Actas do X Congresso Internacional Galego-Português de Psicopedagogia. Braga: Universidade do Minho, 2009
ISBN- 978-972-8746-71-1

COMPARANDO DOIS MÉTODOS: ANGOFF MODIFICADO E GRUPOS DISTINTOS

Maria Cecília Silva
UIED- FCT, Universidade Nova de Lisboa
ceciliasilva@netcabo.pt

Resumo

A avaliação permite averiguar a existência de eventuais diferenças entre examinandos e a sua interpretação pode promover melhorias no processo de ensino/aprendizagem. Para analisar as diferenças entre as cotações médias de cada item num exame, bem como o desempenho global dos alunos, aplicaram-se dois métodos psicométricos: o método dos Grupos Distintos e o método de Angoff Modificado. O exame seleccionado foi o exame nacional de Física (1ª fase - 1ª chamada) realizado no ano lectivo 2002/2003. Numa primeira fase, detectaram-se as diferenças de desempenho entre examinandos de 18 escolas da área da Grande Lisboa. Posteriormente, efectou-se a comparação das cotações dos itens obtidas por esses examinandos com as cotações estimadas por um painel de 25 professores, após reapreciação. As conclusões apontam para a existência de grupos distintos de examinandos e para uma sobreavaliação do desempenho dos alunos, por parte dos professores, em ambos os métodos. Este estudo insere-se numa investigação alargada sobre ao grau de dificuldade dos exames nacionais de Física e Química.

1. Introdução

É um desafio comparar, de forma empírica, dois métodos psicométricos conotados com diferentes objectos de análise: os examinandos e os itens do exame. A escolha destes dois métodos assenta no seu uso generalizado e na possibilidade de obtenção de estimativas, quer ao nível dos itens, quer ao nível dos examinandos. Os resultados obtidos pelos examinandos são o factor de conexão entre os dois métodos (Giraud, Impara e Plake, 2005, p. 223).

A aplicação do método dos Grupos Distintos (*Contrasting Groups Method*), centrado nos examinandos (Kane, 1995, p. 119), permitiu maximizar a discriminação entre dois grupos de alunos internos, evidenciar desvios entre a Classificação Interna Final (C.I.F.) e a Classificação de Exame (C.E.) e, simultaneamente, comparar as respostas aos itens por parte dos dois grupos como sugerem Nijlen e Janssen (2008, p. 46).

Os itens são o foco do método de Angoff Modificado. Na implementação mais simples deste método, um painel de professores estimou a probabilidade de um determinado grupo de alunos responderem correctamente a cada item do exame. A média das estimativas dos professores permitiu:

a) estimar uma classificação para diferenciar o desempenho de dois grupos de alunos;

ISBN- 978-972-8746-71-1

b) comparar a cotação média estimada para cada item, com as cotações médias obtidas por um grupo de examinandos.

O exame seleccionado (Física da 1.ª Fase, 1.ª chamada, de 2003[1]) inclui um conjunto de itens, com tópicos do 10.º, 11.º e 12.º anos de escolaridade, enquadrados pelos objectivos mínimos da disciplina e, nos quais se exigiam definições, aplicações, deduções e cálculos. A escolha deste exame com seis itens de escolha múltipla e dezoito de resposta construída aumentou a complexidade da comparação e impôs constrangimentos à metodologia. Atendendo ao formato dos itens, a comparação combinou duas adaptações:

a) no Método de Angoff Modificado – a variação Angoff Verdadeiro/Falso, sugerida por Impara e Plake (1998, p. 69) para itens de escolha múltipla e a extensão do Método de Angoff, proposta por Hambleton e Plake (1995, p. 41) para os restantes itens;
b) no Método dos Grupos Distintos – uma variação assente nos valores das médias das classificações dos itens (Irwin, Bunckendahl e Poggio, 2007) e uma adaptação do modelo de regressão linear indicado por Cizek e Bunch (2007, p. 109).

O estudo incidiu sobre uma amostra de 275 alunos internos, pertencentes a 18 escolas do distrito de Lisboa.

2. Objectivos

O propósito subjacente a esta comparação é analisar o nível de desempenho dos examinandos internos de um determinado conjunto de escolas, num exame nacional. A análise procura respostas para as seguintes questões:

- Os resultados obtidos nos dois métodos são semelhantes?
- Podem ser detectadas diferenças no desempenho global destes alunos internos?
- Existem diferenças no desempenho item a item que justifiquem uma avaliação diferenciada?

São questões pertinentes porque, todos os anos, sem excepção, debate-se o grau de dificuldade dos exames nacionais em consonância com as expectativas relativas ao ensino aprendizagem e com o desempenho dos examinandos.

3. Métodos

A comparação de dois métodos com operacionalizações diferentes só é viável mantendo os intervenientes e aplicando procedimentos matemáticos idênticos. Para além dos 275 alunos internos já referidos, intervieram neste estudo um conjunto de 25 professores.

ISBN- 978-972-8746-71-1

Os alunos seleccionados proposeram-se a exame como alunos internos e foram considerados possuidores de requisitos básicos para efectuarem este exame e terem uma avaliação positiva. Nesta análise os alunos foram divididos em dois grupos de acordo com a sua classificação interna final (admitindo que os critérios de avaliação utilizados pelos professores na atribuição da C.I.F. foram semelhantes):

Grupo A – constituído por 155 alunos com uma classificação interna final entre 10 e 13 valores;

Grupo B – englobando os restantes 120 alunos internos.

A escolha deste intervalo de C.I.F. justifica-se por duas razões: pelos dados do relatório final do Júri Nacional de Exames, os examinandos internos da área seleccionada tiveram uma média de 13,2 pontos (J.N.E., 2003, p. 61) situando desta forma estes alunos abaixo da média observada e pela proposta de Livingston e Zieky (1982, p. 26) que, sugere percentagens de examinandos semelhantes. Se, no Grupo A, considerássemos apenas examinandos com CIF entre 10 e 12 valores, a distribuição ficaria aquém dos 50%, visto abranger apenas 42% dos alunos internos. A média das classificações de exame dos examinandos internos do distrito de Lisboa igualou os 69.6 pontos (J.N.E., 2003, p. 62), cerca de 53% da média das classificacões internas finais.

O painel de professores foi cuidadosamente seleccionado e incluíu autores de exames, consultores, auditores e professores com larga experiência, em consonância com a posição de Popham (2001, p. 298). Acrescenta-se o facto, do seu número exceder o número mínimo de doze avaliadores considerados necessários para obter níveis aceitáveis de confiabilidade. O critério de selecção atendeu à necessidade de realizar tarefas cognitivas extremamente complexas, nomeadamente: conceptualizar um determinado nível de desempenho, identificar o aluno enquadrado nesse nível, "colocar-se na pele do aluno, nas circunstâncias do exame e, estimar o desempenho desse aluno em itens com diferentes formatos" (Giraud, Impara, e Plake, 2005, p. 310). Os conhecimentos e experiência dos professores foram considerados suficientes para obter uma estimativa credível no método de Angoff Modificado.

A abordagem matemática envolveu em ambos os métodos:

a) o cálculo das médias das classificações dos itens e dos examinandos, com o objectivo de estimar a classificação que distingue os dois grupos e o comportamento dos examinandos face aos itens;
b) o modelo da regressão linear, aplicado aos examinandos (método dos Grupos Distintos) e aos itens (método de Angoff Modificado);

É uma abordagem inovadora porque não incide apenas na resposta a cada item, sugerida por Brandon (2002, p. 168), mas igualmente sobre a classificação global de exame.

ISBN- 978-972-8746-71-1

3.1. Método dos Grupos Distintos

O método dos Grupos Distintos, descrito inicialmente por Berk (1976, p. 4), utiliza a classificação de exame para obter uma classificação (designada na literatura inglesa por *cut score*) que maximiza a distinção entre o Grupo A e o Grupo B.

Existem vários procedimentos para operacionalizar este método. Neste estudo, o primeiro procedimento utilizado foi representar graficamente as classificações de exame obtidas pelos elementos dos grupos, para aferir se o seu comportamento era distinto ou não. Em segundo lugar, calcularam-se os valores das cotações médias dos itens e das classificações de exame obtidas pelos examinandos, para detectar eventuais diferenças na resposta a cada item, entre os elementos dos dois grupos. Seguidamente, analisou-se o comportamento dos alunos internos neste exame, aplicando o modelo da regressão logística binomial (Livingston and Zieky, 1989, p. 121), no qual, a variável de resposta é dicotómica (traduzindo a relação de pertencer ou não a um grupo), com o objectivo de estimar o *cut score* que distinguia os dois grupos.

A regressão logística, por defeito, utiliza a mais baixa das duas distribuições (designada por 0 – pertença ao Grupo A) como distribuição de referência para estimar a mais elevada (designada por 1 – pertença ao Grupo B ou ao grupo de professores avaliadores). Para ambas as regressões, as classificações entraram num único passo, não existindo por isso variação entre o passo, bloco e o modelo.

À semelhança de outro estudo (Silva, 2009, p. 7), na determinação do *cut score* foi aplicada a equação:

$$y = a + b(x),$$

onde y é o valor provável da variável que define a pertença do examinando a um grupo, a é a constante, b o declive da função de regressão e x o valor observado da classificação do examinando.

No contexto típico de uma regressão, o objectivo é determinar o valor de y, associado a um valor conhecido de x, por substituição na equação (Cizek e Bunch, 2007). Neste caso, o que se pretende é conhecer os valores de x, associados a resultados que se situam entre as distribuições do Grupo A e do Grupo B e, as distribuições do Grupo A e do Grupo dos Avaliadores. Como as duas distribuições estão codificadas em 0 e 1, respectivamente, utilizamos o valor $y = 0.56$ na regressões lineares dos Grupos A e B de examinandos internos e, dos Grupos A e Avaliadores. A escolha deste valor implicou considerar as percentagens relativas de pertença a um grupo.

Por último, efectou-se um tratamento das cotações obtidas pelos examinandos em cada item para permitir a sua posterior comparação com as estimativas dos avaliadores no Método de Angoff. Nos seis itens de escolha múltipla a classificação considerada foi 0 no caso de resposta

ISBN- 978-972-8746-71-1

errada e 1, ao invés de 10 pontos, em caso de acerto na opção correcta. As classificações dos restantes dezoito itens de resposta construída foram transpostas para uma escala de 1 a 4. Este tratamento deu origem a uma escala adaptada com valores entre 18 e 96 pontos.

Na conversão das cotações dos items, da escala inicial de 0 a 200 pontos (*Ei*), para a escala adaptada (*Ea*) de 18 a 96 pontos, utilizou-se na seguinte equação:

$$\frac{Ei}{200} = \frac{Ea - 18}{78}.$$

3.2. Método de Angoff Modificado

O método de Angoff é frequentemente utilizado em avaliações ao nível do 12.º ano (Mills e Melican, 1988, p. 264) porque abrange avaliações complexas que envolvem itens com formatos mistos. No âmbito da primeira versão do método proposto por Angoff em 1971, 25 professores avaliadores estimaram a resposta correcta para cada item, dos examinandos do Grupo A[2].

A fim de reduzir a dificuldade da estimativa, aplicou-se a variação Verdadeiro-Falso do método de Angoff aos seis itens de escolha múltipla que constituiam o Grupo I. Esses itens tinham uma pontuação dicotómica (0 ou 10 pontos) e, os professores avaliadores na sua estimativa assinalaram 1, no caso da alternativa correcta e, 0 caso correspondesse a uma opção incorrecta. Nos restantes dezoito itens de resposta construída, pertencentes aos Grupos II e III, considerou-se uma extensão do método de Angoff. O procedimento consistiu em estimar numa escala de 1-4, a classificação provável dos examinandos do grupo A, para permitir o seu tratamento e posterior comparação com os resultados dos examinandos.

O procedimento matemático utilizado para os valores estimados pelos professores avaliadores foi idêntico ao do método dos Grupos Distintos, quer no cálculo das médias e no modelo de regressão logística binomial, quer no uso do mesmo software.

Antes da apresentação dos resultados existem aspectos importantes a salientar relativamente ao desempenho dos alunos e, à selecção e tratamento das classificações de exame:

a) Admitiu-se, implicitamente, que o desempenho dos examinandos estava em consonância com o seu desempenho habitual, pois se tal não acontecesse, o grau de precisão do *cut score* seria inferior (Zieky et al., 2008, p. 130). Para além disso, a comparação entre os examinandos dos grupos só tem sentido se "os examinandos tiverem um desempenho similar em termos de conhecimento e competências no exame" (Livingston, 2006 p. 436);
b) Para evitar perdas de informação e aumentar a precisão da medida, utilizaram-se os valores das classificações de exame individuais atribuidas pelos professores no final do

ISBN- 978-972-8746-71-1

ano lectivo e não dos intervalos considerados na representação gráfica, no cálculo das médias dos dois grupos e na regressão linear relativa aos Grupos A e B.

Em relação aos itens, cada professor avaliador efectuou uma avaliação "cega", isto é, estimou as cotações dos itens para um aluno "minimamente competente" (Angoff, 1971, p. 515), desconhecendo o nível de desempenho específico de cada um dos 155 examinandos do Grupo A.

4. Resultados

Na operacionalização do método dos Grupos Distintos, as classificações de exame foram distribuídas por 10 intervalos relacionados com uma classificação de referência (Figura 1).

Class. de referência	Intervalo da CE	Grupo A	Grupo B
2	[0,24]	8	0
4	[25,44]	22	2
6	[45,64]	36	3
8	[65,84]	35	7
10	[85,104]	32	14
12	[105,124]	14	31
14	[125,144]	8	26
16	[145,164]	0	17
18	[165,184]	0	12
20	[185,200]	0	8

Figura 1. Tabela de frequências das classificações de exame dos 2 grupos.

As frequências das classificações de exame obtidas pelos dois grupos de alunos apontam para um nível de desempenho do grupo A inferior a 100 pontos, o mesmo é dizer, um nível de desempenho negativo.

Os valores das frequências permitiram a construção do Gráfico 1 no qual, o eixo horizontal fornece informação sobre as classificações de exame, distribuídas segundo as classificações de referência e, no eixo vertical, os valores indicam a proporção de cada grupo pertencente ao intervalo representado por cada uma das classificações de referência. Os cálculos foram efectuados admitindo que todos os valores de uma classe se confundem com o seu ponto médio.

ISBN- 978-972-8746-71-1

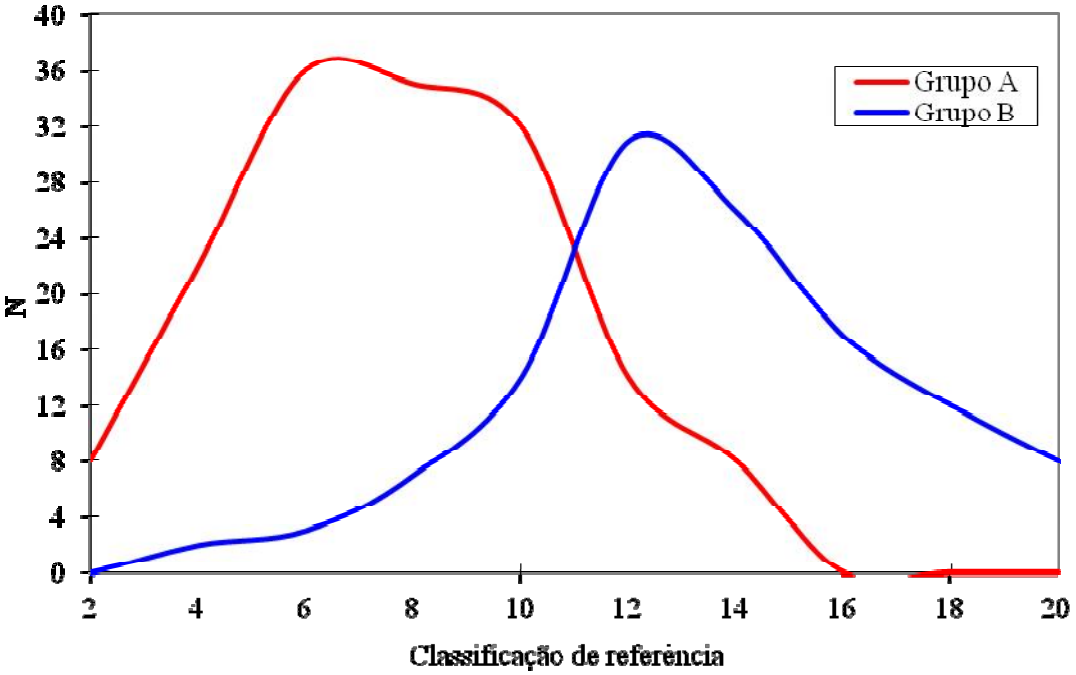

Gráfico 1. Grupo A versus Grupos B

A representação gráfica das duas distribuições de alunos internos confirmou a existência de dois grupos distintos de alunos internos e permitiu obter graficamente um *cut score* - aproximadamente 11 valores - abaixo dos 13 valores esperados.

Os valores das cotações médias dos itens e das médias das classificações de exame (C.E.) obtidas pelos examinandos dos dois grupos A e B (G. A e G. B) estão representados na figura 2.

	Grupo I						Grupo II												Grupo III						
	1	2	3	4	5	6	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	3.1	3.2	3.3	3.4	1.1	1.2	2.1	2.2	3	4	C.E.
G. A	3	4	4	4	2	3	4	5	1	5	1	4	4	5	6	3	3	2	1	2	2	2	1	1	71
G. B	6	5	7	7	5	6	5	7	4	7	3	6	8	8	10	6	4	5	2	3	2	4	3	3	127

Figura 2. Tabela das cotações médias por item (grupos A e B), na escala de 0 a 200 pontos.

Em todos os itens desta amostra, o nível de desempenho médio do Grupo A é inferior ao do Grupo B. Para determinar o *cut score*, numa versão simplificada do método dos Grupos Distintos (Cizek e Bunch, 2007, p. 109) calculou-se o valor médio ponderado das médias das classificações de exame dos dois grupos (99 pontos), o qual correspondeu a 10 valores numa escala de 0 a 20 valores. O valor é inferior aos 11 valores estimados a partir do gráfico.

Efectuando o cálculo do *cut score* com os valores transpostos para a escala de 18 a 96 pontos, os valores das médias das classificações dos Grupos A e B eram, respectivamente, 45 e 66 pontos, resultando um *cut score* de 56 pontos, que correspondia a 100 pontos na escala de 0 a 200 pontos. Verifica-se desta forma a concordância de valores nas duas escalas.

ISBN- 978-972-8746-71-1

Na figura 3 representam-se as cotações obtidas pelos examinandos do Grupo A, transpostas para a escala de 0-1 (itens 1 a 6 do Grupo I) e de 1-4 (itens 1.1 a 3.4 do Grupos II e itens 1.1 a 4 do Grupo III) e as estimativas dos professores avaliadores (A.V.).

	Grupo I						Grupo II												Grupo III					
	1	2	3	4	5	6	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	3.1	3.2	3.3	3.4	1.1	1.2	2.1	2.2	3	4
G.A	0,3	0.4	0.4	0.4	0.2	0.3	3.0	2.8	1.6	2.2	1.4	2.3	1.9	2.2	2.4	1.9	2.6	1.9	1.7	2.0	2.0	2.1	2.0	1.3
Av.	0.5	0.5	0.7	0.9	0.5	0.5	2.6	3.2	1.8	2.4	2.2	2.8	2.4	2.2	3.4	2.8	3.0	2.4	2.3	2.4	2.3	2.5	2.4	2.6

Figura 3. Tabela das cotações médias por item (Grupo A e Grupo dos Avaliadores), na escala de 18 a 96.

Neste procedimento consideraram-se as cotações médias por item dos avaliadores e verificou-se que apenas no item 1.3 as expectativas de obtenção de uma resposta correcta eram inferiores a 50%. Por outro lado, as estimativas dos professores foram superiores aos valores obtidos pelos examinandos, com uma excepção: o itens 1.1 do Grupo II. Uma possível explicação para o facto da média obtida pelos examinandos no item 1.1 ser superior à esperada pelos avaliadores prende-se com a aplicação rotineira das 2 equações paramétricas da cinemática necessárias para a resolução do item. A cotação média dos examinandos do Grupo A no item 2.3 foi a única concordante com as expectativas dos avaliadores.

Considerando as classificações médias do Grupo A (39 pontos) e do Grupo de professores avaliadores (48 pontos), o seu valor médio ponderado é igual a 44 pontos, ou a 68 pontos na escala de 0 a 200 pontos. Este valor é inferior à classificação média de exame do Grupo A (71 pontos), apesar das expectativas dos avaliadores.

Recorrendo ao software SPSS, determinaram-se os valores da constante *a* e do declive *b* da função de regressão linear que permitiram o cálculo dos *cut scores*:

a) da amostra total, considerando as classificações de exame de todos dos elementos da amostra (Grupo A + Grupo B) na escala de 0 a 200 pontos;
b) do conjunto (Grupo A + Avaliadores) na escala transformada de 18 a 96 pontos.

O súmula dos resultados obtidos encontra-se na figura 4.

	constante (a)	declive (b)	*cut score*
Grupo A+ Grupo B	- 5.466	0.053	113 pontos
Grupo A + Avaliadores	-5.552	0.091	70 pontos

Figura 4. Resultados obtidos para a regressão logistica binomial incidindo sobre as classificações de exame dos examinandos internos e do conjunto Grupo A + professores avaliadores

ISBN- 978-972-8746-71-1

O valor do *cut score* do conjunto Grupo A+ Avaliadores, 70, equivale a 137 pontos, na escala de 0 a 200 pontos e, era um valor esperado face às elevadas cotações médias do itens estimadas pelo grupo de avaliadores.

Na análise de resposta ao item aplicou-se o mesmo software, quer aos examinandos dos Grupos A e B, quer aos examinandos do Grupo A e professores avaliadores.

Na figura 5 apresentam-se os resultados desta análise de resposta ao item.

	constante (a)	declive (b)	*cut score*
Grupo A+ Grupo B	- 6.978	0.067	62 pontos
Grupo A + Avaliadores	- 6.241	0.098	69 pontos

Figura 5. Resultados da análise de resposta ao item para aos examinandos internos e para o conjunto Grupo A + professores avaliadores

Os valores dos *cut score* do conjuntos: Grupo A+ Grupo B, 62, equivale a 112 pontos e no Grupo A+ Avaliadores, 69, equivale a 131 pontos, na escala de 0 a 200 pontos.

Comparando os resultados obtidos nas duas regressões lineares não se observam grandes discrepâncias de valores. Uma vantagem do Método de Angoff Modificado relativamente ao Método dos Grupos Distintos é avaliar, item a item, o desempenho dos grupos de examinandos. Esta análise abrange uma amostra restrita de alunos da área da Grande Lisboa. Não é de excluir a obtenção de resultados diferentes, por exemplo, em áreas não citadinas.

5. Conclusões

Podem salientar-se os seguintes aspectos:

A - Uma semelhança entre os resultados obtidos através destes dois métodos;

B - As expectativas dos professores são, em geral, muito elevadas face aos resultados desta amostra, na qual, 56% dos examinandos internos pertenciam ao Grupo A – alunos com C.I.F. entre 10 e 13 valores;

C - Existem diferenças significativas nas cotações médias dos items de resposta construída. A menor cotação dos items relacionados com a componente experimental do currículo é um indicador da necessidade de promover o desenvolvimento de mais competências neste domínio;

D - Não se justifica uma avaliação diferenciada porque os examinandos de ambos os grupos revelam as mesmas dificuldades , sendo o desempenho dos examinandos do Grupo A, como seria de esperar, inferior ao do Grupo B.

O fundamento deste tipo de estudos assenta na concepção que a razão primordial da avaliação da qualidade do ensino é a melhoria das aprendizagens.

ISBN- 978-972-8746-71-1

Os resultados deste estudo sugerem uma maior especialização dos professores avaliadores para promover a obtenção de *cut scores* mais próximos dos valores dos examinandos, bem como a aplicação de outros métodos mistos, considerando diferentes amostras.

6. Notas

[1] O enunciado deste exame, bem como os critérios de correcção estão disponiveis no site do Ministério da Educação, em http://www.gave.min-edu.pt/np3/76.html.

[2] Existe uma versão posterior deste método, sugerida por Angoff. Nesta segunda versão, aplicada nomeadamente por Nijlen e Janssen (2008), os professores avaliadores têm de estimar uma probabilidade para a resolução correcta de cada item, por parte dos examinandos, designados como "minimamente competentes". Para não sobrecarregar mais os professores avaliadores, considerou-se a versão mais simples do método.

7. Referências bibliográficas

Berk R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, 45, 4-9.

Brandon, P. R. (2002). Two versions of the Contrasting-Groups Standard-Setting Method: a Review. *Measurement and Evaluation in Counseling and Development*, 35, 167-181.

Camilli, G. (2006). Test Fairness. In Brenan, R. L. (Ed.), *Educational Measurement* (4th ed. pp. 221-256). Westport, CT: ACE/Praeger.

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting*. Thousand Oaks, London and New Delhi: Sage Publications, p. 107.

Cizek, G. J. & Husband, T. H. (1997). *A Monte-Carlo investigations of the contrasting groups standard-setting method*. Paper presented at the meeting of the American Educational Research Association, Chicago (p. 18).

Cronbach, L. J. (1971). Test validation. In R. L. Torndick (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41-56.

ISBN- 978-972-8746-71-1

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35(1), 69-81.

Irwin, P., Buckendahl, C. W., & Poggio, A. (2007). *Examinee-Centered Standard Setting: An Alternative Approach.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Kane, M. (1995). Examinee-centered vs. task-centered standard setting. In *Proceedings of the Joint Conference on Standard Setting in Large-Scale Assessments* (Vol. 2, pp. 119-139).Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

Livingston, S. A. (2006). Item Analysis. In Downing, S. M. & Haladyna, T. M. (Ed) *Handbook of Test Development* (pp. 421-441). Mahwah, NJ: Erlbaum.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Livingston, S., & Zieky,M. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2,* 121–141.

Mills, C. N., & Melican, G. J. (1988). Comparative Review: Estimating and Adjusting Cutoff Scores: Features of Selected Methods. *Applied Measurement in Education, 1*(3), 264.

Nijlen, D. V. & Janssen, R. (2008). Modeling Judgments in the Angoff and Contrasting-Groups Method of Standard Setting. *Journal of Educational Measurement,* 45 (1), 45–63.

Popham, W. J. (1978). As always, provocative. *Journal of Educational Measurement*, 15, 297-230.

Zieky, M., Perie, M., & Livingston, S. (2008). *Cutscores: A manual for setting performance standards on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Webb, N. L. (2006). Identifying Content for Student Achievement Tests. In Downing, S. M. & Haladyna, T. M. (Ed.) *Handbook of Test Development* (pp. 155-180). Mahwah, NJ: Erlbaum.