran escala resulta compleja, tardada y costosa. Los principales requerimientos para ello, incluyen el desarrollo de rúbricas, la selección y entrenamiento de jueces expertos, y la operación de prolongadas sesiones de evaluación. En consecuencia, tras su aplicación a gran escala, los ensayos de los niños no pudieron ser evaluados por falta de recursos humanos y financieros sino hasta recientemente.

Para ello se hizo una convocatoria a egresados y estudiantes avanzados de la carrera Lengua y Literatura de Hispanoamérica de la UABC y a profesores de la Escuela Normal de Ensenada, a quienes de dio un entrenamiento específico en el manejo de matrices de evaluación de la escritura (Rúbricas), mismo que incluyó la capacitación de dichos especialistas a fin de garantizar índices aceptables de confiabilidad en sus observaciones; además de un entrenamiento en el manejo de procedimientos automatizados para la

captura en línea de sus juicios en una página Web que se diseño para tal efecto (Ver figura 1). Así, 31 jueces calificaron cada uno de ellos 100 ensayos durante aproximadamente un mes.

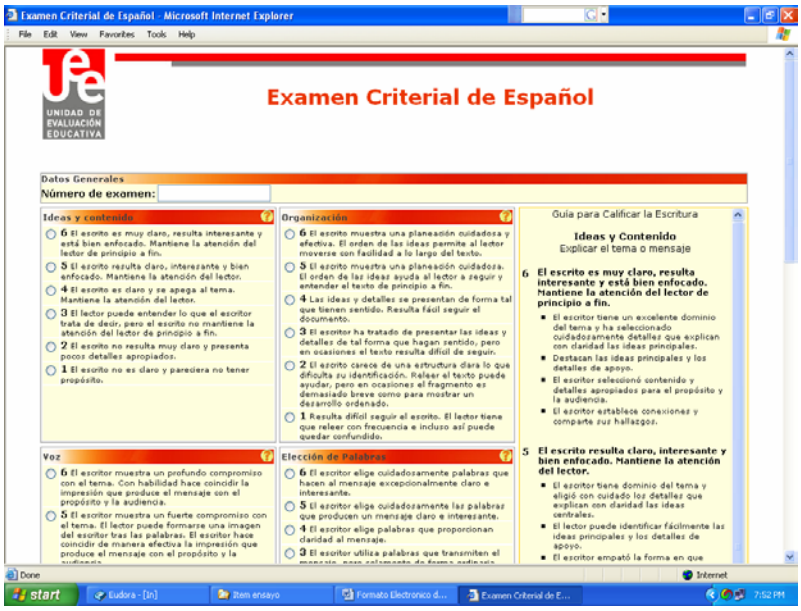

Figura 1 Pantalla de captura en línea de las evaluaciones de los jueces

En cuanto a la matriz de evaluación utilizada, se consideró que dadas ciertas características como su brevedad, respuesta a un instigador y otras relacionadas, para evaluar los escritos de los niños podría utilizarse una rúbrica desarrollada dentro del modelo de escritura de 6+1 rasgos, propuesto a mediados de los años ochenta del siglo pasado por investigadores del laboratorio psicométrico regional del noroeste de los Estados Unidos (NWREL) (Koslow y Bellany, 2005). De manera específica se tradujo y adaptó la rúbrica para evaluar la escritura que fue desarrollada en el estado norteamericano de Oregon para realizar la evaluación estatal de la escritura de niños de K3 a K5 (Wolfe et al, 1993), grados educativos entre los que se encuentra el que corresponde al egreso de la educación primaria en nuestro país. Enseguida se presenta una breve descripción de los rasgos de la rúbrica que fue utilizada para evaluar los escritos producidos por los examinados en respuesta a las preguntas de ensayo del examen de español.

Rasgo de la escritura evaluado	Descripción resumida
Ideas y contenido	***Explicar el tema o mensaje.*** Este rasgo establece el tema del escritor, el foco de su mensaje, junto con los detalles de apoyo que desarrollan y enriquecen dicho tema. Las ideas principales se comunican y apoyan mediante detalles informativos que muestran una exploración del tema apropiada para la audiencia y propósito comunicativo.
Organización	***Planear y emplear conexiones claras de principio a fin.*** Este atributo se refiere a la estructura interna del escrito, la cual incluye el hilo conductor del mensaje central y los patrones que mantienen unificado al escrito. La estructura organizativa puede estar basada en la comparación y el contraste, la cronología de un evento u otro patrón identificable. Cuando la organización es apropiada, el escrito crea en el lector un sentido de anticipación. Los eventos se suceden lógicamente, la información se dosifica de manera que el lector nunca pierde el interés o el panorama de lo que el escritor pretende. Las transiciones mueven al lector de un punto al siguiente.
Voz	***Proyectarse como una persona real.*** La voz es la manifestación del escritor a través de las palabras, su sello personal; el sentido de que una persona real se dirige a nosotros y de que tiene interés en el mensaje. Es el corazón y el alma del escrito; su magia, su humor, su sentimiento, su vida. Este rasgo muestra el interés y compromiso del escritor con el tema. La voz variará de acuerdo con el propósito y el tipo de texto, pero debe ser apropiadamente formal o casual, distante o íntima, dependiendo del propósito o la audiencia.
Elección de palabras	***Elegir con cuidado palabras para formar una imagen en la mente del lector.*** Este rasgo refleja el uso específico que hace el escritor de las palabras y oraciones, a fin de transportar el mensaje de manera interesante, precisa y natural; apropiada para la audiencia y el propósito. La elección de palabras no solo sirve para comunicar de manera funcional, sino que mueve al lector hacia una nueva visión de las cosas e ilumina y expande las ideas. La elección correcta de palabras se caracteriza por la habilidad de emplear con precisión palabras de uso cotidiano.
Fluidez de las oraciones	***Crear oraciones que hagan sentido y que luzcan articuladas cuando se leen en voz alta.*** Este rasgo da cuenta del ritmo y flujo del lenguaje, del sonido de los patrones de palabras y de la variedad de estructuras de las oraciones. ¿Cómo suena el escrito si se lee en voz alta? Esa es la prueba a realizar. Un escrito fluido tiene cadencia, poder, ritmo y movimiento.
Convenciones	***Utilizar de manera correcta la ortografía, puntuación, escritura de párrafos y demás reglas del español.*** Este atributo se refiere a la mecánica del escrito, al apego a las convenciones de la lengua. Un escrito apropiado, usualmente ha sido revisado y editado con cuidado. La caligrafía y limpieza no se califican como parte de este rasgo.

Por su parte, cada uno de los seis rasgos fue evaluado en una escala de seis niveles de ejecución. Con propósitos ilustrativos, a continuación se muestra la escala de ejecución que fue utilizada por los jueces para evaluar el primero de los rasgos de la escritura de los niños.

Ideas y contenido ***Explicar el tema o mensaje***		
6 **El escrito es muy claro, resulta interesante y está bien enfocado. Mantiene la atención del lector de principio a fin.** - El escritor tiene un excelente dominio del tema y ha seleccionado cuidadosamente detalles que explican con claridad las ideas principales - Destacan las ideas principales y los detalles de apoyo - El escritor seleccionó contenido y detalles apropiados para el propósito y la audiencia - El escritor establece conexiones y comparte sus hallazgos	**5** **El escrito resulta claro, interesante y bien enfocado. Mantiene la atención del lector.** - El escritor tiene dominio del tema y eligió con cuidado los detalles que explican con claridad las ideas centrales. - El lector puede identificar fácilmente las ideas principales y los detalles de apoyo. - El escritor empató la forma en que presenta el tema con el propósito y la audiencia. - El escritor establece conexiones y comparte sus hallazgos	**4** **El escrito es claro y se apega al tema. Mantiene la atención del lector.** - El escritor muestra conocimiento del tema y eligió detalles que ayudan a explicar la idea principal. - El lector puede identificar la idea principal y detalles de apoyo. - El lector percibe que el escritor está conciente del propósito y la audiencia. - El escritor establece algunas conexiones y pueden estar presentes nuevos hallazgos
3 **El lector puede entender lo que el escritor trata de decir, pero el escrito no mantiene la atención del lector de principio a fin.** - El escritor tiene cierto dominio del tema; algunas ideas pueden ser claras, mientras que otras no lo son o no parecen encajar. - El escrito no cuenta con suficientes detalles, son demasiado generales o no tienen relación con las ideas. - El lector ubica algunas formas en que el escrito empata con el propósito y la audiencia, pero no siempre resulta claro. - El escritor establece conexiones obvias y predecibles.	**2** **El escrito no resulta muy claro y presenta pocos detalles apropiados.** - El escritor tiene poco dominio del tema; las ideas no resultan claras. - El escrito presenta detalles limitados, que se repiten o que no tienen relación con las ideas. - El lector no está seguro del propósito ni de la idea central del escrito, pero puede hacer algunas suposiciones al respecto.	**1** **El escrito no es claro y pareciera no tener propósito.** - Las ideas del escritor son muy limitadas o pueden dispersarse en diversas direcciones. - Es difícil entender lo que el escritor realmente quiso decir.

Resultados Obtenidos

Estadístico	Total de la rúbrica	Ideas	Organización	Voz	Palabras	Fluidez	Convenciones
Media	18.9769	3.3518	3.1965	3.3545	3.0695	3.1139	2.8907
Desviación estándar	5.85798	1.2218	1.1455	1.2254	1.0873	1.0828	1.0513
N válido = 2993							

Modelo	Estadístico	ideas	organización	voz	palabras	fluidez	convenciones
1	Media	3.36	3.24	3.32	3.09	3.15	2.98
	Desviación estándar	1.271	1.222	1.247	1.165	1.147	1.109
	N válido = 756						
2	Media	3.45	3.34	3.53	3.22	3.26	3.11
	Desviación estándar	1.306	1.145	1.303	1.119	1.065	1.049
	N válido = 747						
3	Media	3.37	3.22	3.43	3.06	3.09	2.78
	Desviación estándar	1.192	1.170	1.244	1.091	1.122	.974
	N válido = 702						
4	Media	3.23	2.97	3.13	2.90	2.96	2.68
	Desviación estándar	1.154	1.048	1.116	.982	1.015	1.064
	N válido = 671						

Análisis de la Confiabilidad

La interpretación valida de los puntajes depende sustancialmente de la confiabilidad de los jueces que calificaron las ejecuciones. La confiabilidad de los jueces es una propiedad del proceso de medición y como tal esta debe ser determinada en cada ejercicio particular del proceso.

Para determinar la confiabilidad se identificaron tres aproximaciones psicométricas aplicables a este proceso.

Aproximación basada en la Consistencia entre Jueces

En esta aproximación se toma como base el cálculo el Coeficiente α de Cronbach como una medida del grado en que los puntajes asignados por jueces múltiples convergen para medir un constructo en común. El coeficiente α es una medida de consistencia interna de las calificaciones. Si el estimado resulta bajo, implica que la variabilidad en los puntajes combinados se debe a variabilidad atribuible al error aleatorio y no al puntaje verdadero en el constructo de interés (Hatcher, L. 1998, pp.130-140)

Para este caso se desagregó la base de datos del Ítem 45 para analizar los puntajes de los seis rasgos en cada Modelo por separado. De tal forma que los puntajes otorgados por los jueces quedan ordenados como sigue:

Los coeficientes obtenidos se muestran continuación:

Modelo	Alfa
1 (n=756)	.935
2 (n=747)	.927
3 (n=702)	.932
4 (n=671)	.909

Como los tamaños de muestra originales son considerables, se seleccionó una muestra aleatoria de 50 casos de cada modelo y se repitió el cálculo para descartar que los coeficientes se inflen artificialmente debido al tamaño de muestra en cada Forma / Modelo. Los resultados con n = 50 son:

Modelo	Alfa
1	.958
2	.923
3	.946
4	.943

Como puede observase en ambos casos los coeficientes Alfa son superiores a .90 lo que definitivamente constituye una razón sólida para concluir que las calificaciones otorgadas por los jueces son consistentes y convergen en un constructo común.

Aproximación basada en el Proceso de Medición

Otra aproximación alternativa a la confiabilidad entre jueces se basa en la información que proporciona el procedimiento de medición en su conjunto. Esta aproximación se implementa realizando un análisis de componentes principales (ACP) sobre los datos de cada escala. Este método es de aplicación óptima cuando la escala que se analiza se ha diseñado para medir un solo constructo unidimensional (e.g. competencia en escritura).

Los puntajes otorgados por los jueces se someten al ACP para determinar la cantidad de varianza compartida que se puede atribuir al primer componente extraído. El porcentaje de varianza explicada por el primer componente proporciona una indicación del grado que los jueces están coincidiendo. Si el

porcentaje de varianza es alto (e.g 60%), se tiene también una indicación de que los jueces están valorando un constructo unidimensional. (Stemler, 2004, p.9).

Se aplicó el procedimiento sobre las bases desagregadas por Forma / Modelo y se obtuvieron los siguientes resultados:

Modelo 1

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.545	75.747	75.747	**4.545**	**75.747**	**75.747**
2	.553	9.222	84.969			
3	.272	4.541	89.510			
4	.239	3.982	93.492			

Extraction Method: Principal Component Analysis.

Modelo 2

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.427	73.778	73.778	**4.427**	**73.778**	**73.778**
2	.556	9.273	83.051			
3	.345	5.753	88.805			
4	.272	4.534	93.339			

Extraction Method: Principal Component Analysis.

Modelo 3

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.485	74.752	74.752	**4.485**	**74.752**	**74.752**
2	.724	12.074	86.826			
3	.254	4.234	91.061			
4	.214	3.568	94.629			

Extraction Method: Principal Component Analysis.

Modelo 4

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.167	69.456	69.456	**4.167**	**69.456**	**69.456**
2	.680	11.338	80.795			
3	.354	5.905	86.700			
4	.314	5.234	91.933			

Extraction Method: Principal Component Analysis.

Como puede verse en los cuatro casos el procedimiento ACP extrajo solo un componente principal con valor eigen superior a 1. En todos los casos el porcentaje de varianza explicado por el componente es superior al 60%.

La ventaja de esta aproximación es que permite asignar puntajes finales a los examinados con base en la dimensión de mayor peso es decir el primer componente principal. La desventaja es que se asume que los puntajes de los jueces están exentos de error de medición. Esto último se debe a que una suposición inicial en todo ACP es que los datos sobre los que se calcula representan resultados sin error asociado a la medición misma

Aproximación a la Confiabilidad de Jueces basada en el Modelo Rasch

El modelo Rasch de facetas múltiples fue propuesto por J.M. Linacre, (1994) para analizar de forma exhaustiva la contribución individual de cada faceta y de cada elemento en un proceso de medición. En el caso particular del ítem de ensayo (145) las tres facetas son: los jueces, los examinados y los ítems de la escala para cada rasgo.

En particular el modelo permite derivar estimaciones de la severidad de los jueces, la habilidad de los examinados y la dificultad de los ítems.

Se presenta a continuación un ejemplo que compara la severidad de los jueces:

"Examen Espanol I45 Modelo1" 04-26-2007 13:30:06
Table 7.3.1 Item Measurement Report (arranged by mN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	PtBis	N Item
301	114	2.6	2.53	.89	.15	1.71	4.1	1.65	3.8	.37	.34	6 Convenciones
327	114	2.9	2.75	.32	.15	.89	-.8	.82	-1.3	1.15	.58	4 Palabras
345	114	3.0	2.89	-.05	.14	.66	-2.7	.71	-2.3	1.28	.64	5 Fluidez
352	114	3.1	2.95	-.19	.14	.86	-1.0	.90	-.7	1.10	.60	2 Organizacion
363	114	3.2	3.03	-.41	.14	.79	-1.5	.83	-1.2	1.15	.59	1 Ideas
371	114	3.3	3.10	-.56	.14	.84	-1.2	.85	-1.1	1.15	.62	3 Voz
343.2	114.0	3.0	2.87	.00	.14	.96	-.5	.96	-.5		.56	Mean (Count: 6)
23.4	.0	.2	.19	.48	.00	.34	2.2	.31	2.0		.10	S.D. (Populn)
25.6	.0	.2	.21	.53	.01	.38	2.4	.34	2.2		.11	S.D. (Sample)

Model, Populn: RMSE .14 Adj (True) S.D. .46 Separation 3.24 Reliability .91
Model, Sample: RMSE .14 Adj (True) S.D. .51 Separation 3.57 Reliability .93
Model, Fixed (all same) chi-square: 65.5 d.f.: 5 significance (probability): .00
Model, Random (normal) chi-square: 4.7 d.f.: 4 significance (probability): .32

Los estadígrafos de ajuste interno (INFIT) y externo (OUTFIT) indican el grado en que el comportamiento de cada juez se ajusta a las expectativas del modelo Rasch tomando en cuenta la calibración individual de severidad obtenida para cada juez. Por regla general los valores Infit y Outfit

MsSq deben mantenerse dentro del intervalo .80 a 1.30. Para los valores estandarizados ZStd el rango aceptable es de -2 a +2.

La aproximación vía FACETS permite además obtener una estimación de la consistencia intrajuez En particular, los valores de INFIT y OUTFIT proporcionan una medida del grado en que las calificaciones otorgadas por los jueces son internamente consistentes al mantenerse dentro de la expectativa que el modelo Rasch crea para cada juez dado su patrón observado de calificaciones emitidas dentro del proceso en su conjunto. Valores de IINFIT Mean Square mayores a 1.3 indican mayor variabilidad intrajuez que la que se esperaría con base al modelo. (Stemler, 2004, P.17).

Esta aproximación tiene la ventaja de no requerir que todos los jueces califiquen todos los ítems para lograr una estimación de la confiabilidad entre jueces. En lugar de ello, los jueces pueden calificar un subconjunto particular de ítems y, mientras exista suficiente conectividad (Linacre, 1994; Linacre, Englehard, Tatem y Myford, 1994) entre jueces y calificaciones, será posible comparar directamente a los jueces.

Un método para computar las diferencias en la severidad de los jueces e incorporarlas al puntaje final que resume la ejecución de cada participante, es mediante el uso del modelo de Rasch de muchas facetas. La tabla presenta un reporte de la medición de los jueces, generado por el programa de computo denominado FACETS (Linacre, 1994). La tabla presenta un indicador del nivel de severidad de cada juez en lo individual, junto con varios estadísticos de ajuste que ayudan a diagnosticar en qué medida cada juez fue consistente con su propio uso de la rúbrica para evaluar los ensayos. La utilidad de la tabla radica en que puede ser comparada simultáneamente la severidad relativa de todos los jueces.

Los índices de severidad de los jueces (measure) son útiles para estimar en qué medida existen diferencias sistemáticas entre jueces respecto a su nivel de severidad.

Además, la tabla 2 proporciona una serie de estadísticos de ajuste que son útiles para interpretar el grado de confiabilidad intrajuez. Los estadísticos infit de los jueces se interpretan de la misma manera en se hace en el caso con los estadísticos infit de los ítems o las personas (Bond y Fox, 2001; Wright y Stone, 1979).

Medias cuadráticas de Infit mayores a 1.3 indican que existe más variación no predicha en las respuestas de los jueces, de lo que cabría esperar según el modelo. Por su parte, medias cuadradas de Infit menores a 0.7 indican que existe menos variación en las respuestas de los jueces de la que podríamos predecir basados en el modelo.

Con los resultados FACETS así como con las dos anteriores aproximaciones a la confiabilidad es posible determinar lo grados de severidad o laxitud de los jueces y sobre todo verificar que estas diferencias se conserven dentro de un limite previsible lo que permite concluir que los resultados obtenidos de la evaluación son interpretables.

Bibliografía

Bond, T.G., Fox, Ch. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. New Jersey: Erlbaum.

Contreras, L. A. (2000). *Desarrollo y Pilotaje de un Examen de Español para la Educación Primaria en Baja California*. Tesis para optar por el grado de Maestro en Ciencias Educativas. México. Baja California. UABC. En: http://eduweb.ens.uabc.mx/egresados/Tesis/indicetesis.htm

Hatcher, H. (1998). *Using the SAS* System *for Factor Analysis and Structural Equations Modeling*. The SAS Institute. Cary, North Carolina.

Kozlow, Michael & Bellamy, Peter (2005). *Research on the 6+1 Trait Writing Model for Improving Student Writing*. Consultado el 22 de febrero de 2007 en: http://www.nwrel.org/ascd05/

Linacre, J.M. (1994). *Many-facet Rasch Measurement*. MESA, Chicago, Ill.

Stemler, S. E. (2004). *A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability*. *Practical Assessment, Research & Evaluation*, 9(4).

Wolfe, B., Dalton, M. and Neuburger, W. (1993). *Oregon Statewide Writing Assessment 1991 and 1992*. ERIC document: ED 366 960.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design.* Chicago, IL: MESA Press