DETERMINAÇÃO DA EXTENSÃO DE TESTES REFERENCIADOS A CRITÉRIOS

Robert K. Walker, Ed.D.*

As primeiras prescrições para a preparação de objetivos comportamentais (Mager, 1962; Popham e Baker, 1976) foram muitas vezes interpretadas como se cada objetivo fosse equacionado com uma única questão de prova. Briggs (1976) distinguiu entre a aprendizagem reprodutiva e a produtiva. Dos oito tipos de aprendizagem propostos por Gagné, os primeiro cinco (respostas clássicas condicionadas, condicionamento operante, cadeias motoras, associação verbal e discriminação múltipla) foram considerados de aprendizagem reprodutiva, e os últimos três (conceitos, utilização de princípios e resolução de problemas), de aprendizagem produtiva. Briggs sugeriu que os objetivos de aprendizagem reprodutiva fossem testados através de uma única questão, enquanto que os objetivos de aprendizagem produtiva normalmente exigiriam mais de uma questão. Gagné e Briggs (1974) ofereceram a opinião de que "é difícil ver por que mais do que três ou quatro exemplos seriam necessários", acrescentando que "o propósito da utilização de múltiplos exemplos é principalmente o de evitar 'erros de mensuração'"

Quando os editores de testes nos Estados Unidos começaram a publicar testes referenciados a critérios, era natural que seguissem o padrão prevalecente de um ou, às vezes, dois itens por objetivo. Millman (1974) chamou a atenção para esta tendência, observando:

> "Além de serem de uma precisão questionável, tais testes podem apenas fornecer informações sobre a habilidade do aluno para responder a um quesito particular e não para desempenhar um conjunto mais generalizado de tarefas. O conhecimento da habilidade de responder a uma questão específica tem um valor limitado, e quase sempre os editores de testes... acharam necessário tentar algum tipo de análise de perfil em que são empregadas respostas a questões adicionais. Em vez de apresentar dados de avaliação

* Professor visitante do Programa de Pós-Graduação em Educação na Universidade Federal do Espírito Santo.

de estudantes numa miscelânea de itens específicos, parece mais interessante que a coleção de itens seja cuidadosamente escolhida de um âmbito bem definido."

Hambleton *et alii* (1978) distinguem entre "estes referenciados a objetivos", por um lado, e "testes referenciados a critérios" ou "testes referenciados a âmbitos" (*domain-referenced tests*), por outro. Eles definem "testes referenciados a objetivos" como "testes constantes de itens igualados a objetivos". Salientam que "com um teste referenciado a objetivos nenhum âmbito de comportamentos é especificado, e os itens não são considerados representativos de nenhum âmbito de comportamento". Estes autores consideram "testes referenciados a critérios" como equivalentes a "testes referenciados a âmbitos", adotando a definição de Popham: "Um teste referenciado a critérios é usado para determinar o status de um indivíduo [designado como escore de âmbito] com respeito a um âmbito bem definido de comportamento (Popham, 1975)".

Geis (1978) propõe uma distinção semelhante entre "enunciado de desempenho total" e "enunciados de amostra". A sua definição de um enunciado de desempenho total, "uma descrição completa do desempenho que o aluno deve poder exibir ao final da instrução", pode, muitas vezes, implicar um tipo mais intensivo de comportamento que o exigido por provas de lápis e papel – chamados "testes referenciados a objetivos" (ou, erroneamente, "testes referenciados a critérios") –; mas o princípio é o mesmo: um "item" por objetivo. Em contrapartida, os "enunciados de amostra" exigem a seleção de itens dentro de um âmbito.

Um argumento contra os "testes referenciados a objetivos" poderia ser levantado do ponto de vista da teoria da aprendizagem. Um dos principais benefícios dos objetivos comportamentais é a maximização da aprendizagem intencional. Por outro lado, a aprendizagem incidental também é importante, tanto como habilidade e hábito para a educação permanente como para a aquisição de conhecimentos e habilidades durante uma dada experiência de aprendizagem. A nossa hipótese seria a de que objetivos demasiadamente específicos tendem a focalizar a atenção do aprendiz exclusivamente na preparação para as questões esperadas, inibindo, desse modo, a aprendizagem incidental(1). Mais amplos "enunciados de desempenho total", na medida em que forem menos triviais, poderão ser mais úteis. O terceiro tipo de objetivo proposto por Geis (1978), "enunciados de conseqüências", parece especialmente eficaz para objetivos que não permitem uma análise completa de tarefas, tais como "falar como um nativo". O avaliador teria apenas que constatar se os nativos conseguem identificar o aluno como estrangeiro.

Alguns pesquisadores e usuários empregam o termo "testes referenciados a âmbitos" somente para referir-se a amostras de itens de um universo de conteúdo explicitamente definido. O âmbito pode ser definido através de uma lista exaustiva ou, desde que isto é usualmente impossível, por algum sistema lógico. São exemplos a abordagem baseada em lingüística de Bormuth (1970); formas de item, introduzidas por Osborn (1968) e Hiveley *et alii* (1973); e a análise de facetas (Guttman, 1969).

Recentemente surgiram dúvidas sobre a viabilidade de desenvolver especificações de âmbitos fora da área da Matemática – onde a maior parte do trabalho importante foi feito até agora – e, mesmo, de estender as especificações de âmbitos aos objetivos mais complexos da Matemática (Hambleton, *et alii*, 1975). Hambleton e seus colegas recomendam a noção de Popham (1974) de um objetivo ampliado que "oferece um excelente balanço entre a clareza alcançada através dos sistemas de geração de itens e a praticalidade dos objetivos comportamentais". Segundo Millman (1974), "um objetivo ampliado é um enunciado expandido de um objetivo educacional que estabelece especificações de fronteiras referentes a situações de prova, respostas alternativas e critérios para correção". Um bom tratamento popular do tópico pode

(1) Seria interessante testar esta hipótese nos Centros de Ensino Supletivo, uma das instituições que mais aplicam o ensino individualizado no Brasil.

ser encontrado em Baker e Popham, 1976. Objetivos ampliados são a base do *"Instructional Objectives Exchange"*, um serviço de intercâmbio de objetivos que iniciou operações em Los Angeles em 1973 (Popham, 1974).

Este trabalho irá considerar a questão do número de itens e o escore de aprovação necessários para evidenciar domínio (mestria) de um objetivo ampliado ou referenciado a um âmbito, ao nível de um dado percentual de acertos. Não tratará diretamente de outro assunto relacionado a este, o dos intervalos de confiança em torno das estimativas do verdadeiro nível de funcionamento do examinando em algum âmbito bem especificado de conteúdo.

AS ABORDAGENS BINOMIAL E BAYESIANA

Quando são usados objetivos ampliados ou referenciados a âmbitos, presume-se que existe um universo de itens demasiado grande para que todas as questões possam ser incluídas numa prova. A questão está em quantos itens devem ser selecionados por amostragem e que escore de aprovação deve ser exigido para evidenciar domínio do objetivo ao nível especificado do critério. Como em qualquer tipo de estimação estatística, dois tipos de erro podem ocorrer: falsos positivos e falsos negativos. Um falso positivo significa que um aluno que na realidade não possui domínio (não alcançaria o critério num teste que incluísse todos os itens do universo) é incluído na categoria dos que possuem domínio. Um falso negativo ocorre quando um aluno cujo nível de funcionamento (o percentual de itens no universo que ele poderia acertar)(2) está ao nível ou acima do nível do critério é classificado como não possuindo domínio.

Millman (1972), aplicando um modelo binomial, construiu uma série de tabelas, mostrando o percentual de alunos que se espera aprovar ou reter erroneamente, dado seu nível de funcionamento, o nível especificado do critério, o número de itens e o escore de aprovação.

> "Por exemplo, suponhamos que o construtor do teste esteja disposto a permitir que, em 15 por cento dos casos, o aluno, cujo escore verdadeiro no âmbito da prova (nível de funcionamento) é 60 por cento, receba um escore tão alto como 70 por cento. Observando a coluna encabeçada "60", vemos que uma prova de 25 itens daria esta precisão. Isto é, existe uma probabilidade de apenas 15 por cento de que um aluno, cujo escore no âmbito da prova é 60 por cento, receba um escore de 70 por cento ou mais numa prova de 25 itens (Millman, 1974)."

Com o critério de domínio fixado em 70 por cento, e observando a restrição acima citada, exigir-se-ia que o aluno acertasse pelo menos 18 entre 25 itens.

Sem dúvida, a implicação de que seria recomendável incluir 25 itens para determinar o domínio de apenas um objetivo choca-se com a realidade de programas de instrução individualizada, especialmente aqueles com uma clientela pouco acostumada a fazer provas. Felizmente, outros modelos foram propostos, resultando em recomendações mais razoáveis.

Novick e Lewis (1974) mostraram que, aplicando as tabelas de Millman, a melhoria de precisão é pequena e flutuante quando se aumenta o número de itens de 8 a 22. Eles acusam Millman de propor a questão errada:

> "Em vez de a probabilidade de um aluno alcançar um escore na prova, dado seu nível verdadeiro, o que precisamos, para tomar uma decisão, é a probabilidade de que o nível verdadeiro de funcionamento de aluno supere o nível especificado do critério, dado seu escore na prova."

Para abordar este problema, Novick e Lewis propõem a aplicação da estatística bayesiana. O modelo proposto exige que o instrutor forneça três tipos de informação: o nível do critério, o conhecimento prévio do nível verdadeiro de funcionamento do aluno em termos probabilís-

---

(2) Chamado também "escore de âmbito" ou "verdadeiro escore de proporção de acertos".

ticos, e as perdas relativas associadas com os dois tipos de erros – falsos positivos e falsos negativos (a razão entre perdas)([3]).

Nas tabelas bayesianas apresentadas por Novick e Lewis (como também nas tabelas de Millman apresentadas no mesmo trabalho), não existe nenhuma relação direta entre o nível do critério e o número de itens exigido, embora a extensão da prova recomendada tenda a ser um pouco menor para um critério de 70 por cento do que para níveis mais altos. Nas tabelas bayesianas, tanto mais alta a razão entre perdas, tanto mais itens são exigidos. Por outro lado, tanto mais alto o nível esperado de funcionamento, tanto menos itens são exigidos. Por exemplo, com o nível esperado de funcionamento igual ao nível do critério, neste caso 80 por cento, a extensão recomendada da prova varia de 7 questões, com escore de aprovação 6 (6/7), para uma razão entre perdas de 1,5 a 1; a 19/22, para uma razão entre perdas de 3 a 1. Entretanto, com o nível esperado de funcionamento igual a 85 por cento e o nível do critério posto a 80 por cento, a extensão recomendada varia de 8/10 a 11/13. Assim, a convicção relativamente forte de que o nível de funcionamento do aluno está acima do nível do critério pode compensar as razões altas entre perdas, possibilitando extensões mais razoáveis de testes.

A estatística bayesiana oferece, sem dúvida, um potencial muito grande para a educação. Os tomadores de decisões instrucionais provavelmente levam implicitamente em consideração convicções anteriores e razões entre perdas, sendo desejável incluí-las explicitamente nos cálculos. Entretanto, a aplicação da estatística bayesiana irá requerer, para o futuro previsível, assessoramento especializado, escasso em relação às necessidades e praticamente inexistente no Brasil. Poderá ser difícil treinar os tomadores de decisões instrucionais a estimarem probabilidades *a priori* e razões entre perdas. Hambleton, *et alii* (1978) assinalam que

> "precisam ser consideradas as dúvidas sobre os ganhos gerais que poderiam resultar, tendo em vista a complexidade dos procedimentos, a robustez dos modelos bayesianos em situações de prova onde as presunções básicas do modelo não forem satisfeitas (por exemplo, com provas muito curtas), e a sensitividade dos modelos bayesianos à especificação de probabilidades *a priori*."

É por estas razões, como também pela compreensão adcional que poderá ser adquirida através da consideração de alternativas, que apresentamos outro modelo derivado do teorema binomial. Este modelo considera a "questão certa" – isto é, qual é a probabilidade de que o nível verdadeiro de funcionamento de um aluno exceda o nível especificado do critério, dado seu escore na prova. Conta, ainda, com a vantagem adicional de validade até para testes muito curtos.

UM MODELO BINOMIAL MODIFICADO

Wilcox (1976) desenvolveu os aspectos essenciais do modelo em questão, relacionando uma solução baseada na teoria clássica dos testes dada por Fhanér (1974) à abordagem de Millman. A inovação básica, que possibilita uma solução viável, é a noção de uma zona de indiferença. Wilcox chama o nível de funcionamento π e o nível do critério π_0. A zona de indiferença é definida por $\pi_0 \pm c$.([4])

([3]) Hambleton, *et alii* (1978) assinalam que estas razões são tipicamente superiores a um: "Por exemplo, com objetivos instrucionais que são pré-requisitos de objetivos mais avançados num currículo, erros do tipo falso positivo (fazendo avançar os examinandos antes de eles estarem prontos) podem ser muito mais graves do que erros do tipo falso negativo (retendo os examinandos, embora eles tenham "dominado" os objetivos em questão)".

([4]) Num artigo mais recente, Wilcox (1977) usa λ e λ_0 em lugar de π e π_0.

"A importância da zona de indiferença é que se um examinando tiver um nível de funcionamento "perto" de π_0, então existe uma perda desprezível na classificação errônea. Entretanto, se o nível de funcionamento do examinando for "substancialmente" maior que π_0 (maior ou igual a $\pi_0 + c$) ou substancialmente menor que π_0 (menor ou igual a $\pi_0 - c$), queremos então estar relativamente seguros de que uma decisão correta será tomada (Wilcox, 1976)."

Wilcox apresenta uma tabela de "Escores de Aprovação e a Probabilidade Mínima de uma Decisão Correta para Valores de π fora da Zona de Indiferença", para os níveis de critério 0,70; 0,75; 0,80; e 0,85; c = 0,05 e 0,10; e número de itens n = 8 . . . 20. Para encontrar a "probabilidade mínima de uma decisão correta", ele calculou α (alfa, a probabilidade de evitar um falso positivo) e β (beta, a probabilidade de evitar um falso negativo).(5) A menor das duas é a probabilidade mínima.(6)

Uma característica desejável da tabela de Wilcox é que a probabilidade mínima de uma decisão correta tende a aumentar com o número de itens n. As exceções são atribuídas à "descontinuidade da função de densidade binomial". Como veremos mais adiante, esta anomalia fica esclarecida quando α e β são apresentadas separadamente, e o número de erros permitido é mantido constante dentro de categorias separadas. Também é interessante notar que, pelo menos nos níveis mais altos de n (onde as descontinuidades foram até certo ponto niveladas), a probabilidade mínima de uma decisão correta é diretamente proporcional ao critério π_0.

Para usar a tabela, fixa-se o critério π_0 e a zona de indiferença $\pm$ c, e desce-se a coluna até encontrar a probabilidade mínima de uma decisão correta requerida. Por exemplo, dado $\pi_0 = 0{,}80$ e c = 0,1, e exigindo-se um grau de certeza de uma decisão correta de pelo menos 75 por cento, precisa-se então de uma prova de nove itens e escore de aprovação oito (8/9), que dá a probabilidade mínima de uma decisão correta igual a 0,7748.

A interpretação das nossas Tabelas I e II é mais clara. Consideremos primeiro a categoria 2 da Tabela I. Nesta categoria, um examinando pode errar no máximo um item e ainda ser considerado possuidor de domínio ($n_0 = n - 1$, onde n_0 é o escore de aprovação). Como veremos, α é diretamente proporcional a n, enquanto β é inversamente proporcional a n. A probabilidade mínima de uma decisão correta é sempre a menor das duas probabilidades associadas com um dado n. Os dois valores mais altos da probabilidade mínima são aqueles que seriam encontrados exatamente de ambos os lados da interseção das curvas representando α e β. Todas as outras coisas sendo iguais, para um dado conjunto de parâmetros (π_0, c e o número de erros permitidos), o número ótimo de itens para um teste corresponderá a uma ou outra destas probabilidades. Neste caso, um teste de 9 itens seria ótimo. Note que chegamos à mesma conclusão acima, através da interpretação da tabela de Wilcox.

Na Tabela I, os níveis de α e β são apresentados para as categorias $n_0 = n$, $n_0 = n - 1$, $n_0 = n - 2$, e $n_0 = n - 3$. Note-se que, quando o número de erros permitido (e a extensão média de prova dentro de cada categoria) aumenta, também aumentam os valores médios tanto de α como de β. Note-se também que, exceto onde $n_0 = n$, a razão de aprovação (n_0/n) aumenta em proporção a n, dentro de cada categoria.

---

(5) No artigo mais recente, Wilcox (1977) define α e β como a probabilidade de *cometer* um erro tipo falso positivo ou falso negativo, respectivamente. Por motivo de coerência com o artigo em que está baseada esta apresentação, conservamos a terminologia anterior.

(6)
$$\alpha = \sum_{x=0}^{n_0-1} \binom{n}{x} (\pi_0 - c)^x (1 - \pi_0 + c)^{n-x}$$

$$\beta = \sum_{x=n_0}^{n} \binom{n}{x} (\pi_0 + c)^x (1 - \pi_0 - c)^{n-x}$$

Aqui, n é o número de itens, n_0 é o escore de aprovação, e x é o número de acertos. $\binom{n}{x}$ é o número de combinações de n itens x a x. O leitor perceberá aqui uma aplicação da expansão binomial.

TABELA I

Extensão de Teste, Escore de Aprovação, e Probabilidades de Evitar um Falso Positivo (α) e um Falso Negativo (β) para Critério $\pi_0 = 0,8$ e Zona de Indiferença $\pm c = 0,1$

	n	n_0	Razão de Aprovação	α	β
Categoria 1	1	1	1,00	0,3000	0,9000
($n_0 = n$)	2	2	1,00	0,5100	0,8100
	3	3	1,00	0,6570	0,7290
	4	4	1,00	0,7599	0,6561
	5	5	1,00	0,8319	0,5905
	6	6	1,00	0,8824	0,5314
	7	7	1,00	0,9176	0,4783
Categoria 2	5	4	0,80	0,4718	0,9185
($n_0 = n-1$)	6	5	0,83	0,5798	0,8858
	7	6	0,86	0,6706	0,8503
	8	7	0,88	0,7447	0,8131
	9	8	0,89	0,8040	0,7748
	10	9	0,90	0,8507	0,7361
	11	10	0,91	0,8870	0,6974
	12	11	0,92	0,9150	0,6590
Categoria 3	10	8	0,80	0,6172	0,9298
($n_0 = n-2$)	11	9	0,81	0,6873	0,9104
	12	10	0,83	0,7472	0,8991
	13	11	0,85	0,7975	0,8651
	14	12	0,86	0,8392	0,8415
	15	13	0,87	0,8732	0,8159
	16	14	0,88	0,9006	0,7892
	17	15	0,88	0,9226	0,7618
Categoria 4	15	12	0,80	0,7031	0,9444
($n_0 = n-3$)	16	13	0,81	0,7541	0,9316
	17	14	0,82	0,7981	0,9174
	18	15	0,83	0,8354	0,9018
	19	16	0,84	0,8668	0,8850
	20	17	0,85	0,8929	0,8670
	21	18	0,86	0,9144	0,8480
	22	19	0,86	0,9319	0,8281

Para um teste de um único item, $\alpha = 0,3000$. Se o professor fosse jogar uma moeda para determinar domínio ou não-domínio, a probabilidade de evitar um falso positivo (ou um falso negativo) seria 0,5. Desde que evitar um falso positivo é normalmente mais importante do que evitar um falso negativo na instrução seqüenciada, isto poderia ser preferível a uma prova de um item, para inferir domínio de um âmbito de itens. Da mesma maneira, a exigência de quatro acertos entre cinco itens, embora intuitivamente atraente como evidência de domínio ao nível de 80 por cento, dá um valor de α inferior ao acaso. Por outro lado, é animador que provas de apenas três ou quatro itens dão probabilidades mínimas respeitáveis. Embora possa haver alguma objeção ao uso de uma razão de aprovação de 100 por cento para evidenciar domínio ao nível de 80 por cento, as economias consideráveis propiciadas por provas curtas poderão justificar seu uso quando probabilidades mínimas mais altas não forem requeridas.

Wilcox (1976) assinala que dois valores de c podem ser especificados: um para o intervalo acima de π_0, e outro para o intervalo abaixo. Desta maneira, podemos efetuar essencialmente a mesma coisa como a razão entre perdas na estatística bayesiana – levar em consideração que erros do tipo falso positivo são normalmente mais graves do que falsos negativos. Vamos chamar o intervalo da zona de indiferença abaixo de π_0 c_1, e o intervalo acima de π_0, c_2. Podemos fixar c_1 a, digamos, 0,05, e c_2 a 0,15. Com $\pi_0 = 0{,}80$, a zona de indiferença se estenderia de 0,75 até 0,95. Em outra palavras, estamos dizendo que classificar um examinando cujo nível de funcionamento é menor que 0,75 como possuidor de domínio ao nível de 0,80 é tão grave como classificar um examinando cujo nível de funcionamento é maior que 0,95 como não possuidor de domínio. Visto que a razão de c_2 a c_1 é, neste caso, três a um, podemos dizer que isto corresponde aproximadamente a uma razão entre perdas de três a um na estatística bayesiana. A Tabela II apresenta os valores de α para $c_1 = 0{,}05$ e de β para $c_2 = 0{,}15$. Como vemos, o valor

TABELA II

Extensão de Teste, Escore de Aprovação, e Probabilidades de Evitar um Falso Positivo (α) (com $c_1 = 0{,}05$), e um Falso Negativo (β) (com $c_2 = 0{,}15$), para Critério $\pi_0 = 0{,}8$

	n	n_0	Razão de Aprovação	α	β
Categoria 1	1	1	1,00	0,2500	0,9500
($n_0 = n$)	2	2	1,00	0,4375	0,9025
	3	3	1,00	0,5781	0,8574
	4	4	1,00	0,6836	0,8145
	5	5	1,00	0,7629	0,7738
	6	6	1,00	0,8220	0,7351
	7	7	1,00	0,8665	0,6983
Categoria 2	5	4	0,80	0,3672	0,9774
($n_0 = n-1$)	6	5	0,83	0,4661	0,9672
	7	6	0,86	0,5551	0,9556
	8	7	0,88	0,6329	0,9428
	9	8	0,89	0,6997	0,9288
	10	9	0,90	0,7560	0,9139
	11	10	0,91	0,8029	0,8981
	12	11	0,92	0,8416	0,8816
Categoria 3	10	8	0,80	0,4744	0,9885
($n_0 = n-2$)	11	9	0,81	0,5448	0,9848
	12	10	0,83	0,6093	0,9804
	13	11	0,85	0,6674	0,9755
	14	12	0,86	0,7189	0,9699
	15	13	0,87	0,7639	0,9638
	16	14	0,88	0,8029	0,9571
	17	15	0,88	0,8363	0,9497
Categoria 4	15	12	0,80	0,5388	0,9945
($n_0 = n-3$)	16	13	0,81	0,5950	0,9930
	17	14	0,82	0,6470	0,9912
	18	15	0,83	0,6943	0,9891
	19	16	0,84	0,7369	0,9868
	20	17	0,85	0,7748	0,9841
	21	18	0,86	0,8083	0,9811
	22	19	0,86	0,8376	0,9778

mais alto da probabilidade mínima de uma decisão correta na primeira categoria (onde $n_0 = n$) ocorre agora com cinco itens, em vez de três, como na Tabela I. Nas outras categorias, o número ótimo de itens é maior do que o número máximo incluído na tabela.

Combinando as Tabelas I e II, podem ser criadas quatro combinações de c_1 e c_2, dando razões de 1:1 (onde $c_2 = 0,1$ e $c_1 = 0,1$), 3:2 (onde $c_2 = 0,15$ e $c_1 = 0,1$), 2:1 (onde $c_2 = 0,1$ e $c_1 = 0,05$), e 3:1 (onde $c_2 = 0,15$ e $c_1 = 0,5$). Se não houver interesse em evitar um falso negativo, c_2 pode ser igualado a $1 - \pi_0$, e apenas a coluna de α precisará ser considerada.

Uma recomendação geral seria a de o tomador de decisões instrucionais escolher a categoria mais alta para a qual haja suficiente tempo e recursos para testagem. As categorias mais altas (aquelas que permitem que o examinando erre mais itens) oferecem maior precisão, ao custo de um maior número de itens. Dentro da categoria escolhida, ele deve selecionar a extensão ótima de prova (aquela com maior probabilidade mínima de uma decisão correta). Normalmente as duas extensões, exatamente de ambos os lados da interseção das curvas de α e β, terão probabilidades mínimas aproximadamente iguais, e a menor das duas poderá ser selecionada.

PROBABILIDADES DE CLASSIFICAÇÃO ERRÔNEA ENVOLVIDAS EM ITENS INDIVIDUAIS

Wilcox (1976) assevera que "nenhuma presunção sobre a homogeneidade dos itens (em conteúdo ou dificuldade) é necessária". Entretanto, se consideramos cada item como representante de uma competência para o mundo real, fica claro que existe a possibilidade de erro. Por exemplo, adivinhar a resposta certa resulta num falso positivo. Deixar de resolver uma equação quadrática por causa de um erro nas somas resulta num falso negativo (na realidade, o examinando possui a habilidade de resolver equações quadráticas). Podemos dizer que cada item contém dentro de si um nível alfa e um nível beta. Embora nenhuma solução matemática tenha sido proposta ainda para permitir o ajustamento preciso dos níveis gerais de alfa e beta na base das probabilidades dos itens, o tomador de decisões instrucionais pode fazer ajustamentos aproximados na base de estimativas dos níveis de alfa e beta dos itens.

Com referência a falsos positivos, Davis e Diamond (1974) assinalam:

> "Com itens de escolha múltipla, a probabilidade de marcar corretamente qualquer item, advinhando-o, é $1/c$, onde c é o número de alternativas. Parece provável que os examinandos raramente adivinhem entre todas as escolhas possíveis dentro de um item. Em vez disso, eles são capazes de eliminar uma ou mais alternativas como incorretas, para que possam adivinhar entre as alternativas que sobrarem. A probabilidade de acertar um item nestas circunstâncias é $1/(c - x)$, onde x é o número de alternativas que podem ser corretamente identificadas como sendo incorretas."

Mesmo com itens de escolha livre, observam Davis e Diamond, os examinandos usualmente precisam selecionar uma resposta apenas dentre algumas poucas que são relembradas por causa do estímulo. Não obstante, a seguinte observação de Skager (1974) parece pertinente:

> "Se o teste incorpora itens tipo produção, onde adivinhar é uma base improvável ou até impossível para uma resposta correta, são requeridos menos itens... uma análise válida de virtualmente qualquer principal âmbito de conteúdo escolar produzirá muitos objetivos que requerem que o aprendiz gere uma resposta em vez de selecionar uma dentre um conjunto de alternativas."

Presumimos que as probabilidades dadas por Wilcox sejam válidas para itens simples de produção (estes não devem ser confundidos com itens que medem a aprendizagem produtiva), nos quais a probabilidade de adivinhar a resposta correta seja virtualmente zero. Se houver uma probabilidade substancial de advinhar a resposta correta, deve-se descer na tabela, selecionando uma extensão um pouco maior que a "ótima", dentro da categoria escolhida. Se a probabilidade de evitar falsos positivos for 50 por cento ou menos, como no caso de itens tipo falso-verda-

deiro, deve-se descer ainda mais, com categorias mais altas (envolvendo testes mais longos e sendo permitido maior número de erros) sendo recomendadas.

Itens de produção complexa envolvem numerosas tarefas que poderiam ser consideradas itens em si, como também a necessidade de integração de tarefas. Tanto maior o número de subitens envolvidos, tanto menor a probabilidade de que um não-possuidor de domínio possa efetuar todos por acaso, sendo assim incorretamente classificado como possuidor de domínio do item. Desta maneira, o nível de alfa de cada item complexo é relativamente alto.

Assim como baixos níveis de alfa dos itens reduzem os verdadeiros níveis de alfa do teste, exigindo, portanto, provas mais longas, altos níveis de alfa dos itens parcialmente compensam baixos níveis de alfa do teste, permitindo provas mais curtas. Isto é conveniente, desde que itens de produção complexa freqüentemente consomem bastante tempo, e o número que pode razoavelmente ser incluído numa prova é limitado.

CONCLUSÕES

Os objetivos instrucionais de um só item deveriam ser banidos? Na maioria dos casos, exceto talvez a nível primário, eles deveriam ser evitados, porque tendem a ser triviais em si, e porque não proporcionam nenhuma evidência convincente de que o examinando que acerta o item seja realmente possuidor de domínio de qualquer âmbito amplo de itens. Entretanto, nos dois extremos do *continuum* da testagem, eles podem ser justificáveis. O desempenho satisfatório de um item de produção altamente complexa poderá propiciar evidência adequada do domínio do objetivo relacionado (muito embora, para demonstrar consistência de desempenho, sejam necessários múltiplos itens ou a repetição do mesmo item).

No extremo oposto, é óbvio que a amostra não pode ser maior do que o universo. Se um objetivo for enunciado de tal maneira que apenas um item simples seja possível, então não haverá possibilidade de generalizar a algum âmbito maior de itens. O teste (ou subteste referenciado a aquele objetivo particular) só poderá ter um único item, e normalmente não teria sentido repeti-lo. Embora a generalizabilidade usualmente seja desejável, existem sem dúvida alguns itens que são tão importantes em si que devem ser especificados em forma de enunciados de desempenho total.

A utilização inteligente das Tabelas I e II deve possibilitar a determinação de extensões de testes e escores de aprovação para objetivos ampliados ou referenciados a âmbitos, com critério 80 por cento e várias zonas de indiferença. Para outros valores de π_0 e c, Wilcox sugere escrever um programa em FORTRAN incorporando bem difundidas sub-rotinas de computador. Uma alternativa é escrever um programa para um calculador programável a fim de avaliar α e β, como foi feito neste caso.

Faz necessário trabalho adicional para quantificar os níveis de alfa e beta para itens individuais e propiciar procedimentos para modificar os níveis de alfa e beta da prova de acordo. Um desafio ainda maior é o de relacionar os custos envolvidos nas provas mais longas aos benefícios resultantes de decisões instrucionais mais precisas. Evidência sobre esta questão ajudaria o tomador de decisões instrucionais a selecionar uma categoria de extensões de prova (número de erros permitidos). Dentro do quadro de referência proposto aqui, esta é a primeira decisão a ser tomada na determinação de extensões de prova e escores de aprovação. Os custos poderão também ser levados em consideração na seleção de uma extensão de prova dentro da categoria escolhida, mas esta decisão deve normalmente ser baseada principalmente nas probabilidades relativas de evitar a classificação errônea.

REFERÊNCIAS BIBLIOGRÁFICAS

BAKER, E. L. & POPHAM, W. J. 1976. Como ampliar as dimensões dos objetivos de ensino. Porto Alegre, Globo.

BORMUTH, J. R. 1970. On the theory of achievement test items. Chicago, University of Chicago Press.

BRIGGS, L. J. 1976. Manual de planejamento de ensino. Cultrix, São Paulo.

DAVIS, F. B. & DIAMOND, J. J. 1974. The preparation of criterion-referenced tests. *In* C. W. Harris, M. C. Alkin & W. J. Popham, eds., Problems in criterion-referenced measurement. Los Angeles, Center for the Study of Evaluation.

PHANÉR, S. 1974. Item sampling and decision-making in achievement testing. *British Journal of Mathematical and Statistical Psychology*, 27:172-175.

GAGNÉ, R. M. & BRIGGS, L. J. 1974. Principles of instructional design. New York, Holt, Rinehart and Winston.

GEIS, G. L. 1978. Three kinds of behavioral objectives: Total, sample and consequence statements. *Educational Technology*, 18, (1):28-31.

GUTTMAN, L. 1969. Integration of test design and analysis. *In* Proceedings of the 1969 invitational conference on testing problems. Princeton, Educational Testing Service.

HAMBLETON, R. K. *et alii* 1975. Criterion-referenced testing and measurement: A review of technical issues and developments. Trabalho apresentado na reunião anual da American Educational Research Association, Washington, D. C., Abril, 1975.

HAMBLETON, R. K. *et alii* 1978. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 48, (1). 1-48.

HIVELEY, W. 1973. Domain-referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project. Monograph Series in Evaluation, Nº 1. Los Angeles, Center for the Study of Evaluation.

MAGER, R. F. 1962. Preparing instructional objectives. Palo Alto, California, Fearon Publishers.

MILLMAN, J. 1972. Determining test length: Passing scores and test lengths for objectives-based tests. Los Angeles, Instructional Objectives Exchange, 1972.

MILLMAN, J. 1974. Criterion-referenced measurement. *In* W. J. Popham, ed., Evaluation in education: Current applications. Berkeley, McCutchan.

NOVICK, M. R. & LEWIS, C. 1974. Prescribing test length for criterion-referenced measurement. *In* C. W. Harris, C. Alkin & W. J. Popham, eds., Problems in criterion-referenced measurement. Los Angeles, Center for the Study of Evaluation.

OSBORN, H. G. 1968. Item sampling for achievement testing. *Educational and Psychological Measurement*, 28: 95-104.

POPHAM, W. J. 1974. Selecting objectives and generating test items for objectives-based tests. *In* C. W. Harris, M. C. Alkin & W. J. Popham, eds., Problems in criterion-referenced measurement. Los Angeles, Center for the Study of Evaluation.

POPHAM, W. J. 1975. Educational evaluation. Englewood Cliffs, N. J., Prentice-Hall.

POPHAM, W. J. & BAKER, E. L. 1976. Como estabelecer metas de ensino. Porto Alegre, Globo.

SKAGER, R. W. 1974. Generating criterion-referenced tests from objectives-based systems: Unsolved problems in test development, assembly and interpretation. *In* C. W. Harris, M. C. Alkin & W. J. Popham, eds., Problems in criterion-referenced measurement. Los Angeles, Center for the Study of Evaluation.

WILCOX, R. R. 1976. A note on the length and passing score of a mastery test. *Journal of Educational Statistics*, 1, (4): 359-364.

WILCOX, R. R. 1977. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. *Journal of Educational Statistics*, 2, (4): 289-308.