

Construção Automática de Instrumentos de Avaliação

André Cruz Mendes

Universidade Estadual de Maringá
Av. Colombo 5790
Maringá, PR
+55-44-3011-4324
andrecruzmedes@gmail.com

Dante Alves Medeiros Filho

Universidade Estadual de Maringá
Av. Colombo 5790
Maringá, PR
+55-44-3011-4324
dantefilho@gmail.com

ABSTRACT

Evaluation is one of the most important phases of the teaching and learning process. By means of it, evaluators get information for verification that the educational objectives are being achieved or not and in what degree this occurs. The construction of data collection instruments and their measurement are highlighted because the quality of assessment is directly linked to the attributes of validity and reliability of these instruments. In this context, this paper presents a technique for the construction of measuring instruments, particularly tests, based on Bloom's taxonomy to determination of educational objectives and Item Response Theory. This technique was named BLIRT (Bloom and Item Response Theory).

RESUMO

A avaliação é uma das mais importantes fases do processo de ensino e aprendizagem. É com ela que se consegue obter informações para que se verifique se os objetivos educacionais estão ou não sendo atingidos e em que grau isto ocorre. Neste processo, a construção de instrumentos de coleta de dados e sua mensuração ganham destaque, pois, a qualidade da avaliação está diretamente ligada aos atributos de validade e fidedignidade desses instrumentos. Nesse contexto, o presente trabalho apresenta uma técnica para a construção de instrumentos de medida, particularmente de testes, fundamentada na taxonomia de Bloom para determinação dos objetivos educacionais e na Teoria da Resposta ao Item. Esta técnica foi denominada BLIRT (*Bloom e Item Response Theory*).

Categories and Subject Descriptors

K.3.1 Computers use in Education

General Terms

Validation of tests, learning assessment

Keywords

avaliação automática, testes digitais, validade em testes

1. INTRODUÇÃO

O processo educacional escolar é uma atividade que possui como uma de suas características a intencionalidade. Este atributo exige que ele ocorra de forma planejada, organizada e sistematizada. A avaliação integra uma de suas etapas e possui funções didático-pedagógicas, de diagnóstico e de controle. Nela é possível obter uma radiografia de como ocorreu ou esteja ocorrendo o processo de ensino e aprendizagem. Permite analisar o processo educacional e verificar se os objetivos foram alcançados e em que nível, além de contribuir para sua melhoria ao fornecer informações para sua retroalimentação.

Na etapa de avaliação do processo educacional, uma de suas fases é a construção de seus instrumentos. Sob a perspectiva de Cronbach [8], a avaliação deve ser compreendida como uma

atividade diversificada, que pode exigir vários tipos de decisões e de informações. Assim, a avaliação não deve ser confundida com a construção de instrumentos de medida [24], mas a forma de avaliar exige a construção de instrumentos de avaliação específicos e apropriados que possuam validade e confiabilidade. A avaliação só é bem sucedida se está devidamente amparada por instrumentos confiáveis, isto é, depende da qualidade dos instrumentos de coleta de dados.

Um dos instrumentos mais utilizados para a coleta de dados para subsidiar a avaliação escolar presencial e a distância é o teste. Nele são colhidas informações que permitem mensuração de desempenho e, por conseguinte, inferência sobre a ocorrência da aprendizagem. Apesar de muito utilizado, não é um instrumento simples de ser construído. Neste sentido é que o presente trabalho apresenta uma técnica para a construção de testes baseada na taxonomia de objetivos educacionais de Bloom [1], [7] e na Teoria de Resposta ao Item (TRI). Esta proposta tem como fito facilitar a construção de instrumentos válidos e fidedignos.

2. A PREOCUPAÇÃO COM A VALIDADE

A avaliação da aprendizagem pode ser um tema controverso dependendo do contexto considerado (histórico-social, biológico, cognitivo, etc.) e do enfoque dado aos objetivos educacionais, que podem ser expressos em termos de competências, habilidades e atitudes [6], [14], [17]. Todavia, a avaliação da aprendizagem, em suas diferentes abordagens, envolve meios de se constatar se estes objetivos são atingidos, coletando e interpretando dados quantitativos e qualitativos de alunos a respeito daquilo que lhes foi ensinado e depois fazendo juízo dessas interpretações com critérios previamente estabelecidos [20].

Embora existam várias maneiras de se avaliar, este termo tem sido constantemente associado a expressões como “fazer prova”, “exame final”, “notas”, “repetir de ano” ou “passar de ano”. Isso é resultado de uma concepção pedagógica ultrapassada, mas tradicionalmente dominante [5]. Ora, a aprendizagem humana não é uma realidade acabada que se dá a conhecer de forma única e precisa em seus múltiplos aspectos, trata-se de um fenômeno multidimensional, no qual estão envolvidas as dimensões humanas, a filosófica, a social, entre outras [15]. De fato, avaliação de aprendizagem não deve estar em desacordo com o seu conceito qualitativo [10]. A avaliação da aprendizagem, portanto, é complexa demais para se resumir à simples realização de provas e atribuição de notas [13].

Todavia, a avaliação da aprendizagem em termos quantitativos se faz necessária em diversos contextos educacionais, de modo que testes não devem ser dispensados. Mesmo com limitações, essa ainda é a ferramenta que mais se aplica, pois constitui meio de constatar quantitativamente se os alunos desenvolveram as capacidades expressas por meio dos objetivos educacionais. Deve-se, no entanto, prezar pela rigorosidade de tais testes para

que as implicações negativas de suas limitações sejam minimizadas. Tamanho rigor em prol da validade dos testes pode ser visado se, na construção dessas ferramentas, procurar-se aplicar princípios como os envoltos na psicometria contemporânea.

Psicometria é uso de medida em psicologia [16] e constitui-se de um conjunto de técnicas que permite a quantificação dos fenômenos psicológicos [9]. A importância maior está no processo de quantificação, que é complexo, pois o que se deseja medir são variáveis de características subjetivas dos indivíduos, não se tratando de algo observável, mas de construtos hipotéticos [9]. Embora não possam ser medidas diretamente, tais variáveis podem ser estimadas por inferência estatística. Isto é o que ocorre, por exemplo, com a inteligência, cujo significado compreende um conjunto de faculdades mentais [12] que podem ser constatadas, por exemplo, por meio de testes.

Assim, os escores obtidos por meio de testes são medidas objetivas que podem representar a “aprendizagem” e que, portanto, podem servir para a sua avaliação. Isso, no entanto, requer que tais medidas correspondam à realidade, porém, devido a equívocos de associação e interpretação, este pode não ser o caso em diversas situações, por exemplo: No final do século XIX, Cattell e Galton supunham que os mais inteligentes eram os mais rápidos na conclusão dos testes, isto é, **o menor tempo deveria ocorrer nos mais capazes** [23]. Os testes então eram demasiadamente sensoriais e focados em habilidades específicas, mas não demorou para outros pesquisadores perceberem que as medidas assim fundamentadas não tratavam do aspecto intelectual humano. Ainda no início do século XX, Alfred Binet e Théodore Simon criticaram essas medidas e desenvolveram o primeiro teste de inteligência que foi bem sucedido: a escala Binet-Simon [23], constituída de uma série de trinta testes ou tarefas de conteúdo e dificuldade variados com o objetivo de avaliar o julgamento e a capacidade de raciocínio de uma pessoa independente da aprendizagem escolar. Em 1911, William Stern utilizou essa escala para elaborar o conceito de Quociente de Inteligência (QI), que foi revisado e adaptado para utilização com fins militares, mas que também embasou o desenvolvimento de novos testes com finalidades educacionais.

Desde as baterias de testes de aptidões múltiplas, desenvolvidas por meio das técnicas de análise fatorial propostas por Charles Spearman, as contribuições para o desenvolvimento de testes para avaliação da aprendizagem são de cunho multidisciplinar, isto é, vêm da psicologia, da educação, da estatística, da matemática, da ciência da computação e áreas afim. Nesse âmbito, a informática contribuiu principalmente para **a automatização e dinamização de testes**. Atualmente compreende-se que existe diversidade de habilidades cognitivas e que, por isso, não é possível que uma medida única seja capaz de representar a inteligência de uma pessoa como um todo. Fala-se em múltiplas inteligências [11]. Embora Howard Gardner, o precursor dessa nova abordagem, desaprove testes como os de QI e tampouco tenha proposto método para quantificar as inteligências múltiplas, ele não impede outros pesquisadores de formularem tais testes.

No contexto de inteligências múltiplas, diferentes tipos de testes podem ser relacionados a tipos específicos de inteligência [3], por exemplo: testes de leitura e interpretação servem para avaliação de linguística, testes de memória visual e motora servem para avaliação da inteligência espacial, testes de destreza manual se aplicam à cinestésica corporal, testes piagetianos servem para avaliação de lógica matemática, etc.

Em todo caso, a preocupação com a validade em testes aplicados à avaliação da aprendizagem se deve à perspectiva de sua subjetividade, pois, embora sejam constituídos de procedimentos sistemáticos na obtenção de amostras de comportamento e na interpretação dos dados, a forma como os testes são administrados e o examinador que os administra podem afetar os resultados [23].

Observa-se, portanto, que a controvérsia do uso de testes para avaliação da inteligência e, consequentemente, da aprendizagem, reside na desconfiança de que as medidas produzidas por tais instrumentos sejam capazes de representar de maneira adequada o que realmente se deseja medir. Essa desconfiança é encontrada tanto na construção de um teste (o instrumento construído é correto?) quanto na sua administração e respectiva interpretação dos resultados (o que os números realmente significam e como eles podem ser trabalhados?). Se utilizados adequadamente, testes podem produzir resultados que servem de base para inferência estatística sobre indivíduos ou grupos, mas se mal utilizados, por incompetência ou maldade, podem prejudicar pessoas. Desse modo, os instrumentos de teste para avaliação da aprendizagem devem ser construídos e administrados com respeito às competências de interesse no contexto em que as pessoas se encontram.

Quando a natureza da avaliação é quantitativa, os testes devem prezar pela **exatidão, precisão e fidedignidade dos resultados** [22]. Fidedignidade e validade são requisitos que se aplicam à mensuração. Medidas são fidedignas quando são replicáveis e consistentes, são válidas quando representam precisamente algum atributo. Uma régua que mede em milímetros é mais precisa do que uma régua que mede apenas em centímetros, porém, se uma régua produz uma medida de 2cm e 3mm quando a medida real deveria ser 2cm e 5mm, então essa régua é um instrumento de medida que peca na exatidão, isto é, o valor produzido por ela não representa a realidade. Além disso, se essa régua se expande ou se comprime devido à temperatura e, por isso, produz medidas que variam de acordo com o ambiente, então essa régua também não é fidedigna.

2.1. Exatidão em testes

O instrumento construído deve permitir a constatação adequada das competências, habilidades e atitudes de pessoas em seus diferentes níveis. Um teste de matemática, por exemplo, deve contemplar questões que abordem todo o conteúdo de interesse em diversos pontos de vista. Se as questões forem elaboradas por um único avaliador, provavelmente acarretarão vícios de um único ponto de vista sobre o conteúdo. Desse modo, é recomendável que um teste para avaliação da aprendizagem seja elaborado sob a óptica de **vários avaliadores** ao invés de um só.

O teste também deve abranger todos os **níveis cognitivos** da aprendizagem do conteúdo de interesse. Uma taxonomia de objetivos educacionais pode servir de base para a compreensão desses níveis. A taxonomia de Bloom [1], por exemplo, é uma maneira de representá-los. Com base nessa taxonomia, em seu domínio cognitivo, um avaliador pode elaborar questões que servem para constatar se uma pessoa é capaz de “lembrar”, “entender”, “aplicar”, “analisar”, “avaliar” ou “criar” sobre determinado conjunto de conhecimentos. Nota, isso não se trata dos níveis de dificuldade, mas da complexidade de cognição envolvida no processo de aprendizagem. Um processo cognitivo mais complexo não é necessariamente mais difícil, por exemplo, em alguns casos “lembrar” pode ser mais difícil do que “analisar”.

Além disso, deve-se considerar a forma como os resultados do teste são administrados. Os resultados da maioria dos testes são

expressos em escores, que são números com sentidos específicos. As limitações nesse contexto devem ser compreendidas para que não haja equívoco nas inferências realizadas a partir desses escores. Por exemplo, é comum a ocorrência de erros de inferência com escalas ordinais – como quando se realiza cálculo de média aritmética sobre medidas desse tipo [23].

Prezar pela exatidão em testes, portanto, implica em elaborá-los com questões que abrangem diversos níveis cognitivos e que contemplam todo o conteúdo de interesse, bem como administrar os resultados considerando adequadamente suas limitações de contexto, de modo a evitar equívocos na interpretação.

2.2. Fidedignidade em testes

Para que sejam replicáveis e consistentes, testes devem possuir claros critérios de avaliação, que não produzam divergências de resultados em contextos diferentes. Assim, quanto mais objetivos, e menos subjetivos, forem esses critérios, mais fidedigno tende a ser um teste. Além disso, deve-se considerar a forma de aplicação do teste e a linguagem com a qual as questões são apresentadas às pessoas, que podem influenciar os resultados.

Fidedignidade em testes, portanto, implica em minimizar ambiguidades, propiciando a correta interpretação de comando com clareza nos enunciados (“pegadinhas”, por exemplo, podem comprometer a validade), adequando a linguagem ao público alvo e administrando os resultados com critérios objetivos.

2.3. Precisão em testes

O instrumento de avaliação deve discriminar adequadamente quem sabe mais de quem sabe menos, mesmo que a diferença seja pequena. Por exemplo, é aceitável dizer que uma pessoa que obteve desempenho de 90% em determinado teste é mais inteligente (no quesito específico avaliado) do que outra pessoa que obteve desempenho de apenas 10% no mesmo teste, mas essa afirmação poderia ser controversa caso os desempenhos dessas mesmas pessoas fossem respectivamente 60% e 59%. Todavia, se o teste realizado for suficientemente preciso, então diferenças pequenas entre resultados podem ser significativas.

Intuitivamente, pode-se pensar que quanto mais questões um teste possui, maior será sua precisão. Entretanto, isso não é via de regra, pois existem outros fatores a serem considerados. Um teste com muitas questões pode gerar cansaço e desmotivação, que podem enviesar os resultados e comprometer a exatidão. Além disso, um teste pode possuir uma grande quantidade de questões e todas serem difíceis. Um teste assim serve apenas para discriminar quem sabe muito dos que não sabem tanto, isto é, os escores obtidos não serviriam para discriminar quem sabe pouco de quem sabe “mais ou menos”. Por isso, visando à precisão, um teste deve ser constituído de um **número adequado de questões** que contemplem **todos os níveis de dificuldade**, desde as mais fáceis às mais difíceis. Isso possibilitaria ao avaliador constatar em que nível se encontra a aprendizagem de uma pessoa que o responde. Além disso, deve-se considerar a forma como as questões de um teste estão distribuídas em termos estatísticos, pois muitas questões fáceis podem aumentar os escores, enquanto muitas questões difíceis os diminuem. Os testes, portanto, devem ser **balanceados** em termos de dificuldade. A **distribuição normal**, por exemplo, pode ser utilizada para descrever a localização de um escore em relação a uma amostra e também para inferir derivações de intervalos de confiança que avaliam adequadamente os escores obtidos e as diferenças entre eles [23].

Para tanto, a administração do teste e os métodos nele empregados também devem prezar pela precisão. Tradicionalmente,

a avaliação da aprendizagem por meio de testes se baseia na análise de um somatório de constatações de um comportamento esperado. Esse somatório compreende aquilo que se chama de escore (a nota de uma prova). Isso é o que fundamenta a Teoria Clássica dos Testes (TCT), todavia, recentes teorias procuram não se prender aos escores dos testes para quantificar variáveis psicológicas, como a Teoria da Resposta ao Item (TRI) [2], [16]. Fazendo uma analogia, em questão de precisão na avaliação da aprendizagem, se a TCT é uma régua centimetrada, então a TRI é milimetrada.

Prezar pela precisão em testes, portanto, implica em elaborá-los com adequada quantidade de questões, de modo a abranger todas as partes do conteúdo de interesse com distribuição normal de dificuldade, e administrando-as com métodos estatísticos que possibilitem uma análise profunda dos resultados.

3. BLIRT

Com base nos conceitos e desafios expostos no presente trabalho, propõe-se uma técnica fundamentada na taxonomia de Bloom [1], [7] e na TRI [2], [16] para construção de testes de avaliação da aprendizagem compostos por questões objetivas, que seguem distribuição normal de dificuldade, abrangem amplo domínio cognitivo e minimizam vícios de avaliação acarretados pelo ponto de vista do avaliador. O nome BLIRT é um acrônimo de “*Bloom*” e “*Item Response Theory*”. Esta técnica tem cinco fases: produção, revisão, testagem, classificação e escolha.

3.1. Primeira fase: PRODUÇÃO

Testes de aprendizagem elaborados por um único avaliador podem acarretar alguns vícios de avaliação pelo fato de se sujeitarem a um único ponto de vista do conteúdo. A perspectiva dos testes, no entanto, pode ser ampliada se forem elaborados sob o ponto de vista de vários avaliadores ao invés de um só. Desse modo, a primeira fase da técnica BLIRT compreende a produção coletiva de questões candidatas à composição do teste, realizada por vários avaliadores. Quanto mais avaliadores produzirem questões para o teste, maior a minimização de vícios de avaliação. Para fins quantitativos, o grau de minimização desta técnica pode ser expresso pelo logaritmo na base 2 da quantidade de avaliadores. Por exemplo, se apenas um avaliador produzir questões para o teste, então o grau de minimização é 0, pois $\lg(1) = 0$. Seguindo a lógica, dois avaliadores constituem grau 1 de minimização, pois $\lg(2) = 1$. Quatro avaliadores: grau 2, pois $\lg(4) = 2$. Oito avaliadores: grau 3. Dez avaliadores: grau de minimização aproximado de 3,32, e assim por diante.

Além de minimizar os vícios de avaliação do avaliador, as questões do teste construído devem cobrir diversas capacidades cognitivas, afinal a aprendizagem não se limita à mera retenção de informações na memória, mas também à habilidade de utilizá-las em processos mentais mais complexos. Logo, uma taxonomia de objetivos educacionais pode ser usada para categorizar as questões do teste quanto ao estímulo dessas habilidades. Nesta técnica, BLIRT, o domínio cognitivo da taxonomia de Bloom [1] serve de base para tal categorização. Assim, toda questão produzida por um avaliador deve ser classificada, por ele mesmo, em um dos seguintes níveis cognitivos:

- **Lembrar:** questões que servem para constatar se uma pessoa possui determinada informação em sua memória. Exemplo: “Qual é a capital do Brasil? a) São Paulo; b) Rio de Janeiro; c) Brasília.”. O fato de uma pessoa responder corretamente essa questão é indicio de que ela memorizou o nome da capital, mas não significa que tenha entendido o que é uma capital.

- **Entender:** questões que servem para constatar se uma pessoa, além de lembrar-se de um conceito, é capaz de expressar seu significado com palavras diferentes de sua definição original ou relacioná-lo a explicações coerentes. Exemplo: “*O que significa ‘água mole em pedra dura tanto bate até que fura’?*” **a)** *Que a água é um instrumento adequado para furar pedras;* **b)** *Que pedras não são resistentes à água;* **c)** *Que a persistência leva pessoas a atingirem seus objetivos.*”. O fato de uma pessoa identificar a resposta correta neste caso é indicio de que ela entendeu o significado da expressão apresentada e é capaz de relacioná-la a expressões coerentes ao seu contexto.
- **Aplicar:** questões que servem para constatar se uma pessoa, além de entender determinados conceitos, é capaz de resolver problemas que os envolvem em contextos diversos. Exemplo: “*Se $x = 12$, qual das alternativas mais se aproxima do valor da expressão $\sqrt[3]{\frac{3(x-1)^2+15}{25}}$?*” **a)** 2,47; **b)** 3,47; **c)** 4,47.” Se uma pessoa é capaz de encontrar a resposta correta para essa questão, mesmo se as alternativas ou a variável x apresentarem valores diferentes, então isso significa que, além de ter entendido os conceitos de soma, subtração, multiplicação, divisão, exponenciação e radiciação, essa pessoa também é capaz de aplicá-los na resolução de problemas que os envolvem.
- **Analisar:** questões que servem para constatar se uma pessoa é capaz de realizar comparações entre diversas entidades e identificar qual ou quais delas se sobressaiem de acordo com determinado atributo. Exemplo: “*Qual dos veículos de transporte a seguir é mais rápido?*” **a)** Bicicleta; **b)** Carro; **c)** Avião.”. Uma pessoa capaz de identificar a resposta correta dessa questão, além de ter entendido o que são os veículos de transporte apresentados, é capaz de compará-los por meio do atributo “velocidade máxima”. Analisar, portanto, trata-se de um processo cognitivo que seleciona dentre um conjunto de alternativas aquela que se sobressai em um atributo específico.
- **Avaliar:** questões que servem para constatar se uma pessoa é capaz de fazer juízo de diversas entidades em determinado contexto considerando seus atributos. Exemplo: “*Qual dos veículos de transporte a seguir é mais adequado para o funcionário de uma empresa ir de casa ao trabalho?*” **a)** Bicicleta; **b)** Carro; **c)** Avião.”. Para responder essa questão corretamente, uma pessoa precisa conhecer o contexto do percurso da casa do funcionário ao trabalho e analisar uma série de atributos de cada entidade, como velocidade, custo, conforto, *status*, etc. Diante disso, uma decisão deve ser tomada, o que acarreta o processo cognitivo de julgamento de valores.

A taxonomia de Bloom no domínio cognitivo ainda possui um último nível, chamado “**Criar**”, que se refere à capacidade de uma pessoa utilizar de criatividade em conjunto com processos de memorização, entendimento, resolução de problemas, análise e avaliação. Questões dessa categoria geralmente representam uma forma de incentivo para uma pessoa realizar um trabalho teórico ou prático de produção de material técnico, didático ou artístico envolvendo a exposição intrínseca de determinados conceitos. Devido à complexidade desse nível cognitivo, sua constatação é essencialmente subjetiva. Todavia, a técnica BLIRT, aqui proposta, se limita à produção de questões objetivas. Assim, são produzidos apenas itens para avaliação dos níveis cognitivos: “lembrar”, “entender”, “aplicar”, “analisar” e “avaliar”.

Recomenda-se, nesta fase, a produção de uma quantidade 5 vezes maior do que a desejada para compor o teste final. Por exemplo, para construir um teste de 12 questões, recomenda-se

que nesta fase sejam produzidas mais de 60 questões. Uma vez categorizadas, todas elas alimentam um banco de questões, que deve ser submetido à segunda fase desta técnica.

3.2. Segunda fase: REVISÃO

As questões armazenadas no banco são candidatas para a composição do teste a ser construído. Nesta fase, todas as questões produzidas devem ser revisadas. Um revisor é uma pessoa que conhece bem o conteúdo a ser avaliado por meio do teste e a forma adequada como as questões devem ser categorizadas nos diferentes níveis cognitivos da taxonomia de Bloom. Além disso, o perfil do público que se sujeitará ao teste deve ser considerado nessa revisão, isto é, a linguagem das questões deve ser adequada a quem for respondê-las futuramente.

Nota-se, portanto, que a revisão das questões é subjetiva, porém recomenda-se que esta tarefa seja feita ao mesmo tempo por um par ou um pequeno grupo de revisores. As questões devem ser discutidas uma a uma sobre a forma e a linguagem com que são apresentadas. Além disso, os revisores devem verificar se o enunciado possui clareza e se existe ambiguidade nas alternativas de resposta. As alternativas devem ser inequívocas. Por último, os revisores devem verificar se as questões foram categorizadas corretamente de acordo com a classificação adotada nesta técnica, que se baseia no domínio cognitivo da taxonomia de Bloom. No fim desta fase, várias questões podem ter sido corrigidas, reclassificadas ou, até mesmo eliminadas. As questões avaliadas como adequadas para o teste são, então, submetidas à próxima fase desta técnica.

3.3. Terceira fase: TESTAGEM

As questões consideradas como adequadas na revisão devem ser testadas para que se saiba o quão difícil cada uma delas é. Nesta fase, um público menor, porém com perfil equivalente ao das pessoas que serão sujeitas ao teste, deve ser convidado para responder todas elas. Um número de respondentes superior a 30 é suficiente para atender aos fins desta técnica, pois, como acreditam os estatísticos, amostras com mais de 30 pessoas se aproximam de uma distribuição normal [10]. Contudo, é importante ressaltar que a interseção desse conjunto de respondentes com as pessoas que se sujeitarão ao teste futuramente deve ser nula para preservar sua validade.

Enfim, as questões devem ser corrigidas de maneira dicotômica (ou certo, ou errado, sem meio termo). Os dados dessas respostas servem de base para classificação das questões na próxima fase da técnica BLIRT.

3.4. Quarta fase: CLASSIFICAÇÃO

Na TRI, as questões de um teste são chamadas de itens. De acordo com a teoria [2], [16], a dificuldade (b) é um dos parâmetros de um item, que representa o valor de habilidade (θ) que uma pessoa precisa ter para que sua probabilidade de respondê-lo corretamente seja 50%, ou melhor, a metade entre as probabilidades máxima e mínima de acerto (parâmetro c), também conhecida como chute.

Embora complexos, os valores dos parâmetros dos itens de um teste podem ser estimados com o auxílio de programas de computador [4], [18], [19], [21] a partir de um conjunto de respostas dadas a esses mesmos itens. Geralmente, os valores de dificuldade (parâmetro b) são expressos com números (\mathbb{R}) entre -3 e 3 . Nesta fase, portanto, os dados das respostas servem de entrada para algum *software* capaz de estimar os parâmetros dos itens com base na TRI. O valor do parâmetro b estimado para cada item serve para classificá-lo quanto à sua dificuldade.

Contudo, essa classificação também considera a média e o desvio padrão dos valores estimados. Isso é feito da seguinte maneira:

- 1º: Calcula-se a média aritmética simples (M) e o desvio padrão (S) de todos os valores de dificuldade (b) estimados.
- 2º: Calcula-se o valor do “passo” (P), definido nesta técnica como sendo 75% do desvio padrão ($P = 0,75 \times S$). Isso deve ser feito quando a quantidade de itens a comporem o teste é menor do que 30, para evitar que muitos itens sejam classificados com a mesma dificuldade. Se esta técnica for usada para construir um teste com mais de 30 itens, então o valor do passo pode ser igual ao valor do desvio padrão ($P = S$).
- 3º: Calcula-se o desvio (D) de cada item, que é a diferença entre seu valor de dificuldade e a média aritmética ($D = b - M$).
- 4º: Converte-se o desvio (D) de cada item em “passos” ($D \div P$).
- 5º: Enfim, cada item deve ser classificado de acordo com o intervalo numérico em que se encontra o seu respectivo valor de desvio convertido em “passos”. Os intervalos de desvios e as categorias de dificuldades dos itens considerados nesta técnica são apresentados na tabela 1.

Tabela 1. Categorias de dificuldade dos itens.

Desvio (em “passos”)	Categoria
Menor que -3	Facilima
entre -3 e -2	Muito Fácil
entre -2 e -1	Fácil
entre -1 e 0	Meio Fácil
entre 0 e 1	Meio Difícil
entre 1 e 2	Difícil
entre 2 e 3	Muito Difícil
Maior que 3	Difícilima

Tabela 2. Exemplo de classificação de itens de acordo com a dificuldade.

Item	Dificuldade (b)	Desvio ($D = b - M$)	Em passos ($D \div P$)	Intervalo	Classificação
Item 1	-2,92	-2,54	-1,58	entre -2 e -1	fácil
Item 2	-1,85	-1,47	-0,92	entre -1 e 0	meio fácil
Item 3	-0,58	-0,20	-0,13	entre -1 e 0	meio fácil
Item 4	1,12	1,50	0,93	entre 0 e 1	meio difícil
Item 5	2,35	2,73	1,70	entre 1 e 2	difícil
Média (M)	-0,38				
Desvio padrão (S)	2,14				
Passo ($P = 0,75 \times S$)	1,61				

A classificação de cada item quanto à sua dificuldade serve de base para a última fase desta técnica, em que o teste é enfim construído. Com o intuito de exemplificar esse procedimento, a tabela 2 apresenta a classificação de um conjunto de apenas cinco itens de acordo com as dificuldades estimadas para cada um deles.

3.5. Quinta fase: ESCOLHA

Visando à distribuição normal de dificuldade dos itens e considerando os valores de “passo” ($0,75 \times S$), intervalos e as categorias de classificação dos itens nesta técnica, o teste a ser construído deve ser formado de acordo com a seguinte distribuição:

- 1,20% **facilimas** e 1,20% **difícilimas**;
- 5,45% **muito fáceis** e 5,45% **muito difíceis**;
- 15,97% **fáceis** e 15,97% **difíceis**;
- 27,37% **meio fáceis** e 27,37% **meio difíceis**.

Por exemplo, um teste com 16 questões deve conter:

- nenhuma questão facilima e nenhuma difícilima;
- uma questão muito fácil e uma muito difícil;
- três questões fáceis e três difíceis;
- quatro questões meio fáceis e quatro meio difíceis.

Assim, dentre as questões que foram produzidas por vários avaliadores, depois revisadas, testadas e classificadas quanto à dificuldade, devem ser escolhidas algumas que formem um subconjunto com distribuição normal de dificuldade e que, além disso, cubram todos os níveis cognitivos considerados na fase de produção. O resultado final da escolha dessas questões é a composição de um teste que serve como instrumento adequado para a avaliação da aprendizagem em diversos contextos educacionais ou de pesquisa. A figura 1 ilustra um exemplo real do processo de construção de um teste por meio desta técnica.

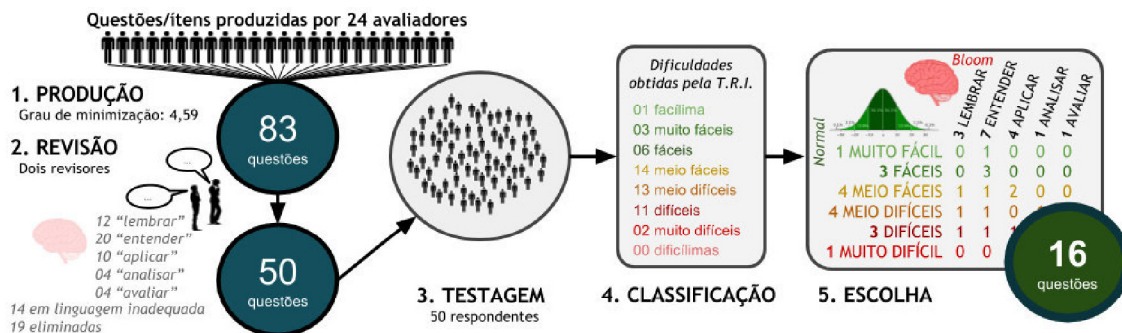


Figura 1. Exemplo real do processo de construção de um teste de aprendizagem com a técnica BLIRT.

O teste construído nesse exemplo foi aplicado a mais de 700 pessoas em um experimento para avaliação do efeito da videoconferência sobre a aprendizagem no contexto de um curso de extensão a distância de formação docente para produção de objetos de aprendizagem, ofertado pela Universidade Estadual de Maringá em parceria com o Ministério da Educação em 2013. Os resultados obtidos por meio do teste construído foram satisfatórios quanto à validade e mostram-se precisos. A tabela 3 é uma forma de apresentar a composição desse teste, considerando a classificação dos itens quanto ao nível de dificuldade e aos objetivos educacionais do domínio cognitivo da taxonomia de Bloom.

Tabela 3. Composição do teste construído no exemplo

	Lem.	Ent.	Apl.	Ana.	Ava.	Total
Muito Fáceis	0	1	0	0	0	1
Fáceis	0	3	0	0	0	3
Meio fáceis	1	1	2	0	0	4
Meio difíceis	1	1	0	1	1	4
Difíceis	1	1	1	0	0	3
Muito Difíceis	0	0	1	0	0	1
Total	3	7	4	1	1	16

4. CONCLUSÕES

Construir testes é um trabalho complexo, que deve considerar diversos aspectos subjetivos da aprendizagem humana. Desse modo, a avaliação da aprendizagem em termos quantitativos pode ser controversa, todavia, os testes ainda são as ferramentas que mais se aplicam para avaliação, pois constituem meio de constatar quantitativamente o desenvolvimento das competências expressas nos objetivos educacionais. Assim, o presente trabalho revisou os desafios encontrados no desenvolvimento de testes e propôs, com base na taxonomia de Bloom e na Teoria da Resposta ao Item, uma técnica para construção de testes que preza pela validade e confiabilidade dos resultados, denominada BLIRT. Espera-se que esta técnica sirva de subsídio ao zelo pela validade no processo de avaliação em contexto educacional e também em pesquisas.

Para tanto, sugere-se como trabalho futuro que essa técnica seja implementada em Ambientes Virtuais de Aprendizagem (LMS), tal como o Moodle, que conta com módulos para elaboração de bancos de questões e questionários, aplicáveis tanto em contexto presencial como a distância. Assim, a construção de testes de aprendizagem mais confiáveis pode ser automatizada e viabilizada em diversos contextos.

5. AGRADECIMENTOS

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) por bolsa de mestrado concedida a André Cruz Mendes.

6. REFERÊNCIAS

- [1] Anderson, L. W.; Krathwohl, D. R. (2001) A Taxonomy for Learning, Teaching, and Assessing: a Revision of Bloom's Taxonomy of Educational Objectives. 2ª ed., Harlow, UK: Longman. ISBN: 080131903X, 9780801319037.
- [2] Andrade, D. F. & Tavares, H. R. & Valle, R. C. (2000) Teoria da Resposta ao Item: Conceitos e Aplicações. São Paulo: Associação Brasileira de Estatística.
- [3] Armstrong, T. (2009) Describing Intelligences in Students. In: _____. Multiple Intelligences in the Classroom. 3ª ed., Alexandria, USA: Association for Supervision and Curriculum Development – ASCD. Cap. 3, p. 32-43. ISBN: 1416607897, 9781416607892.
- [4] Assessment Systems Corporation (2014). Xcalibre [software]. Site oficial do desenvolvedor <http://assess.com/xcart/product.php?productid=415> [8 Set 2014].
- [5] Barbosa, J. R. A. (2008) A Avaliação da Aprendizagem como Processo Interativo: Um Desafio para o Educador. Democratizar, v. 2, nº 1. Instituto Superior de Educação da Zona Oeste / Faetec / Sect – RJ.
- [6] Barbosa, J. R. A. (2011) Didática do Ensino Superior. 2ª ed. Curitiba: IESDE Brasil S.A. ISBN 8538719246, 9788538719243.
- [7] Bloom, B. et al. (1956) Taxonomy of education objectives. New York: David McKay.
- [8] Cronbach, L. J. (1996) Fundamentos da testagem psicológica. 5.ed. Porto Alegre: Artes Médicas.
- [9] Erthal, T. C. S. (2009) Manual de Psicometria. 8ª ed. Rio de Janeiro: Jorge Zahar.
- [10] Fenton, N. E. (1991) Software Metrics: A Rigorous Approach. Londres: Chapman & Hall.
- [11] Gardner, H. (1998) Multiplicity of Intelligences. Scientific American.
- [12] Houaiss, A. (2009) Dicionário Houaiss da Língua Portuguesa / Antônio Houaiss, Mauro de Salles Villar, Francisco Manoel de Mello Franco. Rio de Janeiro: Objetiva: Instituto Antônio Houaiss de Lexicografia.
- [13] Libâneo, J. C. (1994) Didática. São Paulo: Cortez.
- [14] Libâneo, J. C. (2004) A Identidade Profissional dos Professores e o Desenvolvimento de Competências. In: _____. Organização e Gestão da Escola: Teoria e Prática. 5ª ed., Rio de Janeiro: Editora Alternativa. Cap. 4, p. 73-94.
- [15] Mizukami, M. G. N. M. (1986) Ensino: As Abordagens do Processo. São Paulo: EPU.
- [16] Pasquali, L. (2010) Instrumentação Psicológica: Fundamentos e práticas. Porto Alegre: Artmed.
- [17] Perrenoud, P. (1999) Construir Competências é Virar as Costas aos Saberes? Pátio, Porto Alegre: ARTMED, ano 3, n. 11, p. 15-19.
- [18] R-Project. (2014) CRAN – Package ltm [software]. Latent Trait Models under IRT. Disponível em <http://cran.r-project.org/web/packages/ltm> [8 Set 2014]
- [19] R-Project. (2014) CRAN – Package mirt [software]. Multidimensional Item Response Theory. Disponível em <http://cran.r-project.org/web/packages/mirt> [8 Set 2014]
- [20] Santos, J. F. S. (2006) Avaliação no Ensino a Distância. Revista Iberoamericana de Educación, v. 38, n. 4. ISSN: 1681-5653.
- [21] Scientific Software International (2014). Bilog-MG [software]. Site oficial do desenvolvedor <http://www.ssicentral.com/irt/> [8 Set 2014].
- [22] Sellitz, C. & Wrightsman, L. S. & Cook, S. (1987) Métodos de Pesquisa nas Relações Sociais, 2ª ed. São Paulo: EPU.
- [23] Urbina, S. (2007) Fundamentos da Testagem Psicológica. Porto Alegre: Artmed.
- [24] Vianna, H. M. (1989) Introdução à avaliação educacional. São Paulo: IBRASA.