Capítulo

8

Teoria de Resposta ao Item

Ana Liz Souto O. Araújo (UFPB), Jucelio S. Santos (UFCG),
Monilly Ramos A. Melo (UFCG), Wilkerson L. Andrade (UFCG),
Dalton D. S. Guerreiro (UFCG) e Jorge Cesar A. de Figueiredo (UFCG)

analiz@dcx.ufpb.br, jucelio.soares.santos@gmail.com,
monilly.ramos@gmail.com, wilkerson@computacao.ufcg.edu.br,
dalton@computacao.ufcg.edu.br, abrantes@computacao.ufcg.edu.br

Objetivo do Capítulo

Este capítulo tem o objetivo de apresentar diretrizes para guiar a elaboração e a avaliação de itens dicotômicos ou de múltipla escolha em pesquisas na área de Informática na Educação por meio da Teoria Clássica dos Testes (TCT) e da Teoria de Resposta ao Item (TRI). Ao final da leitura deste capítulo, você deve ser capaz de:

- Compreender a importância de elaborar itens confiáveis para a coleta de dados;
- Entender a TCT e a TRI como referenciais para análise dos resultados dos itens;
- Aplicar os princípios da TCT e da TRI nos dados coletados;
- Identificar as limitações da TCT e como a TRI complementa a análise;
- Conhecer os conceitos de Testes Adaptativos Informatizados.

Era uma vez... *Pedro está realizando uma pesquisa sobre uma nova metodologia de ensino de algoritmos para alunos do Ensino Médio. Ele preparou o material didático e ministrou o curso. Ao final do estudo, Pedro elaborou e aplicou um instrumento surpresa, de múltipla escolha, para medir a eficácia da metodologia no aprendizado de algoritmos. Pedro ficou feliz, pois os alunos acertaram todas as questões, e concluiu que a metodologia foi um sucesso. Olhando os resultados com mais calma, Pedro observou um comportamento estranho: alguns alunos que tiveram desempenho baixo no curso, se deram bem no instrumento aplicado. Pedro se questionou se os alunos poderiam ter acertado as questões pelo chute. Ele ainda ficou pensando que elaborou questões fáceis demais para o instrumento. Pedro se questionou se poderia considerar a metodologia eficaz baseado apenas no resultado do seu instrumento. Que boas práticas podem ser usadas para elaborar instrumentos que possam avaliar de maneira fidedigna o aprendizado?*

1 Introdução

A teoria da medida tem por objetivo usar números na descrição de fenômenos naturais (PASQUALI, 2004). O uso de números permite quantificar os fenômenos naturais em estudos científicos. A quantificação é realizada por meio de instrumentos e técnicas de medida que favorecem a compreensão desses fenômenos quando atrelados a métodos científicos. Em pesquisas na área de Educação e Computação, desejamos ter métodos para coletar dados e avaliar habilidades cognitivas dos alunos relacionados aos nossos temas de investigação científica.

A coleta de dados em uma pesquisa pode ser feita por meio de itens dicotômicos ou de múltipla escolha. Itens dicotômicos são aqueles que possuem como resultado apenas duas possibilidades: certo ou errado. Itens de múltipla escolha também podem ser considerados como dicotômicos, uma vez que o resultado geral de cada item pode ser resumido a certo ou errado. Esses itens permitem quantificar, de forma numérica, a estimativa de uma habilidade (COUTO; PRIMI, 2011).

Para colher dados corretos, precisamos conhecer os parâmetros de medida do item, saber como elaborar itens mais confiáveis e que mensurem a habilidade que desejamos avaliar. Um parâmetro designa uma característica fundamental de algo. Os parâmetros mais elementares referentes à análise de um item são a dificuldade e a discriminação. Esses parâmetros são estimados a partir de dados coletados de uma amostra de sujeitos por meio de análise estatística.

A dificuldade de um item pode ser definida em termos do percentual de pessoas que respondem corretamente a um item (PASQUALI, 2001). Por exemplo, em um instrumento, se o item 1 é respondido de forma correta por 75% dos sujeitos, ele é considerado mais fácil em comparação ao item 2, que foi respondido corretamente por 20%.

A discriminação indica o poder do item em distinguir pessoas com diferentes níveis de habilidade examinada. Quanto mais próxima for a magnitude de habilidade que o item puder diferenciar, mais discriminativo o item será, ou seja, mais poder o item tem de diferenciar pessoas com habilidades próximas (PASQUALI, 2001). Ao longo deste capítulo, veremos como estimar os parâmetros de dificuldade e discriminação de um item.

A Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI) são dois referenciais utilizados para construção, validação e avaliação de instrumentos em Psicologia e em Educação (PASQUALI, 2004). No contexto educacional, TCT e TRI podem ser empregadas para avaliação de construtos cognitivos. Construtos são habilidades ou características do sujeito que não podem ser mensuradas diretamente, ou seja, necessitam de um instrumento para serem mensuradas. Por este motivo, o instrumento precisa ser bem construído, de forma que possibilite a medição com menor erro possível. Por exemplo, as habilidades de escrita, aritmética e leitura são construtos que só podem ser mensurados via comportamento do sujeito por meio das respostas que

ele dá a um conjunto de itens. Na psicologia, o Teste de Desempenho Escolar (TDE) (STEIN, 1994) avalia as habilidades de escrita, aritmética e leitura em escolares de 1ª a 6ª séries do Ensino Fundamental.

Neste capítulo você conhecerá como utilizar a TCT e a TRI para construir, validar e avaliar itens como instrumentos de coleta de dados sobre construtos cognitivos para sua pesquisa. Iniciamos apresentando os princípios da TCT e da TRI. Nós focamos aqui em apresentar o modelo logístico unidimensional de três parâmetros, que considera a dificuldade, a discriminação e o acerto ao acaso. Entretanto, esse não é o único modelo da TRI, portanto você pode se aprofundar consultando a seção de Leituras Recomendadas. Em seguida, falamos como Testes Adaptativos Informatizados selecionam itens mais adequados para medir a habilidade dos sujeitos. Escolhemos esta abordagem para que você possa conhecer primeiro as teorias, suas limitações e possibilidades, para então compreender como utilizá-las.

Ressaltamos ainda que você não precisa se preocupar em realizar cálculos complexos à mão no papel. Atualmente existem vários recursos computacionais que realizam toda a parte de cálculos estatísticos. Alguns exemplos são SPSS[1], *Item Response Theory Library* (libirt)[2] com *add-on* para MS Excel e vários pacotes no R, como por exemplo, *mirt* (CHALMERS, 2012) e *ltm* (RIZOPOULOS, 2006).

2 Teoria Clássica dos Testes

TCT considera o escore total de um instrumento como medida para avaliar o desempenho de uma pessoa. O escore total consiste na soma do escore verdadeiro da pessoa somado aos possíveis erros de medida cometidos durante a aplicação do instrumento. O escore verdadeiro seria a medição da habilidade na situação perfeita, ou seja, situação na qual a pessoa irá responder ao instrumento apenas com base no seu conhecimento, sem possibilidade de acerto ao acaso, sem nenhum distrator externo ou interno, além do instrumento não possuir nenhum erro e medir em sua totalidade apenas a habilidade estudada, sem nenhum viés (HUTZ; BANDEIRA; TRENTINI, 2015). Entretanto, não podemos saber o escore verdadeiro de uma pessoa porque não existe a situação perfeita e toda medição inclui erros durante o processo. Erros de medida acontecem quando não podemos controlar todos os fatores que influenciam no processo de medição ou temos erro sistemático, devido a fatores do sistema que interferem na medição (PASQUALI, 2001). Na prática, procuramos minimizar o erro e elaborar itens confiáveis. Portanto, na TCT consideramos o número de acertos no instrumento como escore total.

Outro aspecto da TCT é que ela utiliza normas para interpretar os escores de um instrumento. Essas normas constituem-se um referencial para que possamos interpretar e classificar os escores, por exemplo, situar a posição de um sujeito no construto medido pelo instrumento ou comparar os escores de dois sujeitos (PASQUALI, 2001). Essas

[1] Disponível em: https://www.ibm.com/br-pt/marketplace/spss-statistics

[2] Disponível em: http://psychometricon.net/libirt/

normas são construídas a partir de uma amostra da população, grande o suficiente para ser representativa. Além disso, existem diferentes técnicas de amostragem, como a estratificada, que podem ser tomadas como referência para construção das normas (HUTZ; BANDEIRA; TRENTINI, 2015). Assim, essa amostra é tomada como referência para comparar e classificar os sujeitos que irão futuramente responder ao instrumento. Por exemplo, o TDE possui uma norma que permite avaliar e classificar o desempenho dos alunos nas habilidades de escrita, aritmética e leitura.

Por meio do escore, podemos utilizar algumas medidas para avaliar a qualidade dos itens e do instrumento, como por exemplo o coeficiente de correlação ponto bisserial e o coeficiente alfa de Cronbach. Essas medidas podem ser estimadas por meio de pacotes do R, como *Classical Test Theory* (CTT) (WILLSE, 2018). A seguir, apresentamos o conceito e a principal aplicação do coeficiente ponto bisserial e da consistência interna do instrumento.

2.1 Coeficiente ponto bisserial

O coeficiente ponto bisserial pode ser aplicado como uma medida da capacidade de discriminação do item em relação ao resultado do teste no contexto educacional. A discriminação indica a capacidade do item em diferenciar pessoas com pouca habilidade de pessoas com muita habilidade na tarefa testada (PASQUALI, 2001). O coeficiente permite estimar quais são os itens tal que, se o sujeito avaliado acertar esse item, ele tem mais chances de atingir melhor resultado no instrumento.

O coeficiente ponto bisserial é a correlação de Pearson entre a resposta de um item dicotômico e o escore do instrumento. Essa correlação é calculada entre uma variável categórica (a resposta dicotômica) e uma variável numérica (o escore). Quanto maior for o coeficiente, mais forte é a correlação daquele item com o escore e indica que aquele item é importante para o resultado total do instrumento. O coeficiente varia de -1 a 1 e valores mais próximos de 1 são mais discriminativos. Por exemplo, se o item 4 possui coeficiente igual a 0.75 e o item 6 possui coeficiente igual a 0.55, significa que o item 4 é mais discriminativo que o item 6 e os sujeitos que acertarem o item 4 têm mais chances de obter um bom resultado geral.

2.2 Consistência interna

A consistência interna de um instrumento indica a confiabilidade do instrumento. Ela consiste em examinar a homogeneidade dos itens que compõem o instrumento. O escore total do teste e o escore de cada item são usados para calcular a consistência interna, examinando a significância dos itens que compõem o instrumento.

A forma mais tradicional de calcular a consistência interna é por meio do coeficiente alfa de Cronbach. O valor do coeficiente pode variar entre 0 e 1. Valores mais próximos de 1 indicam que o instrumento possui consistência interna adequada. Valores de alfa entre 0.70 e 0.80 são considerados aceitáveis por alguns autores (BAKER, 2001; PASQUALI, 2004), mas com ressalvas. Valores abaixo de 0.70 não

são aceitáveis e podem significar que as questões que compõem o instrumento precisam ser reavaliadas.

2.3 Limitações da TCT

A TCT possui duas limitações principais: os resultados dependem da amostra de sujeitos que responderam ao instrumento e não há distinção de sujeitos que acertam a mesma quantidade de itens (SARTES; SOUZA-FORMIGONI, 2013). Quando se utiliza a TCT para estimar um construto, o resultado depende do instrumento utilizado e da amostra padronizada, a qual é tomada como referência. Essa amostra precisa ser boa o suficiente para representar toda a população. Por exemplo, a amostra padronizada é tomada como modelo para classificar o sujeito com avaliação alta, mediana ou baixa na habilidade pesquisada. Além disso, um item pode se tornar mais fácil ou mais difícil dependendo se os sujeitos selecionados para responder ao instrumento são mais ou menos habilidosos no construto investigado. Portanto, o parâmetro de dificuldade do item depende da amostra da pesquisa e da dificuldade do construto.

Na TCT, o escore total é a medida da estimativa da habilidade, não permitindo diferenciar sujeitos que acertaram o mesmo número de questões. Isso acontece porque não se estima os parâmetros individuais dos itens, como, por exemplo, a dificuldade de cada item. Como consequência, se dois sujeitos acertarem o mesmo número de questões, eles terão a mesma estimativa da habilidade, mesmo se responderem itens diferentes, com parâmetros de dificuldade diferentes (ANDRADE; LAROS; GOUVEIA, 2010).

3 Teoria de Resposta ao Item

A TRI é uma teoria estatística utilizada pela psicometria e pela área educacional para construção, avaliação e validação de instrumentos (ANDRADE; LAROS; GOUVEIA, 2010; PASQUALI, 2004). Os modelos matemáticos envolvidos dependem do modelo logístico adotado e da dimensão do instrumento. Nesta seção, apresentamos o modelo logístico unidimensional de três parâmetros. Ressaltamos que toda a análise estatística da TRI pode ser realizada por meio de recursos computacionais, como pacotes no R que serão usados na Seção Exemplo Ilustrativo.

A TRI considera a resposta dada ao conjunto dos itens como o elemento capaz de fornecer estimativas para a habilidade avaliada. A estimativa da habilidade, chamada de Theta (θ), está relacionada com a probabilidade de o sujeito responder corretamente aos itens, considerando um ou mais parâmetros (BAKER; KIM, 2017). Por este motivo, a TRI também é conhecida como Teoria do Traço Latente, pois trata construtos como sendo compostos de dimensões, isto é, propriedades de diferentes magnitudes que podem ser mensuradas. Assim, a habilidade avaliada pode ser chamada de construto ou traço latente, pois são termos considerados sinônimos (PASQUALI; PRIMI, 2003).

Na TRI, a habilidade Theta está relacionada com a probabilidade de acertar um item por meio de funções matemáticas. Essas funções matemáticas precisam ser

estimadas e são denominadas de Curvas Características do Item (CCI). A CCI é o gráfico que representa a relação entre a habilidade estimada e o desempenho nos itens (BAKER, 2001). Podem ser usados diferentes modelos matemáticos, dependendo do número de parâmetros envolvidos, da dimensionalidade ou do tipo de itens presente no instrumento. Considerando o modelo logístico unidimensional de três parâmetros, a Equação 1 descreve a probabilidade do sujeito com habilidade Theta de acertar o item *j* dependendo da discriminação *a*, da dificuldade *b* e da chance de acertar pelo chute *c*.

$$P(\theta) = c_j + (1 - c_j)\frac{1}{1 + e^{-a_j(\theta - b_j)}}$$

Equação 1: Modelo logístico de três parâmetros

Slope, threshold e asymptote: parâmetros da TRI

Os nomes dos parâmetros da TRI aparecem em inglês na maioria dos pacotes do R, como no *add-on* Eirt para o MS Excel. Por isso, é importante que você tenha conhecimento dos termos e sinônimos. O parâmetro de discriminação é chamado de *slope*, mas também pode aparecer como *inclination* ou *dispersion*. Já o parâmetro de dificuldade é chamado de *threshold*, bem como *location* ou *position*. Por último, o parâmetro de acerto ao acaso é chamado de *asymptote* (PASQUALI; PRIMI, 2003).

O modelo da Equação 1 é conhecido como modelo logístico unidimensional de três parâmetros, pois usa a função de distribuição acumulada da distribuição logística. Existem outras distribuições que também são usadas na TRI, como função de distribuição normal padrão ou funções de distribuição acumulada contínua. Já o modelo que avalia os parâmetros de discriminação e dificuldade é conhecido como modelo logístico de 2 parâmetros. Neste caso, na Equação 1, o valor de *c* (chance de acertar pelo chute) é considerado zero. Por último, o modelo logístico de 1 parâmetro avalia apenas a dificuldade do item e é também conhecido como modelo Rasch. Neste último caso, o valor da discriminação é fixado em 1. A escolha por um desses modelos depende do ajuste dos dados coletados do mundo real ao modelo.

O gráfico da Figura 1 apresenta três CCI, cada uma correspondendo a um item, modelados a partir da Equação 1, que usaremos como exemplos para ilustrar o comportamento dos parâmetros. O valor de habilidade (Theta θ) pode assumir qualquer número real (BAKER, 2001). Geralmente, o eixo x que representa a escala de habilidade (Theta θ) é avaliado de -3 a +3, mas para facilitar a visualização gráfica, mostramos de -4 a +4. O eixo y representa a probabilidade de resposta correta do item que varia de 0 a 1. As linhas pontilhadas indicam o nível de dificuldade quando a probabilidade de resposta correta for de 50%, pois o parâmetro de dificuldade e a habilidade (Theta θ) estão na mesma escala.

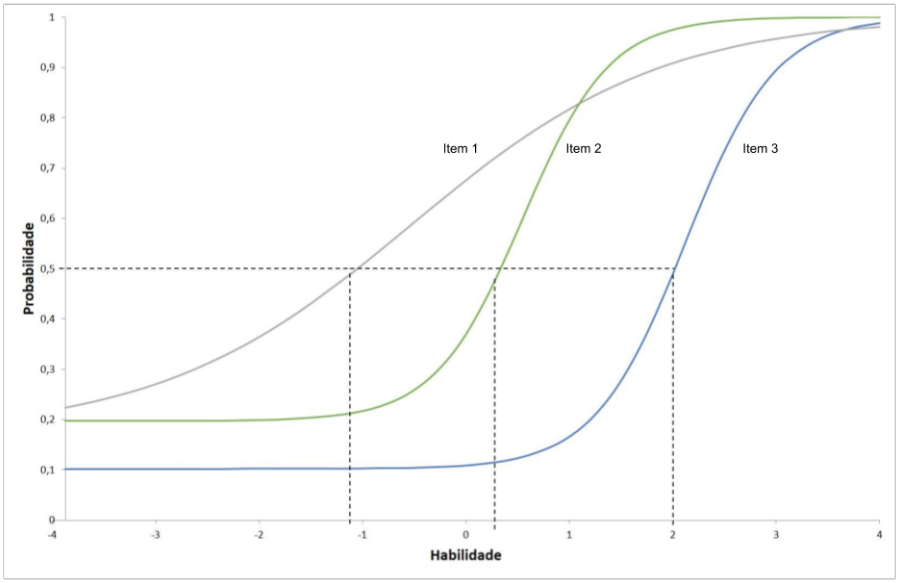


Figura 1: Exemplo de Curva Característica do Item

O parâmetro de discriminação *a* é estimado junto com os outros parâmetros usando a estimação por máxima verossimilhança ou usando alguma característica pontual da distribuição *a posteriori* dos parâmetros. O parâmetro de discriminação é proporcional à inclinação da reta tangente da CCI no ponto *b*, em que *b* é a dificuldade do item. O índice de discriminação pode variar entre 0 (nada discriminativo) até 4 (extremamente discriminativo), mas na prática geralmente varia até 2. Quanto mais inclinada for a curva, maior a discriminação do item. Na Figura 1, os itens 2 e 3 em verde e azul possuem aproximadamente a mesma discriminação e são ambos mais discriminativos em relação ao item 1 em cinza. Observe que os itens 2 e 3 em verde e azul, respectivamente, possuem forma mais próxima de "S" em comparação ao item 1 em cinza.

O parâmetro de dificuldade *b* é estimado na escala de habilidade (eixo x) quando a probabilidade de resposta correta do item é de 50% (eixo y). O índice de dificuldade varia entre -3 (itens fáceis) a +3 (itens difíceis), passando pelo valor 0 que indica um item de dificuldade mediana. Quando um item é considerado fácil, o valor da dificuldade é encontrado em níveis de habilidade mais baixos e a CCI fica posicionada mais à esquerda. Já para itens considerados mais difíceis, o valor da dificuldade é encontrado em níveis de habilidade mais altas e a CCI fica posicionada mais à direita. Na Figura 1, observamos que o item 3 em azul é o mais difícil, pois está posicionado mais à direita e o parâmetro de dificuldade é 2. Já o item 1 em cinza é o mais fácil, pois está posicionado mais à esquerda e o parâmetro de dificuldade é aproximadamente -1.2. Os itens 2 e 3 em verde e azul, mesmo possuindo índices de discriminação próximos, apresentam valores de dificuldade diferentes.

O parâmetro de probabilidade de acerto ao acaso *c* é estimado pela probabilidade de um sujeito sem habilidade no assunto do instrumento acertar um item pelo chute.

Esse parâmetro pode variar de 0 a 1, ou seja, de 0% a 100%. Valores acima de 40% de probabilidade de chute são considerados críticos e o item deve ser revisto. Esse parâmetro é representado na origem da curva em relação ao eixo das ordenadas, a sua extensão é proporcional ao valor de desvio deste ponto em relação ao valor 0. Na Figura 1, os itens 1 e 2 em cinza e verde possuem aproximadamente a mesma probabilidade de chute e são ambos com mais chances de acertar ao acaso em relação ao item 3 em azul.

Dentre as vantagens, podemos destacar que a TRI pode estimar habilidades diferentes para sujeitos que acertaram a mesma quantidade de itens. Isso é possível por meio dos parâmetros individuais de cada item. Outra vantagem é que os parâmetros de dificuldade dos itens possuem valores próximos independente da amostra (HUTZ; BANDEIRA; TRENTINI, 2015). Todavia, pode existir variação se a estimação for realizada com muito erro, como por exemplo, uma amostra composta por sujeitos com pouca variação nos níveis de habilidade (PASQUALI; PRIMI, 2003).

4 Testes Adaptativos Informatizados

Os instrumentos de coleta de dados feitos em papel e lápis têm como principal vantagem a aplicação a um grupo de sujeitos uma única vez, coletivamente. Por exemplo, testes psicológicos utilizam, em sua maioria, respostas escritas para avaliar construtos (PASQUALI, 2001). Entretanto, na maioria dos casos, a correção é manual e pode estar vulnerável a erros humanos. Além disso, testes aplicados em papel precisam conter itens de baixa, média e alta dificuldade no intuito de cobrir todo o espectro de habilidade Theta na população. Assim, os instrumentos precisam ter um número maior de itens e o tempo de aplicação é mais extenso. Consequentemente, os sujeitos precisam responder itens que não correspondem à sua habilidade (ou muito abaixo ou muita acima da sua habilidade).

Recursos computacionais permitem não só automatizar o processo de aplicação e correção do instrumento, como também podem selecionar os itens a serem respondidos. Esta abordagem é chamada de Teste Adaptativo Informatizado (TAI). A primeira vantagem de usar TAI é a automatização para aplicar e corrigir os itens do instrumento. O próprio programa no computador, *tablet* ou celular realiza a correção e armazena os resultados, minimizando possíveis erros manuais durante a correção de um instrumento no papel. A segunda e principal vantagem é a seleção de itens mais adequados a cada sujeito, ou seja, a possibilidade de construir um instrumento sob medida (COSTA, 2009). Assim, o uso de TAI evita que o sujeito precise responder itens que não correspondam com sua habilidade, pois a estimativa da habilidade é atualizada após cada resposta de um item e um novo (próximo) item sempre é escolhido usando a estimativa mais recente da habilidade.

Para construir um TAI precisamos de um banco de itens e um algoritmo de seleção (SEGALL, 2005). Um banco de itens é uma base de dados de questões que possui informações descritivas, psicométricas e outras relevantes para cada item. A parte descritiva contém o enunciado do item, a opção correta e as opções incorretas, quando for o caso de itens de múltipla escolha. A parte de informação psicométrica

apresenta os parâmetros estimados dos itens, tanto os da TCT quanto os da TRI. Por último, as outras informações relevantes podem ser, por exemplo, o conteúdo que cada item mede e a dificuldade teórica. Já o algoritmo de seleção ajusta a escolha dos itens que irão compor o instrumento.

Uma vez criado o banco de itens e implementado o algoritmo de seleção, podemos solicitar ao programa que extraia os itens para compor o instrumento segundo os requisitos desejados. Por exemplo, você tem um banco de itens para avaliar uma habilidade qualquer. Você pode solicitar ao TAI um instrumento de 15 itens que apresentem dificuldade média de 0 com amplitude de -2 e +2 e índice de discriminação entre 0.8 e 1.8. Caso seu banco de itens seja grande o suficiente, o TAI pode montar vários instrumentos com 15 itens diferentes que satisfaçam esses requisitos. Assim, você pode aplicar indiferentemente esses instrumentos a quaisquer sujeitos, e as informações extraídas desta aplicação são comparáveis, quaisquer que sejam os 15 itens selecionados pelo algoritmo. O mesmo aluno pode testar sua habilidade com o mesmo TAI em momentos distintos, pois a seleção de diferentes itens, mas com os mesmos parâmetros, evita que o aluno repita o item e se lembre da resposta.

Caso você deseje minimizar a quantidade de itens do instrumento e o tempo de aplicação, você pode implementar um algoritmo de seleção sem um número fixo de itens. O sujeito vai respondendo itens sucessivos até o programa conseguir estimar a habilidade dele. A condição de parada, ou seja, a quantidade de itens respondidos pelo sujeito, pode ser estipulada ou por uma quantidade mínima/máxima de itens respondidos e/ou por um valor limite de erro padrão.

5 Construção do instrumento

A construção de um instrumento é um processo com várias etapas e pode ser melhor orientado quando temos apoio de uma equipe multidisciplinar para guiar em todas as etapas. Essa equipe pode ser composta por profissionais especialistas na habilidade que será analisada, por estatísticos e por psicólogos psicometristas. Quando o instrumento é computadorizado, como TAI, precisamos também de profissionais da Computação. Nesta seção, veremos as etapas envolvidas no processo de produção de um instrumento.

5.1 Definição do instrumento e elaboração de itens

A primeira etapa na elaboração de um instrumento consiste na formalização do objetivo, na definição da dimensão e das habilidades que serão necessárias para respondê-lo. O objetivo do instrumento indica seu propósito geral, que pode ser, por exemplo, estimar uma nota ou classificar indivíduos. A dimensão do instrumento indica a quantidade de habilidades avaliadas. Neste capítulo, abordamos apenas itens para instrumento unidimensional, mas é possível também construir instrumentos multidimensionais (WISE; KINGSBURY, 2000). Habilidade é a característica a ser avaliada, ou seja, aquilo que o instrumento deseja mensurar. As habilidades que o

instrumento se propõe a medir precisam ser definidas de forma operacional, ou seja, de forma que possam ser observadas na prática.

Quando passamos para a etapa de elaborar itens, precisamos ter domínio sobre a teoria da habilidade que será estimada por meio do item. Todos os itens devem ser elaborados de forma que o examinando possa expressar a habilidade respondendo o item. Com isso em mente, partimos para a escrita do enunciado, a descrição das alternativas, no caso de itens de múltipla escolha, e a indicação da resposta correta.

A elaboração de itens envolve também a definição do tipo e a quantidade de itens que irão compor o instrumento. A definição do tipo do item depende do objetivo do instrumento. Por exemplo, quando se deseja mensurar a proficiência, recomenda-se utilizar um formato de resposta de escolha múltipla. Se o instrumento for computadorizado, o item pode ser formado por recursos multimídias como imagens, animações e sons, por exemplo.

A quantidade de itens a serem elaborados depende do tipo do item. Se os itens elaborados pertencerem ao instrumento de proficiência, deverá ser adotada uma taxa para controlar a exposição deles e uma grande quantidade de itens deverá ser elaborada. Recomenda-se que a quantidade de itens elaborados seja no mínimo três vezes mais do que a quantidade de itens que o instrumento terá, devido ao processo de calibração que provavelmente irá eliminar alguns itens do instrumento (PASQUALI, 2001). Quanto maior for a quantidade de itens no banco, melhor será o instrumento, pois haverá itens mais adequados para um determinado nível de habilidade.

Após os itens serem elaborados, eles precisam ser avaliados teoricamente. A análise teórica pode ser realizada por juízes, ou seja, por especialistas na área pesquisada. Os juízes verificam se os itens são bem compreendidos (análise semântica) e se são adequados para medir a habilidade desejada (análise de conteúdo) (PASQUALI, 2001).

A análise semântica confere se os itens são inteligíveis para todos os sujeitos. Os itens precisam ser de fácil compreensão para todos, até mesmo para os sujeitos com mais baixos traços da habilidade. A análise semântica visa assegurar que a dificuldade de compreensão dos itens não seja ser um fator complicador que possa interferir na resposta do sujeito.

A análise de conteúdo busca assegurar que os itens se refiram à habilidade que se deseja estimar. A quantidade de juízes pode variar, mas recomenda-se que sejam no mínimo três (HUTZ; BANDEIRA; TRENTINI, 2015). Uma concordância de 80% entre os juízes pode ser uma referência para decidir se o item se refere à habilidade e incluí-lo no instrumento (PASQUALI, 2001). Se houver concordância menor que 80%, o item deve ser excluído do instrumento. No caso de se contar com três juízes, é necessário que todos os três concordem a respeito de incluir o item no instrumento. Caso sejam apenas dois juízes, o nível de concordância é de 66,6%, não atingindo os 80% necessários. Entretanto, mais importante que a quantidade é a qualificação do juiz escolhido na especificidade da área da pesquisa.

Os itens podem ser enviados para análise dos juízes, conforme esquema sugerido na Tabela 1 (HUTZ; BANDEIRA; TRENTINI, 2015). O esquema tem a forma de uma tabela na qual cada item é descrito na íntegra e o juiz pode assinalar se o item é ruim, regular ou bom para avaliar o construto da pesquisa em questão. Para cada item, o juiz pode ainda fazer sugestões de alteração. Usando o esquema apresentado na Tabela 1, nós só podemos considerar para a contagem da concordância os itens que forem julgados como "bons" pelos especialistas.

Itens	**Ruim**	**Regular**	**Bom**
Item 1			
Sugestão de alteração			
Item 2			
Sugestão de alteração			
Item 3			
Sugestão de alteração			

Tabela 1: Avaliação de itens por juízes, adaptado de (HUTZ; BANDEIRA; TRENTINI, 2015)

5.2 Calibração do banco e seleção de itens

Após a elaboração dos itens, passamos para a etapa de construir e calibrar o banco de itens. Nesta etapa, os itens já passaram pela análise semântica e análise de conteúdo dos juízes. Agora, esses itens irão compor um banco de itens.

A calibração do banco de itens consiste na aplicação dos itens, na coleta dos dados, na escolha do modelo de resposta e do método de calibração, além da elaboração e interpretação da escala. Na aplicação dos itens precisamos ter uma amostra suficiente de respondentes. O tamanho da amostra de respondentes depende da quantidade de itens do banco, assim, quanto mais itens, maior deve ser a amostra. Por exemplo, para um banco composto por 20 itens, precisamos de uma amostra de no mínimo 200 respondentes, ou seja, pelo menos dez sujeitos para cada item. Para a obtenção da amostra, o instrumento pode ser aplicado no formato tradicional em papel e lápis ou em uma versão computadorizada. Todos os sujeitos deverão responder a todos os itens nessa etapa.

Geralmente o banco de itens é muito grande para que todos os sujeitos respondam a todos os itens. Uma forma de resolver isso é usar técnicas do bloco incompleto balanceado e métodos de equalização (BEKMAN, 2001). Se houver divergência na dimensionalidade do construto do instrumento, podemos escolher um modelo de resposta que pondere essa dimensionalidade, exclua os itens que colaboram para a existência de construtos indesejáveis ou produza mais itens que considerem os construtos desejados e, por último, colete mais amostras para calibrar esses novos itens.

A escolha do modelo de resposta consiste na verificação de qual modelo da TRI se ajusta ao instrumento. Devemos verificar se o ajuste foi adequado e, se necessário, substituir o modelo ajustado. Um modelo mal ajustado não fornecerá parâmetros

constantes para os itens e para as habilidades. Se as estimativas dos parâmetros dos itens por meio da TRI estiverem inconsistentes, por exemplo, apresentando valores absurdos ou erro padrão elevado, isso pode ser causado devido ao tamanho inadequado da amostra.

O método de calibração do banco dos itens consiste em estimar os parâmetros dos itens utilizando critérios da TCT e da TRI. As análises por meio dos índices da TCT ajudarão na eliminação dos itens inadequados e podem ser realizadas usando o pacote *CTT* do R. O processo de calibração dos itens por meio da TRI e a análise dos parâmetros estimados são feitos de forma consecutiva, ou seja, os itens são calibrados e, em seguida, analisados. Essa análise ajudará na decisão da eliminação de itens inadequados. Os parâmetros podem ser estimados por máxima verossimilhança usando o pacote *ltm* do R ou usando uma abordagem bayesiana, usando o pacote *bairt* do R. Após a eliminação desses itens inadequados, as análises por meio dos indicadores da TCT e da dimensionalidade devem ser refeitas para verificar se não foram afetadas pela eliminação dos itens, e o processo de calibração deve ser refeito para verificar se os itens restantes estão adequados e não foram afetados pela eliminação dos outros itens. Um banco de itens é considerado bem calibrado se as estimativas dos parâmetros dos itens forem adequadas e os erros padrões forem baixos.

As estimativas com valores críticos dos parâmetros de discriminação, dificuldade e chute implicam que o item seja retirado do banco. Índice de discriminação abaixo de 0.30 é considerado inadequado para um item ter o poder de diferenciar sujeitos com diferentes estimativas de habilidade. Índice de dificuldade abaixo de -2.95 ou acima de 2.95 também são considerados inadequados pois a escala de habilidade varia de -3 a 3 na prática. Por último, probabilidade de acerto ao acaso acima de 0.40 também é considerado um valor crítico. Em todos esses casos, recomendamos que o item seja excluído do banco de itens para que a estimativa da habilidade não seja comprometida (VENDRAMINI; DIAS, 2005).

Após a calibração final dos itens, devemos fazer algumas verificações no banco. Precisamos verificar se a quantidade de itens que permaneceram no banco de itens é suficiente para a aplicação do instrumento. Também temos que checar se os itens abrangem todo o conteúdo, se estão bem distribuídos e fornecem informação adequada em toda extensão do traço latente avaliado (itens fáceis, médios e difíceis), conforme o objetivo do instrumento. A construção da escala de habilidade é efetuada após a calibração e equalização dos itens. A escala tem o objetivo de fornecer uma interpretação qualitativa dos valores obtidos pela aplicação do modelo da TRI, possibilitando, assim, a interpretação pedagógica dos valores das habilidades (BRITO; MOTTA, 2014).

A seleção dos itens para compor o instrumento pode ser feita por meio de um algoritmo, considerando que nosso instrumento seja adaptativo, ou seja, um TAI (SEGAL, 2005). O algoritmo tem como requisitos a existência de um banco de itens já calibrado, a definição de um critério de seleção do item inicial, o método de seleção dos demais itens, o critério de parada e a taxa de exposição dos itens (YAN; MAGIS, 2016).

O algoritmo geral para seleção de itens pode seguir os passos abaixo:

Passo 1: Estabeleça uma estimativa inicial para habilidade (Theta). Pode ser a mesma estimativa para todos ou ser um número aleatório entre -1 e 1;

Passo 2: Selecione um item usando Theta;

Passo 3: Atualize a estimativa de Theta (aqui usamos estimação por máxima verossimilhança);

Passo 4: O critério de parada está satisfeito? Se sim, o teste acabou. Se não, volte ao passo 2. O critério de parada pode ser quando o erro padrão do cálculo de Theta for menor que 0,01 ou quando o examinado atingir o número mínimo ou máximo de itens aplicados, o que ocorrer primeiro.

O critério de seleção do item inicial determina o grau de informação prévia do examinado. Em geral, caso não tenhamos nenhuma informação sobre o examinado, adotamos um nível de estimativa de habilidade inicial mediano que pode ser fixo, por exemplo, o nível de estimativa de habilidade 0, centrado na média ou um valor aleatório dentro de um intervalo, como entre -1 e + 1 (BAKER, 2001). No caso do valor fixo para estimativa da habilidade inicial, o primeiro item será sempre o mesmo para todos os examinandos. A seleção de um valor aleatório contribui para o controle da exposição dos itens, mas poderá diminuir um pouco a eficiência do instrumento.

O algoritmo seleciona os itens mediante o critério de seleção de itens e do método de estimação da habilidade. Há vários critérios de seleção dos itens que podem ajudar no controle da taxa de exposição dos itens (ANDRADE; TAVARES; VALLE, 2000). Primeiro, podemos selecionar o item usando a estimativa mais recente da habilidade. Neste caso, apresentamos a Máxima Informação de Fisher, mas existem outras, como a Informação de Kullback-Leibler. A Máxima Informação de Fisher seleciona itens procurando maximizar a informação na estimativa atual da habilidade e nos procedimentos Bayesianos que selecionam itens, minimizando a variância *a posteriori*. Segundo, realizamos o procedimento de estimar a habilidade após o indivíduo responder a um item por máxima verossimilhança (SEGAL, 2005). O procedimento de Máxima Verossimilhança é um processo iterativo, usado para estimar a capacidade de um examinado (BAKER, 2001). O processo começa com algum valor *a priori* para a habilidade do examinado e os valores conhecidos dos parâmetros do item. Esses valores são utilizados para calcular a probabilidade de resposta correta a cada item. Em seguida, obtemos um ajuste para a estimativa da habilidade, que melhora de acordo com as probabilidades calculadas da resposta do item. O processo é repetido até que o ajuste se torne pequeno o suficiente para que a alteração na capacidade estimada seja negligenciável, resultando em uma estimativa da habilidade do examinado. Esse processo é então repetido separadamente para cada examinando do teste.

A taxa de exposição dos itens controla a exibição deles para que não se tornem

conhecidos e isso comprometa a confiabilidade do instrumento. O controle da exposição do item fará com que o instrumento não alcance o seu melhor desempenho. Infelizmente, em algum momento do instrumento, o item selecionado não será o melhor para estimar a habilidade devido às restrições necessárias definidas no algoritmo adaptativo. A desvantagem de utilizar a taxa de exposição dos itens é a perda na precisão do instrumento. A precisão do instrumento está relacionada com a informação do item, ou seja, quanto maior a informação que ele fornece, maior a precisão na estimativa da habilidade. Para que essa perda seja praticamente imperceptível, é necessário que o banco de itens possua uma boa quantidade de itens de qualidade, isto é, itens com uma boa quantidade de informação, distribuídos ao longo da escala. Além disso, quanto menor for o valor fixado para a taxa de exposição do item, maior deverá ser a quantidade de itens no banco.

O critério de parada determina o momento em que o instrumento finaliza, ou seja, quando para de selecionar novos itens (SEGAL, 2005). Um critério de parada é considerar que a habilidade estimada alcance um nível mínimo aceitável de precisão, ou seja, um determinado erro padrão mínimo. Outro critério muito utilizado é estabelecer uma quantidade fixa de itens no instrumento. Se essa quantidade de itens for grande, em muitas situações, poucos itens podem ser o suficiente para estimar o traço latente com precisão ou classificar o examinando acima ou abaixo de um ponto de corte.

5.3 Análise da precisão e da validade

Após a construção do algoritmo de seleção de itens, precisamos avaliar o algoritmo mediante algum controle psicométrico de qualidade para verificar a precisão e a validade do instrumento. Podemos fazer esta avaliação por meio de dados empíricos ou simulações.

Vários algoritmos para um mesmo instrumento podem ser projetados, combinando diferentes itens, critérios de seleção de itens, métodos de estimação do traço latente, critérios de parada e restrições (SEGALL, 2005). Por exemplo, podemos formar um banco de itens com parâmetros aceitáveis e outro banco mais rígido, selecionando apenas itens com desempenho acima do aceitável. Nos casos em que a quantidade de itens é baixa, podemos montar bancos com diferentes tamanhos e testá-los por meio de simulações para verificar até que ponto os itens menos adequados interferem na qualidade do instrumento. Os algoritmos projetados serão testados e avaliados. Pode ser analisado o desempenho de um único instrumento ou comparado o desempenho de várias opções de instrumento. Na verificação do desempenho de um único instrumento, devemos escolher os critérios de avaliação, realizar a simulação e analisar o instrumento. Se as análises mostraram que o instrumento é adequado, ele está pronto para ser implementado e utilizado, caso contrário, o instrumento deverá ser reformulado.

Se o objetivo do instrumento for a estimação do traço latente utilizando um nível de precisão como critério de parada, alguns critérios podem ser considerados para esta análise. Em relação à precisão e à validade (de conteúdo, de construto e de predição),

temos o erro padrão médio de estimação, a raiz quadrada do erro quadrado médio, o desvio empírico médio, a eficiência, as correlações entre a habilidade simulada e a habilidade estimada, entre outros procedimentos provenientes da TCT (MUÑIZ; HAMBLETON, 1999). Os resultados são analisados verificando o quanto o algoritmo testado conseguiu recuperar da habilidade simulada, ou seja, se as habilidades estimadas são as mesmas ou próximas das habilidades simuladas. Os resultados são analisados segundo os critérios que foram selecionados anteriormente.

As comparações entre os algoritmos são feitas buscando verificar qual deles obteve o melhor desempenho. A decisão sobre qual algoritmo escolher é tomada baseada nos critérios que foram selecionados anteriormente. Um instrumento é válido quando ele mede o que presumivelmente deve medir.

6 Exemplo Ilustrativo

Nosso pesquisador Pedro percebeu que seu instrumento pode ter tido resultados questionáveis porque ele não elaborou nem analisou os itens seguindo algum método. Ele entendeu que é importante seguir um método para obter resultados mais confiáveis. Então, Pedro leu este capítulo, se aprofundou nos conceitos e resolveu elaborar um instrumento seguindo as diretrizes da TCT e TRI.

Pacotes R para TCT e TRI

Existem diversos pacotes R para estimar parâmetros da TCT e TRI. Listamos os principais pacotes R que você pode usar para estimar esses parâmetros. A especificação de todos os pacotes, juntamente como as funções, pode ser encontrada em <https://cran.r-project.org/>.

- ***CTT*** (*Classical Test Theory Functions*) - O pacote *CTT* fornece funções para estimar parâmetros da TCT para itens dicotômicos e politômicos, como confiabilidade (alfa de Cronbach) e estatística dos itens.
- ***ltm*** (*Latent Trait Models*) - O pacote *ltm* apresenta funções para estimar parâmetros da TRI pela Máxima Verossimilhança Marginal. Para dados dicotômicos, inclui modelos Rasch, modelo 2PL e modelo 3PL de Birnbaum.
- ***irtoy*** (*A Collection of Functions Related to Item Response Theory*) - O pacote *irtoy* oferece um conjunto de funções para calcular parâmetros básicos da TCT e da TRI para itens dicotômicos.
- ***bairt*** (*Bayesian Analysis of Item Response Theory Models*) - O pacote *bairt* fornece funções para estimar parâmetros de modelos de 2 e 3 parâmetros usando o estimador Bayesiano.
- ***catR*** (*Generation of IRT Response Patterns under Computerized Adaptive Testing*) - O pacote *catR* oferece funções para usar em TAI, como estimativa da habilidade, seleção do primeiro e do próximo item no TAI, bem como regras para condições de parada no TAI.
- ***mstR*** (*Procedures to Generate Patterns under Multistage Testing*) - O pacote *mstR* apresenta funções para teste computadorizado (*Multistage Testing*).

Pedro iniciou definindo o objetivo do seu instrumento e o conceito da habilidade

que ele deseja medir. Assim, o objetivo do instrumento que Pedro está desenvolvendo é estimar a habilidade em construir algoritmo. Na pesquisa, algoritmo foi definido como a capacidade de uma pessoa em descrever passos a serem seguidos em sequência para atingir um objetivo. Em seguida, Pedro elaborou um conjunto de trinta itens que aplicassem esse conceito, pois ele deseja que seu instrumento final tenha no mínimo dez itens, ou seja, ele elaborou três vezes mais itens do que o número mínimo de itens que ele deseja que tenha no instrumento final. Ele lembrou que o processo de calibração provavelmente irá eliminar alguns itens do instrumento.

Pedro validou teoricamente os itens elaborados. Ele convidou três pesquisadores, especialistas na área de pesquisa em algoritmos para alunos do Ensino Médio, para julgarem se os itens trabalham a habilidade de construir um algoritmo. A cada um dos especialistas, Pedro forneceu os itens em um arquivo para que o especialista afirmasse se o item está adequado ou não e se há alguma sugestão de mudança ou correção no item.

Quando Pedro recebeu os dados dos especialistas sobre os itens, organizou uma tabela e verificou o grau de concordância entre os especialistas. Como foram três especialistas, para um item ter grau de concordância maior que 80%, Pedro precisou considerar que todos os três especialistas concordaram. Caso Pedro considerasse apenas a concordância de dois especialistas, seria apenas 66,6% de concordância (e não 80%). Após avaliar o grau de concordância de cada item, nesta etapa, Pedro precisou excluir dez itens que foram apontados pelos especialistas como não adequados para medir a habilidade de construir algoritmos. Assim, ele entendeu que nem todos os itens elaborados por ele eram satisfatórios para estimar a habilidade de algoritmo segundo os especialistas na área.

Prosseguindo agora com vinte itens, Pedro realizou uma aplicação com uma amostra de alunos e, em seguida, organizou os dados para análise baseada na TCT e na TRI. Pedro utilizou o R Studio e o pacote *ltm* para estimar os parâmetros. Para estimar a correlação bisserial de um item com o escore do instrumento, Pedro usou a função *biserial.cor( )*, passando como parâmetros o escore `rowSums(dados.itens)`, um item `dados.itens[[1]]` e informando que a questão é dicotômica `level=2`, conforme exemplo de código R:

```
> library(ltm)

> biserial.cor(rowSums(dados.itens), dados.itens[[1]],level=2)
[1] 0.7599672
```

O resultado da correlação bisserial do item 1 com o escore foi de 0.75, indicando que esse item tem forte correlação com o resultado do teste pela TCT.

Em seguida, Pedro calculou o índice de confiabilidade alfa de Cronbach por meio da função *cronbach.alpha( )* do pacote *ltm*. O código a seguir mostra a função *cronbach.alpha( )* passando como parâmetro os itens `dados.itens`.

```
> cronbach.alpha(dados.itens)

Cronbach's alpha for the 'dados.itens' data-set

Items: 20
Sample units: 100
alpha: 0.869
```

O resultado mostra que o valor de alfa foi 0.869. Esse valor é considerado aceitável para consistência interna do instrumento (BAKER, 2001; PASQUALI, 2004).

Pedro continuou com a análise e estimou os valores dos parâmetros de discriminação, de dificuldade e de acerto ao acaso. Ele utilizou o pacote *ltm* e a função *tpm( )*, a qual é utilizada para a estimação dos parâmetros dos itens de um instrumento. O código a seguir mostra a chamada da função *tpm( )*, passando as respostas dos itens `dados.itens` e o `control` por meio de lista de valores de controle com elementos que são otimizados por uma cadeia de caracteres '`nlminb`'. Em seguida, mostramos a saída da função:

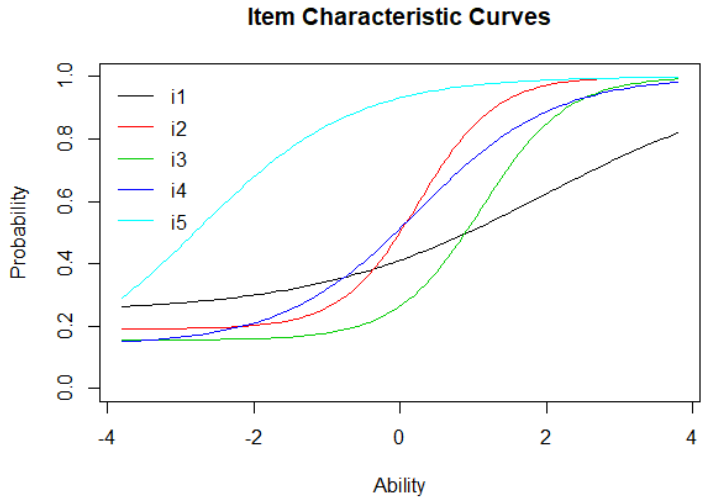
```
> library(ltm)
> dados.tpm <- tpm(data = dados.itens, control = list(optimizer
= "nlminb"))
> dados.tmp

Coefficients:
     Gussng  Dffclt  Dscrmn
i1    0.244   1.982   0.635
i2    0.191   0.254   1.896
i3    0.156   1.114   1.718
i4    0.140   0.246   1.078
i5    0.022  -2.765   0.935
i6    0.022  -0.474   0.561
i7    0.493   0.798   2.318
i8    0.732   0.004   3.159
i9    0.077   0.437   0.958
i10   0.591   0.529   1.317
i11   0.000   0.975   0.581
i12   0.173   2.198   2.814
i13   0.215   2.352   2.249
i14   0.300   2.589   0.785
i15   0.001   0.995   0.958
i16   0.194  -0.243   1.363
i17   0.578   0.605   1.987
i18   0.067   0.361   1.518
i19   0.000  -1.089   1.219
i20   0.327  -3.617   0.264
```

Na saída da função, a primeira coluna corresponde aos itens do instrumento enumerados na ordem do arquivo de origem. Já a coluna `Dscrmn` corresponde ao parâmetro de discriminação *a*. A coluna `Dffclt` mostra os parâmetros de dificuldade *b*. Por último, a coluna `Gussng` corresponde ao parâmetro de acerto ao acaso *c*. Analisando os números, Pedro observou que os itens i7, i8, i10 e i17 possuem o parâmetro de acerto ao acaso *c* maior do que o valor crítico de 0.40 (colocamos em negrito para facilitar sua visualização). Além disso, o item i20 possui os parâmetros de discriminação *a* e dificuldade *b* abaixo dos valores críticos (colocamos em negrito para facilitar sua visualização). Portanto, os itens i7, i8, i10, i17 e i20 foram eliminados nesta etapa, pois possuem estimativas de parâmetros fora do padrão esperado. Após a retirada desses itens, Pedro estimou novamente os parâmetros e os valores estavam todos aceitáveis. Agora Pedro conta com quinze itens.

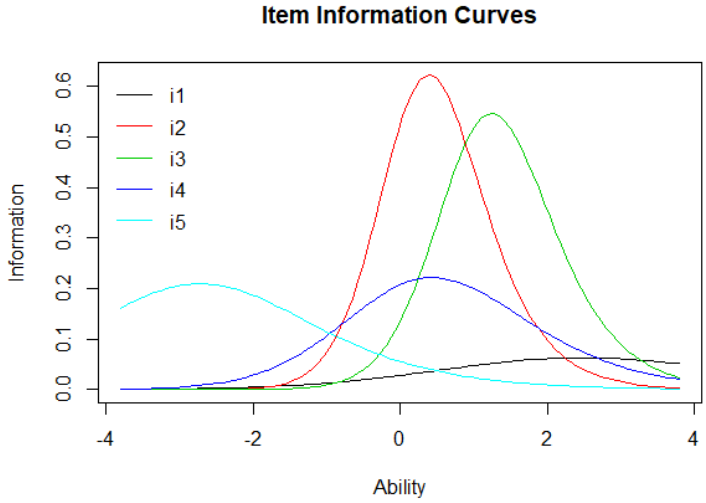
Logo em seguida, Pedro elaborou as CCI dos cinco primeiros itens do instrumento por meio da função *plot( )* do pacote *ltm*, a qual é utilizada para plotagem de gráficos no R. O código a seguir mostra a chamada da função *plot( )*, passando os parâmetros dos itens *dados.tpm*. Em seguida, exibimos o gráfico gerado que mostra uma representação das CCI dos cinco primeiros itens:

```
> plot(dados.tpm, items=1:5, legend=TRUE)
```



Pedro também plotou as funções de informação dos cinco primeiros itens do instrumento. Pedro utilizou novamente a função *plot( ),* a qual é utilizada para plotagem de gráficos no R. O código a seguir mostra a chamada da função *plot( )*, passando os parâmetros dos itens `dados.tpm` e o tipo da função `'IIC'`. Em seguida, mostramos o gráfico gerado contendo as funções de informação dos cinco primeiros itens do instrumento, as quais fornecem a quantidade de informação de cada item em uma determinada região do traço latente (habilidade):

```
> plot(dados.tpm, type='IIC', items=1:5, legend=TRUE)
```



Continuando a análise, Pedro fez a estimativa da habilidade para todos os examinados. Ele utilizou o pacote *irtoys* e a função *eap( )*, a qual é utilizada para estimar as habilidades dos examinados pela expectativa *a posteriori*, ou seja, este estimador depende da média *a priori* de todos os examinados garantindo a mesma estimativa. O código a seguir mostra a chamada da função *est( )* que estima novamente os três parâmetros da TRI `model="3PL"` usando a `engine="ltm"` no formato adequado para serem parâmetros na próxima função. Por sua vez, a função *eap( )* estima a habilidade Theta de cada examinando, passando as respostas dos itens `dados.itens`, os parâmetros dos itens estimados anteriormente `dados.itens.par` pela função *est( )* e a escolha de um objeto de quadratura produzida na normal `qu=normal.qu()`. Mostramos a saída da função referente aos seis primeiros examinados usando a função `head(hab)`:

```
> library(irtoys)

> dados.itens.par <- est(dados.itens, model="3PL", engine="ltm")
> hab <- eap(dados.itens, dados.itens.par, qu=normal.qu())
> head(hab)

              est       sem  n
[1,]  0.61104042 0.4771454 15
[2,]  1.05883705 0.3089617 15
[3,]  1.92623806 0.4680334 15
[4,] -1.71143568 0.6046471 15
[5,]  0.47840008 0.3540985 15
[6,] -0.79040561 0.4448174 15
```

O resultado acima mostra os valores para os seis primeiros alunos. A coluna `est` representa a estimativa da habilidade para cada aluno. Já a coluna `sem` mostra o erro

padrão da estimativa. Por último, a coluna `n` mostra o número de itens do instrumento, que nesta etapa corresponde a 15 itens.

Em seguida, Pedro construiu o banco de item formado por quinze itens e implementou o algoritmo de seleção para seu instrumento adaptativo. Para o algoritmo, Pedro utilizou os procedimentos de Máxima Informação de Fisher e da Máxima Verossimilhança. O critério de seleção do primeiro item foi o que tinha dificuldade mais próxima de 0 e o critério de parada foi quando o erro padrão fosse menor que 0.01 ou quando o número máximo de itens administrados for dez, o que ocorrer primeiro.

Por último, Pedro estimou novamente as habilidades para todos os examinados utilizando o algoritmo adaptativo. Em seguida, ele calculou a correlação da habilidade simulada pelo algoritmo e a habilidade estimada com todos os itens para verificar a precisão do algoritmo adaptativo. O código a seguir mostra a chamada da função *cor( )*, passando como parâmetros os dois valores das habilidades dos examinados estimada pelo instrumento sem algoritmo `hab$esp` e com algoritmo `hab.al$esp`, e por último o método Spearman para calcular a correlação `method ="spearman"`:

```
> cor(hab$esp, hab.al$esp, method ="spearman")
> [1] 0.8459782
```

As correlações das habilidades tiveram valores altos indicando que o algoritmo está adequado para estimar habilidades dos examinados. Finalmente, Pedro possui um banco de itens calibrado e um algoritmo de seleção adequado para aplicar na pesquisa da nova metodologia de ensino de algoritmos.

7 Resumo

Neste capítulo foram apresentadas diretrizes para guiar a elaboração e avaliação de um instrumento para pesquisas na área de Informática na Educação por meio da TCT e da TRI. Ambas as teorias são utilizadas para validar e avaliar itens como instrumentos de coleta de dados sobre construtos cognitivos. A TRI é composta por um conjunto de modelos matemáticos que considera o item como unidade fundamental de análise (e não o escore total como na TCT) e procura representar as chances de um indivíduo dar uma resposta a um item em relação a seus parâmetros (discriminação, dificuldade e chute ao acaso, como no caso de modelo logístico de três parâmetros) e a estimativa da habilidade Theta do indivíduo. Em seguida, apresentamos o TAI que pode informatizar o processo de aplicação e correção de instrumentos, embora seu maior destaque seja selecionar itens mais adequados ao sujeito de acordo com sua habilidade Theta.

O TAI é administrado pelo computador ou tablet (por meio de um algoritmo de seleção) que procura encontrar itens próximos do nível de habilidade de cada sujeito examinado. Essa habilidade é estimada iterativamente durante a administração dos itens presentes em um banco de itens do instrumento. O instrumento precisa ser definido mediante a formalização do objetivo, da definição da dimensão e das habilidades que serão necessárias para respondê-lo.

A seleção de itens que irão compor o TAI (banco de itens) primeiro envolve a participação de especialistas para avaliar teoricamente cada item, se são bem compreendidos (análise semântica) e se são adequados para medir a habilidade desejada (análise de conteúdo), e segundo a calibração dos itens. Os itens que tiverem mais de 80% de concordância entre os especialistas irão para o processo de calibração. Esse processo de calibração envolve a aplicação dos itens (coleta dos dados) e a estimativa dos parâmetros da TCT e TRI de acordo com o modelo que melhor se adequa aos dados. A seleção dos itens para compor o instrumento pode ser feita por meio de um algoritmo que tem como requisitos a existência de um banco de itens já calibrado, a definição de um critério de seleção do item inicial, o método de seleção dos demais itens, o critério de parada e a taxa de exposição dos itens.

Uma forma de avaliar o algoritmo mediante algum controle psicométrico de qualidade é verificar a precisão e a validade do instrumento. Podemos fazer essa avaliação por meio de dados empíricos ou simulações. No final deste capítulo, apresentamos um exemplo ilustrativo como forma de aplicar a teoria vista no capítulo em um cenário de uma pesquisa fictícia.

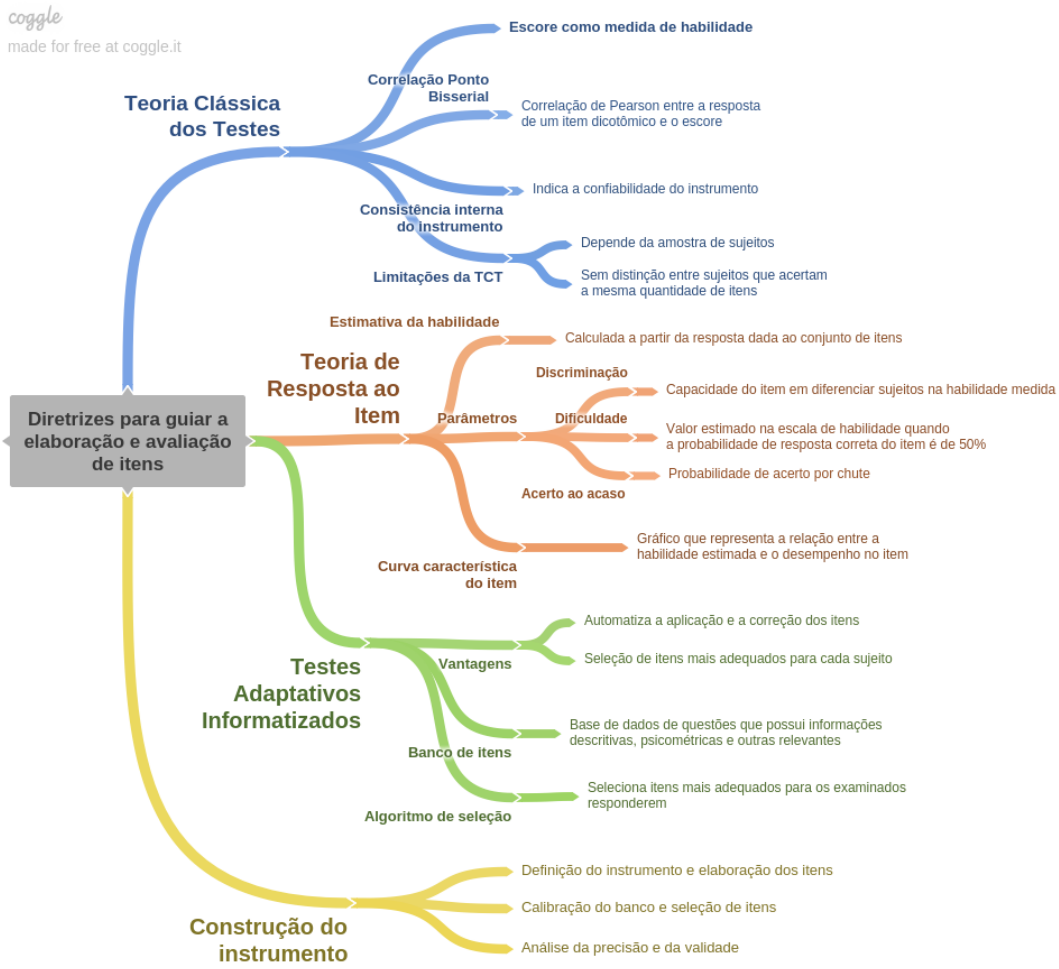


Figura 2: Mapa mental

8 Leituras Recomendadas

- **The Basics of Item Response Theory Using R.** (BAKER; KIM, 2001). Neste livro você encontrará aprofundamento dos conceitos sobre a TRI. Você terá maiores detalhes sobre como interpretar a Curva Característica do Item. Você conhecerá a função de informação do item e a função de informação do teste, além de como utilizá-las para selecionar itens que apresentam maior informação em uma determinada magnitude da habilidade.
- **Computerized adaptive and multi-stage testing with R.** (YAN; MAGIS, 2017). Este livro apresenta tanto a teoria como um guia prático para construção de TAI (*Computerized Adaptive Testing* - CAT). Os autores descrevem como utilizar os pacotes do R *catR* e *mstR* por meio de exemplos práticos.
- **Psicometria: Teoria dos testes na Psicologia e na Educação.** (PASQUALI, 2004). Este livro apresenta a teoria psicométrica envolvida na elaboração de itens para um instrumento de coleta de dados.
- **Uso da Teoria de Resposta ao Item em Avaliações Educacionais: Diretrizes para Pesquisadores.** (ANDRADE; LAROS; GOUVEIA, 2010). Neste artigo são discutidos os pressupostos, os modelos e as aplicações da TRI em avaliações educacionais de larga escala.

9 Artigos exemplos

- **Explorando Teoria de Resposta ao Item na Avaliação de Pensamento Computacional: um Estudo em Questões da Competição Bebras**. (ARAUJO et al., 2018). Neste artigo exploratório você verá a TRI sendo aplicada para estimar parâmetros dos itens de uma prova usada para disseminar Pensamento Computacional.
- **Recomendação de Jogos na Aprendizagem da Matemática baseada na Análise Diagnóstica e Teoria de Resposta ao Item.** (BRITO; MOTTA, 2014). Neste artigo você encontrará a aplicação da TRI para validar um sistema de recomendações de jogos educacionais para ensino da matemática em turmas do ensino médio.
- **Modelagem de Usuários Baseada em Estilo de Aprendizagem, Teoria da Resposta ao Item e Lógica Fuzzy para Sistemas Adaptativos Educacionais.** (CARNEIRO; LIMA NETO; SILVEIRA, 2013). Neste artigo você encontrará uma aplicação da TRI em conjunto com lógica fuzzy na proposta de um mecanismo de modelagem de perfil cognitivo para personalização de treinamentos online de acordo com o ritmo de aprendizagem de cada aluno.
- **Practical Issues in Developing and Maintaining a Computerized Adaptive Testing Program.** (WISE; KINGSBURY, 2000). Neste artigo você verá uma prática de implementação de instrumentos adaptativos, apresentando procedimentos essenciais para profissionais que lidam com medição, como manutenção dos itens no TAI, procedimento para administração do teste e segurança do TAI.

10 Checklist

Resumidamente, para utilizar a TCT e a TRI em sua pesquisa usando um TAI, você necessitará realizar as etapas brevemente descritas a seguir e também ilustradas na Figura 3:

- Elaborar itens de acordo com seus objetivos de pesquisa;
- Validar teoricamente os itens com especialistas no assunto;
- Revisar os itens e excluir aqueles que não tiveram concordância maior que 80% entre os especialistas. Se necessário, construir mais itens e validar teoricamente com especialistas;
- Aplicar o instrumento;
- Analisar os dados segundo os princípios da TCT e TRI e excluir itens com parâmetros críticos;
- Analisar os dados coletados segundo os princípios da TCT e da TRI. Se necessário, aplicar o instrumento novamente;
- Construir e calibrar o banco de itens. Se necessário, analisar dados do instrumento novamente;
- Implementar algoritmo de seleção de itens;
- Analisar precisão e validade do instrumento. Se necessário, voltar para a etapa anterior;

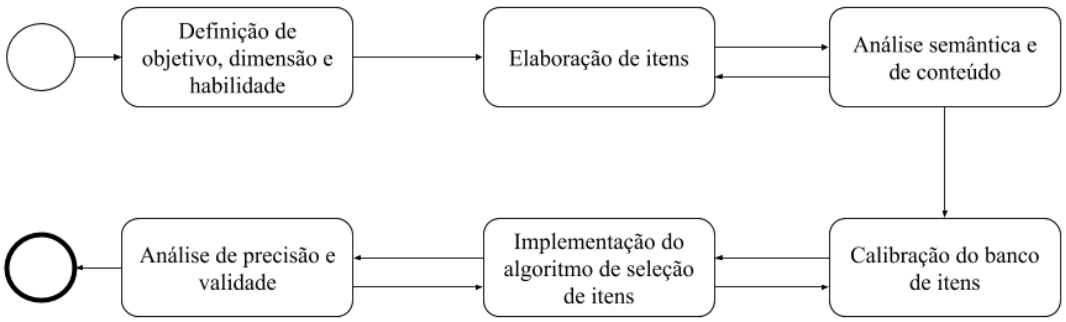


Figura 3. Fluxograma de atividades para construção de itens para o instrumento

11 Referências

ANDRADE, D.; TAVARES, H.; VALLE, R. C. **Teoria da Resposta ao Item**: Conceitos e Aplicações. São Paulo: ABE, 2000.

ANDRADE, J. M.; LAROS, J. A.; GOUVEIA, V. V. O Uso da Teoria de Resposta ao Item em Avaliações Educacionais: Diretrizes para Pesquisadores. In: **Avaliação**

Psicológica, v. 9, n. 3, p. 421-435, 2010.

ARAUJO, A. L. S. O. et al. Explorando Teoria de Resposta ao Item na Avaliação de Pensamento Computacional: um Estudo em Questões da Competição Bebras. In: 29° SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, **Anais...** Fortaleza, 2018.

BAKER, F. B. **The Basics of Item Response Theory**. Washington: Eric, 2001.

BAKER, F. B.; KIM, S. **The Basics of Item Response Theory Using R**. Springer, 2017.

BEKMAN, R. M. Aplicação dos Blocos Incompletos Balanceados na Teoria da Resposta ao Item. **Estudos em Avaliação Educacional**, São Paulo, n. 24, jul-dez, 2001.

BRITO, W.; MOTTA, C. L. R. Recomendação de Jogos na Aprendizagem da Matemática baseado na Análise Diagnóstica e Teoria de Resposta ao Item. In: 25° SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, **Anais...** Dourados, 2014.

CARNEIRO, R. E.; LIMA NETO, F. B.; SILVEIRA, D. S. Modelagem de Usuários Baseada em Estilo de Aprendizagem, Teoria da Resposta ao Item e Lógica Fuzzy para Sistemas Adaptativos Educacionais. In: 24° SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, **Anais...** Campinas, 2013.

CHALMERS, R. P. mirt: A Multidimensional Item Response Theory Package for the R Environment. **Journal of Statistical Software**, v. 48, n. 6, p. 1-29, 2012.

COSTA, D. R. **Métodos Estatísticos em Testes Adaptativos Informatizados.** Dissertação, Universidade Federal do Rio de Janeiro. Instituto de Matemática. Pós-Graduação em Estatística, Rio de Janeiro, março, 2009.

COUTO, G.; PRIMI, R. Teoria de Resposta ao Item: Conceitos Elementares dos Modelos para Itens Dicotômicos. **Boletim de Psicologia**, v. 61, n. 134, p. 1-15, 2011.

HUTZ, C. S.; BANDEIRA, D. R.; TRENTINI, C. M. **Psicometria.** Artmed Editora, 2015.

MUÑIZ, J.; HAMBLETON, R. Evaluación psicométrica de los tests informatizados. In: OLEA, J.; PONSODA, V.; PRIETO, G. (Ed.). **Tests informatizados:** Fundamentos y aplicaciones. Madrid: Pirámide, 1999, p. 23-52.

PASQUALI, L. Psicometria: **Teoria dos testes na Psicologia e na Educação**. Petrópolis: Editora Vozes, 2004.

PASQUALI, L.; PRIMI, R. Fundamentos da teoria da resposta ao item: TRI. **Avaliação Psicológica**, v. 2, n. 2, p. 99-110, 2003.

PASQUALI, L. **Técnicas de exame psicológico–TEP:** manual. São Paulo: Casa do Psicólogo, v. 23, 2001.

RIZOPOULOS, D. ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. **Journal of Statistical Software**, v. 17, n. 5, p. 1-25,

2006.

SARTES, L. M.; SOUZA-FORMIGONI, M. L. Avanços na Psicometria: da Teoria Clássica dos Testes à Teoria de Resposta ao Item. **Psicologia: Reflexão e Crítica**, v. 26, n. 2, p. 241-250, 2013.

SEGALL, D. O. **Computerized Adaptive Testing.** Encyclopedia of Social Measurement, Elsevier Inc. v. 1, p. 429-438, 2005.

STEIN, L. M. **TDE - Teste de Desempenho Escolar:** manual para aplicação e interpretação. São Paulo: Casa do Psicólogo, p. 1-17, 1994.

VENDRAMINI, C. M. DIAS, A. S. Teoria de Resposta ao Item na análise de uma prova de estatística em universitários. **Psico-USF**, v. 10, n. 2, p. 201-210, jul./dez. 2005.

WILLSE, J. T. CTT: **Classical test theory functions** (R package version 2.3.2). Disponível em: <https://CRAN.R-project.org/package=CTT>. Acesso em: 25 jun. 2018.

WISE, S. L.; KINGSBURY, G. G. Practical issues in developing and maintaining a computerized adaptive testing program. **Psicológica**, v. 21, n. 1, 2000.

YAN, D.; MAGIS, D. **Computerized adaptive and multi-stage testing with R**. Springer, 2016.

12 Exercícios

1) Após a leitura do capítulo, disserte sobre a importância de elaborar instrumentos considerando os princípios da TRI em pesquisas na área de Informática na Educação.

2) Explique quais as limitações que a TCT apresenta por considerar o escore a partir de uma amostra padronizada.

3) Esboce as etapas para elaboração do instrumento aplicada ao seu tema de pesquisa.

4) Um pesquisador aplicou um instrumento contendo 30 itens dicotômicos para avaliar a habilidade de raciocínio lógico em turmas do 1° ano do ensino médio. Os dados foram armazenados no formato csv em https://goo.gl/FZuEH1. Ajude o pesquisador a analisar seus dados usando a TRI. Assim, baixe os dados coletados no link acima e execute as etapas abaixo:

a) Calcule o coeficiente alfa de Cronbach dos 30 itens. O que o valor do coeficiente indica sobre a confiabilidade do instrumento?

b) Estime os parâmetros de discriminação, dificuldade e acerto ao acaso dos 30 itens. Há valores críticos? Quais? O que devemos fazer com os itens que apresentam valores críticos?

c) Após remover os itens que apresentam parâmetros com valores críticos, estime novamente os parâmetros com os itens atuais. Os valores são aceitáveis?

d) Plote as CCI e as funções de informação dos itens atuais.

e) Use os itens atuais para estimar a habilidade dos estudantes que responderam esses itens.

5) Descreva um algoritmo de seleção dos itens para um TAI e justifique a sua escolha para o critério de parada.

Sobre os autores

Ana Liz Souto Oliveira de Araújo

http://lattes.cnpq.br/7788932431434287

Doutoranda em Ciência da Computação na área de Educação e Computação da Universidade Federal de Campina Grande (UFCG). Mestre em Sistemas e Computação pela Universidade Federal do Rio Grande do Norte (UFRN). Professora adjunta do Departamento de Ciências Exatas da Universidade Federal da Paraíba (UFPB).

Jucelio Soares dos Santos

http://lattes.cnpq.br/4603605800436333

Mestre em Ciência da Computação, área de Educação em Computação pela Universidade Federal de Campina Grande (UFCG). Professor substituto do Centro de Ciências Exatas e Sociais Aplicadas da Universidade Estadual da Paraíba (UEPB).

Monilly Ramos Araujo Melo

http://lattes.cnpq.br/3029090915971102

Doutora em Psicologia, área de concentração - Psicologia Cognitiva, pela Universidade Federal de Pernambuco (UFPE). Professora Adjunta do Curso de Psicologia da Universidade Federal de Campina Grande (UFCG). Pesquisadora e Coordenadora do Laboratório de Neuropsicologia Cognitiva e Inovação Tecnológica - LabNEUROCIT (UFCG/Cnpq), Pesquisadora do Grupo de Estudos e Pesquisas em Psicologia Cognitiva e Cultura (Psicologia Cognitiva/UFPE/Cnpq) e do Grupo de Pesquisa em Educação e Psicometria/UFPB/Cnpq.

Wilkerson de Lucena Andrade

http://lattes.cnpq.br/3697205933296303

Doutor em Ciência da Computação pela Universidade Federal de Campina Grande (UFCG). Professor adjunto do Departamento de Sistemas e Computação da Universidade Federal de Campina Grande (UFCG). Orientador de mestrado e doutorado da Universidade Federal de Campina Grande (UFCG) na área de Educação em Computação.

Dalton Dario Serey Guerrero

http://lattes.cnpq.br/2050632960242405

Doutor em Engenharia Elétrica pela Universidade Federal de Campina Grande (UFCG). Professor adjunto do Departamento de Sistemas e Computação da Universidade Federal de Campina Grande (UFCG). Orientador de mestrado e doutorado da Universidade Federal de Campina Grande (UFCG) na área de Educação em Computação.

Jorge Cesar Abrantes de Figueiredo

http://lattes.cnpq.br/1424808046858622

Doutor em Engenharia Elétrica pela Universidade Federal da Paraíba (UFPB). Professor titular do Departamento de Sistemas e Computação da Universidade Federal de Campina Grande (UFCG). Orientador de mestrado e doutorado da Universidade Federal de Campina Grande (UFCG) na área de Educação em Computação. Diretor do Centro de Engenharia Elétrica e Informática da UFCG.