# Capítulo 5

# Mensuração e Testes

André Luís Alice Raabe (UNIVALI)
James Roberto Bombasar (UNIVALI)

***Objetivo do Capítulo***

Este capítulo tem o objetivo de apresentar aos pesquisadores de Informática na Educação as técnicas para a construção rigorosa de instrumentos de mensuração de construtos teóricos presentes em suas pesquisas com ênfase na construção de testes. Ao final da leitura deste capítulo, você deve ser capaz de:

- Conhecer as principais metodologias de construção de testes e entender as vantagens da utilização de cada uma delas.
- Conhecer os critérios de qualidade de testes e suas respectivas técnicas de aferição.
- Saber criar e selecionar testes para uso em pesquisas em Informática na Educação.
- Saber identificar os critérios mais importantes para a medida de qualidade de um teste, considerando as características do teste e da variável em estudo.


***Era uma vez...*** *um mestrando fez um software educacional para ajudar crianças com discalculia a aprimorarem suas habilidades de cálculo e superarem suas defasagens. Ele desenvolveu e aperfeiçoou o software com ajuda de especialistas e resultados de experimentação com usuários. Como parte de sua pesquisa ele decidiu que queria mensurar o quanto os estudantes com discalculia avançaram em suas habilidades com uso do software. Para isso elaborou um pré-teste e um pós-teste com questões a serem respondidas antes e depois do uso do software. No entanto, como estava próximo do prazo final do mestrado, ele não pode investir tempo em garantir a qualidade do instrumento de medida. Para sua surpresa, no exame de qualificação um dos revisores analisou cuidadosamente os instrumentos elaborados e afirmou que eles não eram adequados para mensurar a discalculia, e comprometeriam os resultados da pesquisa. Os instrumentos teriam que ser refeitos em um curtíssimo espaço de tempo. Um bonito trabalho quase teve um desfecho indesejável, por descuido com a mensuração do principal construto envolvido, a discalculia.*

# 1 Construtos e a cultura da mensuração através de testes

A palavra **construto** remete a um conhecimento definido de forma precisa e que representa um conceito teórico passível de ser estudado, isolado e inter-relacionado com outros conceitos. Quando estes são definidos de forma imprecisa ou ampla demais dificultam a sua mensuração. Portanto construtos que são muito abrangentes, como por exemplo a aprendizagem, tendem a ser difíceis de mensurar. Em contrapartida, os mais pontuais como: retenção, memória, motivação e engajamento serão mais fáceis de aferir.

A definição objetiva de um construto trata-se de uma delimitação que possibilita que ele seja operacionalizado através de intervenções, traduzido em instrumentos de medida e discutindo comparativamente com outros construtos que podem ser semelhantes ou diferentes. Nesta direção, a definição objetiva é um pré-requisito para que se possam construir instrumentos de medida que possuam qualidades desejáveis para se tornarem robustos e confiáveis que são a validade e a fidedignidade, vistos em detalhes adiante neste capítulo.

A cultura de mensuração através de testes e em especial testes educacionais baseia-se fortemente nos conhecimentos advindos da área de psicometria e tem forte influência da pesquisa em Psicologia e Educação norte americana. A pesquisa em Educação no Brasil, em geral, segue uma tradição mais qualitativa que não se preocupa em mensurar construtos de forma precisa e objetiva, pois tendem a abarcar a maior quantidade possível de aspectos relacionados a um determinado conceito ou fenômeno e não necessariamente mensurar aspectos pontuais e objetivos. É mais comum encontrarmos a preocupação com a construção de instrumentos de medida em pesquisas educacionais de orientação quantitativa ou mista.

Testes padronizados vêm sendo utilizados de forma crescente para avaliação de programas educacionais no Brasil. São exemplos as avaliações do IDEB e a Prova Brasil. Estes testes utilizam técnicas que possibilitam aferir a comparabilidade entre diferentes amostras em períodos distintos. Esta comparabilidade só é possível graças ao grau de confiança que existe no instrumento de medida, ou seja, o que o teste mensura é estável e a variação é decorrente dos indivíduos e seus diferentes níveis de habilidade.

Os Estados Unidos da América utilizam amplamente testes padronizados na avaliação de suas políticas educacionais, o que tem gerado muitas críticas devido ao fato dos testes não considerarem particularidades regionais e também por criarem a tendência de ensinar para o teste. A variedade e disponibilidade de instrumentos e testes padronizados nos Estados Unidos evidenciam o grau de difusão desta prática de mensuração.

# 2 A construção de testes

Muitos princípios podem ser utilizados para construção de testes, desde técnicas simples que são abordadas na formação inicial de professores e também são de conhecimento da maioria dos docentes, até técnicas mais robustas ligadas geralmente à preocupação de pesquisadores que necessitam demonstrar a validade e utilidade do teste construído. A

Teoria da Resposta ao Item que tem sido utilizada para construção de testes padronizados será abordada em profundidade em outro capítulo desta obra e por isso não será abordada neste capítulo.

O que as técnicas buscam garantir é que o teste construído seja uma operacionalização adequada do construto em questão. Por exemplo, se um determinado professor está buscando melhorar a habilidade de seus estudantes em resolver problemas de geometria, a construção de um teste para tal fim deve considerar uma série de passos que vão desde a escolha de quais problemas de geometria serão incluídos, o formato das questões e escalas de resposta, qual o grau de dificuldade desses problemas e ainda podem avançar no sentido de escolher problemas que tenham maior potencial de diferenciar os estudantes proficientes no assunto daqueles que não estão neste mesmo patamar.

## 2.1 Escores, Questões e Escalas de Resposta

A construção de testes e instrumentos de mensuração educacional é uma tarefa que demanda muitas escolhas. Desde a escolha da linguagem, o propósito do teste, o conteúdo do teste, o layout do teste, o tempo estimado de realização, a quantidade de questões, os tipos de questões, as escalas de resposta e assim por diante. Este é um tema bastante extenso, e não será possível aprofundá-lo neste capítulo. Apenas algumas noções fundamentais serão abordadas, para um maior aprofundamento sugerimos o capítulo 19 de Cohen et al. (2007).

Qualquer instrumento de mensuração educacional necessita gerar uma pontuação denominada escore. Este escore pode ser bruto, ou seja, o simples somatório de acertos, ou pode ser ponderado conforme a importância de determinadas questões. O valor final do escore deve representar em que grau o sujeito avaliado possui a qualidade avaliada, no entanto, a interpretação deste resultado depende se o valor obtido será comparado com outros sujeitos (ex: estar entre os 10% melhores) ou então com requisitos pré-definidos (ex: nota mínima de aprovação).

Uma decisão importante na construção de instrumentos voltados a coleta de dados para pesquisas é a escolha dos formatos de questões e escalas de respostas. Existem muitos tipos de questões que podem ter diferentes escalas de respostas. Em Fraenkel e Wallen (2007) são exemplificados diversos tipos de instrumentos e formatos de questões que podem ser utilizados em pesquisas educacionais. A classificação a seguir busca sintetizar as escolhas mais comuns.

Questões Objetivas: Questões onde a resposta pode apenas ser certa ou errada, não existindo meio termo. São exemplos as questões de múltipla escolha, verdadeiro e falso, associação de itens e respostas curtas (lacunas)

Questões de Escala ou Frequência: Questões onde a resposta é um valor em um intervalo ou escala. São exemplos questões em que o sujeito assinala uma valor em uma escala discreta (ex: escala Likert) ou contínua (ex: um valor do intervalo [0, 10]). Podem ainda representar a frequência de algum evento (ex: número de horas de estudo).

Questões Subjetivas: Questões que necessitam da avaliação subjetiva de uma pessoa com expertise no tema. São exemplos as questões que solicitam a resolução de problemas, e questões com respostas discursivas abertas. A subjetividade da correção cria um risco de viés para pesquisas, por isso recomenda-se o uso de critérios de correção bem definidos (usualmente chamados de rubricas de avaliação) e também a avaliação por mais de uma pessoa buscando aferir a concordância entre as avaliações e ajustá-las quando necessário.

Os testes podem combinar diferentes tipos de questões com escalas de respostas diferentes, no entanto isto deve ser feito com cuidado para evitar a sobrecarga cognitiva do respondente e minimizar a ocorrência de falhas de preenchimento.

## 2.2 Matriz de conteúdos

A formulação de uma matriz de conteúdos é uma das técnicas mais simples e úteis para garantir que um determinado teste contemple todos os aspectos do construto, permitindo ainda que este seja ponderado conforme os elementos constituintes que são considerados mais importantes na construção de uma escala numérica que represente este construto.

Consiste basicamente em listar os conteúdos que devem estar presentes em um determinado teste e em que quantidade. O exemplo a seguir representa um teste sobre conhecimentos de laços de repetição (loops), um construto do domínio da programação.

| Tipo de problema | Qtd | % |
|---|---|---|
| Número de repetições definido no enunciado | 1 | 8,33 |
| Número de repetições definido pelo usuário | 2 | 16,67 |
| Repetir até digitação de um código de fim | 2 | 16,67 |
| Uso de variável Somadora | 3 | 25,00 |
| Uso de mais de um contador | 2 | 16,67 |
| Com desvio dentro do loop | 2 | 16,67 |
| Total | 12 | 100 |

Tabela 1: Planejamento das questões de um teste educacional sobre laços de repetição

A matriz permite auxiliar o construtor do teste a planejar os conteúdos que serão abordados e em que quantidade. Garante também que nenhum aspecto do construto esteja sendo negligenciado na construção do teste. Esta informação dá maior evidência ao peso que cada conceito terá no escore final do estudante e será informação útil para aferir a validade do teste.

O uso da matriz de conteúdo pode ainda ser associado com o grau de dificuldade de cada questão, uma forma robusta de fazer esta associação é com uso de uma classificação de dificuldade de cada questão. A taxonomia de Bloom e associados tem sido amplamente utilizada para este fim e será abordada a seguir.

## 2.3 Uso da taxonomia de Bloom

A taxonomia de objetivos educacionais, Bloom et al. (1956), é o resultado de uma iniciativa liderada pelo pesquisador da Universidade de Chicago, Benjamin Bloom. O projeto iniciou em reunião da Associação Americana de Psicologia (APA), ocorrida em 1948, em que os presentes manifestaram a necessidade de construção de um quadro teórico de referência que facilitasse a comunicação entre examinadores educacionais.

Após ampla discussão, definiram que ela deveria ter a forma de uma classificação de objetivos educacionais para auxiliar a formulação de currículos, planos de aula e avaliações. A partir daí os proponentes reuniram-se anualmente com intuito de avaliar e aprimorar a classificação, envolvendo também professores e especialistas em Educação. Apenas em 1956, após ampla validação é que a taxonomia foi publicada contemplando os domínios Cognitivo, Afetivo e Psicomotor, no entanto a taxonomia de domínio cognitivo é a que se tornou mais amplamente conhecida e utilizada.

A taxonomia de Bloom, como é corriqueiramente denominada, é uma estrutura de organização hierárquica de objetivos educacionais. O domínio cognitivo se divide em seis categorias: conhecimento, compreensão, aplicação, análise, síntese e avaliação. Estas categorias são ordenadas da mais simples para a mais complexa. Além disso, a taxonomia é uma hierarquia cumulativa, onde uma categoria mais simples é pré-requisito para a próxima categoria mais complexa. A Figura 1 apresenta a taxonomia e a descrição de cada categoria.

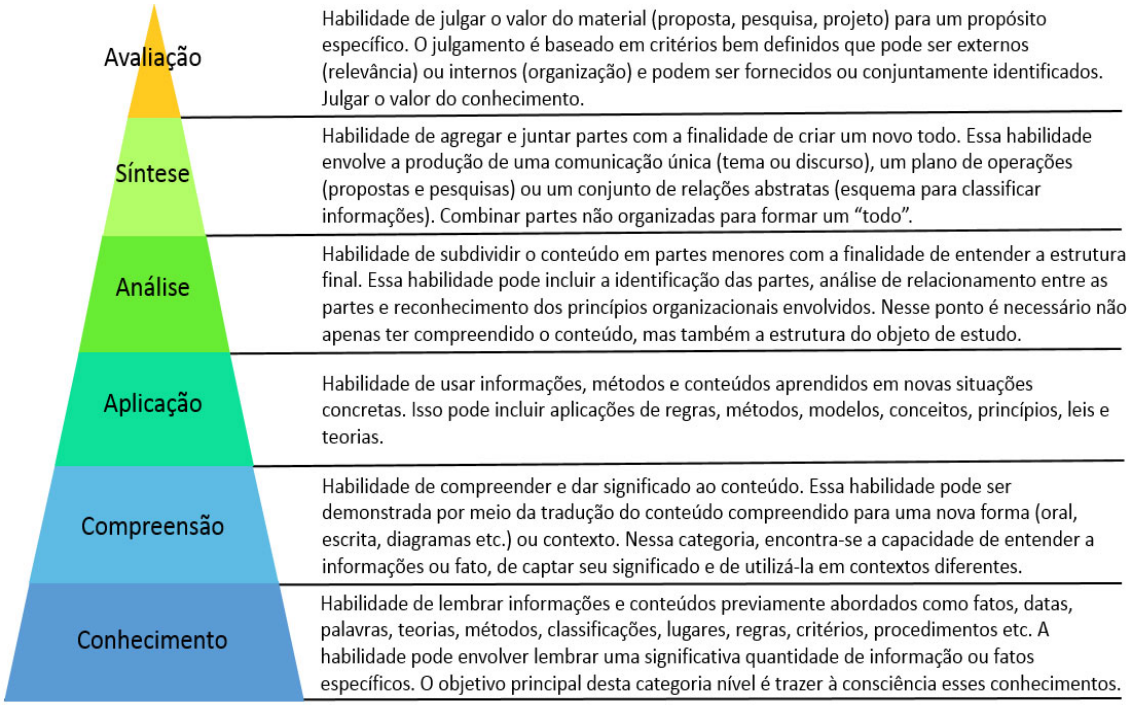



Figura 1: Taxonomia de Bloom

Quarenta anos depois de ter sido divulgada Anderson e Sosniaks (1994) realizaram um significativo trabalho de retrospectiva da utilização da taxonomia, o qual

motivou a rever os pressupostos teóricos da Taxonomia de Bloom uma vez que novos conceitos, recursos e teorias foram incorporados ao campo educacional. A partir deste trabalho uma visão revisada da Taxonomia foi publicada em Anderson et al. (2001), mas não será explorada neste capítulo por não estar amplamente consolidada como a versão original.

O benefício em utilizar a taxonomia de Bloom como referência para elaboração de testes é que que ela possibilita que se avalie o grau de demanda cognitiva necessária para responder cada questão de um teste. Por exemplo, questões que se referem a fatos e conceitos que precisam ser lembrados ficam no nível mais baixo da taxonomia, o nível conhecimento, e são, portanto, de menor demanda cognitiva. Já questões que necessitam que o estudante realize uma síntese a partir de informações providas possuem uma demanda mais alta e são, portanto, mais difíceis de serem realizadas.

Portanto a taxonomia de Bloom possibilita enriquecer a matriz de conteúdos incluindo a dificuldade cognitiva de cada questão, ou ainda motivando a elaboração de questões de determinada categoria da taxonomia a fim de equilibrar o teste. O Quadro 1 exemplifica o planejamento de um teste seguindo esta estratégia.

| Tipo de problema/ Categoria | Conhecimento | Compreensão | Aplicação | Análise | Síntese | Avaliação | total |
|---|---|---|---|---|---|---|---|
| Número de repetições definido no enunciado | | 1 | | | | | 1 |
| Número de repetições definido pelo usuário | | | | 1 | | | 1 |
| Repetir até digitação de um código de fim | | 1 | 1 | | | | 2 |
| Uso de variável Somadora | | | 1 | 1 | | 1 | 3 |
| Uso de mais de um contador | 1 | | | | 1 | | 2 |
| Com desvio dentro do loop | | | 1 | | | 1 | 2 |
| Total | 1 | 2 | 3 | 2 | 1 | 2 | 11 |

Quadro 1: Questões de um teste classificadas conforme a taxonomia de Bloom

Na tabela pode-se verificar que buscou-se equilibrar a quantidade de questões de cada categoria da taxonomia, sem perder com isso a distribuição original de quantidade de questões por conteúdo.

Uma vantagem adicional do uso da taxonomia está em prover a comparabilidade entre a dificuldade de diferentes testes. O que pode ser bastante útil em delineamentos de pesquisa que contêm pré-teste e pós teste, e também para a avaliação de formas paralelas de um teste a fim de aferir sua fidedignidade (visto em detalhes a seguir)

## 2.4 Design centrado em evidência

Evidence-Centered Design (ECD) é uma metodologia para a construção de avaliações educacionais criada por Robert J. Mislevy e colaboradores (MISLEVY; ALMOND; LUKAS, 2003). É uma metodologia muito útil para a mensuração de novos construtos e para desenvolvedores de testes com pouca ou nenhuma experiência. No ECD, o processo de concepção, desenvolvimento e aplicação de avaliações é dividido

em cinco camadas de atividades: i) análise do domínio; ii) modelagem do domínio; iii) framework conceitual de avaliação; iv) implementação da avaliação; v) entrega da avaliação.

A camada de análise do domínio está preocupada com a coleta de informações sobre o domínio que terão implicações diretas para a avaliação, incluindo conceitos, terminologias e formas representacionais utilizadas por pessoas que trabalham no domínio. Para o ECD, cada avaliação é uma amostra de Conhecimentos, Capacidades e Habilidades (CCH). Assim, é necessário identificar os CCH mais importantes para o domínio de interesse.

Um ponto bastante interessante do ECD é a camada de modelagem do domínio. Ela articula os CCH identificados na análise do domínio na forma "Se X, então Y porque Z", onde "X" representa a observação de um comportamento ou produto deste comportamento, "Y" representa a reivindicação de algum CCH e "Z" representa a garantia de que o comportamento ou produto observado demonstra a posse (ou não) do CCH. Por exemplo, uma pessoa capaz de realizar operações aritméticas com números decimais (X) demonstra boas habilidades matemáticas (Y) porque operações aritméticas com números decimais exploram construtos importantes no domínio da matemática (Z).

As últimas camadas do ECD estão focadas em questões de operacionalização da avaliação, tais como tipos de tarefas, modelos de mensuração e cálculo da pontuação.

## 3 Medidas de qualidade de testes

A qualidade de testes é uma investigação cujo desenho depende de vários aspectos, tais como o construto/ que o teste pretende medir, o número de dimensões deste, bem como a precisão que o pesquisador deseja conseguir. Isto significa que é pouco provável que a qualidade de diferentes testes possa ser aferida da mesma forma. Como consequência, não existe um passo a passo bem definido quando falamos em medida de qualidade de testes.

Por exemplo, se você traduzir um teste amplamente aplicado e validado para a língua nativa dos estudantes que irão realizar o teste, você precisará verificar se as questões da versão traduzida que você criou serão interpretadas da mesma forma que as questões do teste original. Por outro exemplo, se você deseja utilizar somente parte das questões de um teste, você precisará verificar se as questões selecionadas conseguem refletir por completo (e somente) o construto em estudo. Você pode ainda se deparar com situações nas quais é necessária a criação de um novo teste, pois o construto em estudo é complexo ou ainda pouco explorado.

Seja qual for o cenário de pesquisa, existem critérios que irão auxiliá-lo tanto na criação de testes de qualidade como na aferição da qualidade de testes já existentes. Os critérios de qualidade de testes podem ser classificados em duas grandes perspectivas: critérios de validade e de fidedignidade.

## 3.1 Validade e fidedignidade

Sob a perspectiva da validade, a principal preocupação de um pesquisador em relação à qualidade do teste a ser aplicado pode ser resumida na questão "o teste está medindo o que ele se propõe a medir?". Um teste de boa qualidade deve refletir bem o domínio do construto em estudo e ser capaz de identificar quem são as pessoas que apresentam melhores capacidades no domínio. Por exemplo, um pesquisador que cria um teste para mensurar o construto "habilidades de cálculo" deve ser capaz de responder bem a perguntas como: Quais são as habilidades de cálculo? Todas as habilidades de cálculo são exploradas pelo teste? Os alunos que obtém bons escores no teste são efetivamente os alunos com melhores habilidades de cálculo?

Sob a perspectiva da fidedignidade, por sua vez, a preocupação do pesquisador é se o teste mede com precisão. Ou seja, o pesquisador que criou o teste para mensurar a variável "habilidades de cálculo" deve ser capaz de responder bem a perguntas como: se teste for aplicado novamente com os mesmos participantes e diferentes avaliadores, os escores serão os mesmos? As questões que medem as mesmas capacidades apresentam escores semelhantes? Para que você possa melhor compreender a diferença entre validade e fidedignidade, apresentamos na Figura 2 uma representação visual clássica das duas perspectivas.

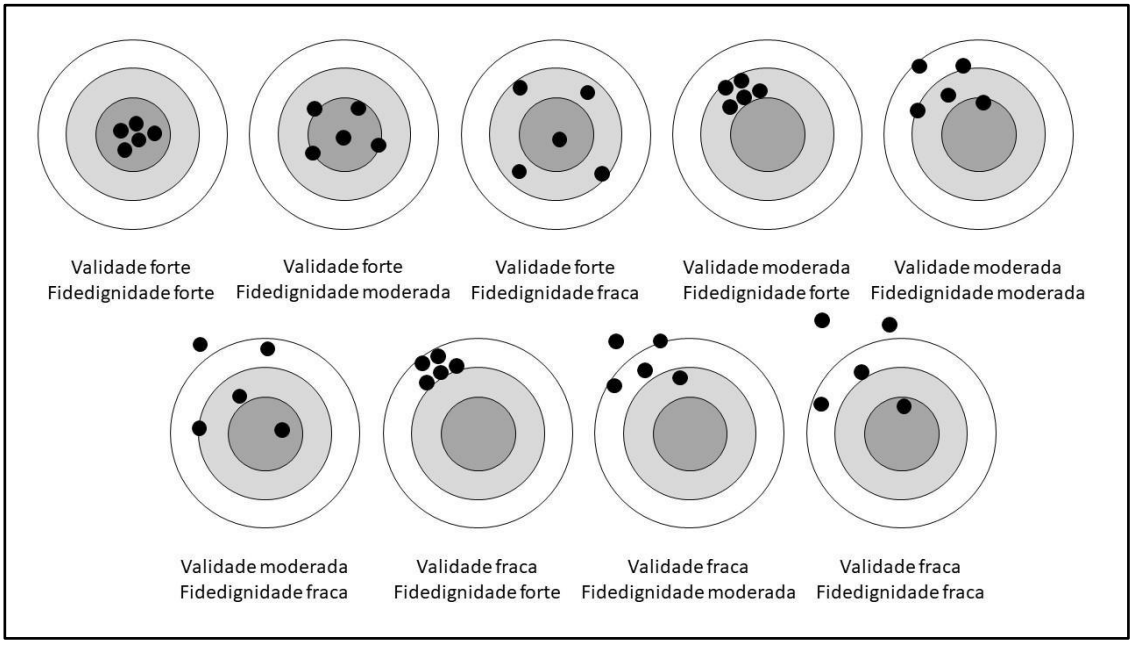



Figura 2: Representação visual de validade e fidedignidade

Considere que cada ponto é a média dos escores obtidos em uma aplicação do teste, e que a parte mais escura do alvo representa o construto que o teste pretende avaliar. Quanto mais dispersas forem as médias dos escores, menor será a fidedignidade do teste, e quanto mais distantes do alvo forem as médias, menor será a validade do teste. Para que um pesquisador tenha certeza de que todas as médias de escores estão concentradas no centro do alvo os critérios de validade e fidedignidade que devem ser considerados.

## 3.2 Tipos de validade

Já sabemos que a validade de um teste representa o grau em que ele mede o que se propõe a medir. Para um teste ser considerado totalmente válido, três critérios devem ser considerados. Em primeiro lugar, as capacidades que o pesquisador deseja medir devem estar bem traduzidas no teste. Ou seja, todos os tópicos importantes do domínio devem estar contemplados e a linguagem deve ser adequada. Em segundo lugar, é necessário que os escores do teste (variáveis observáveis) sejam evidências válidas dos construtos teóricos (variáveis latentes). O teste deve realmente medir a capacidade pretendida.

Por fim, um teste válido deve ser capaz de dizer quem são os estudantes com melhores capacidades no domínio. Como explicar, por exemplo, que os estudantes com os escores mais baixos em um teste que mede "habilidades de cálculo" são os mesmos estudantes que conseguiram excelentes notas na disciplina de matemática? Assim, a validade total de um teste pode ser obtida com a soma de três tipos de validade: validade de tradução, validade de construto e validade de critério.

### 3.2.1 Validade de tradução

Ao analisar a validade de um teste, a primeira preocupação que você deve ter é com o que está sendo apresentado para as pessoas que irão realizar o teste. Um teste com **tradução válida** deve apresentar conteúdo e aparência válidos. A **validade de conteúdo** está preocupada em saber se conteúdo do teste é uma amostra adequada do domínio. Ou seja, se os itens do teste são apropriados e relevantes. Por exemplo, um teste de aritmética deve explorar as operações de soma, subtração, multiplicação e divisão. A **validade aparente ou de face** diz respeito à linguagem, a forma com que o conteúdo está sendo apresentado. Por exemplo, um teste planejado para adultos e depois utilizado em crianças não possui validade aparente.

A elaboração de conteúdo de testes que exploram construtos bem definidos na literatura é muitas vezes uma tarefa simples. É bem provável que ninguém irá contestar que a aritmética envolve operações de soma, subtração, multiplicação e divisão. Assim, para criar um teste de aritmética, basta o pesquisador criar algumas questões de cada tipo de operação, inspiradas ou extraídas de livros ou outros testes já utilizados. A matriz de conteúdos apresentada na seção 2.1 é uma técnica que auxilia na realização desta tarefa.

O problema é que nem sempre os construtos que desejamos explorar estão bem definidos ou possuem uma definição simples como a da aritmética. Por exemplo, qual seriam as definições e as dimensões das capacidades de abstração, decomposição e paralelização, que se encontram dentro de uma capacidade mais ampla, denominada pensamento computacional? Será que elas possuem aspectos em comum? Como avaliá-las separadamente? São perguntas que provavelmente surgirão na mente do pesquisador.

Uma estratégia de validação de conteúdo bastante utilizada é a técnica Delphi, na qual um corpo de juízes formado por experts (especialistas, mestres, doutores) com vasta experiência no domínio realizam a avaliação do teste. A técnica baseia-se no

princípio de que as decisões de um grupo estruturado de experts são mais precisas se comparadas às provenientes de grupos não estruturados ou individuais.

Por sua vez, a validação de aparência de um teste pode ser realizada por meio de uma aplicação piloto do teste a um grupo reduzido de indivíduos do público alvo. Após a aplicação do teste, faça perguntas aos indivíduos, por exemplo: você teve dificuldade em compreender algum enunciado? Ficou indeciso entre duas ou mais alternativas de alguma questão? Se sim, por quê?

### 3.2.2 Validade de construto

A validade de construto ou de conceito é um tipo de validade que dá significado às pontuações dos testes. Ou seja, está preocupada em saber se os escores de um teste podem ser utilizados para extrair inferências corretas sobre a conceito que o teste pretende medir. A validade de construto vai mais além que a validade de conteúdo. Ela deseja entender melhor as questões cognitivas e psicológicas que estão sendo medidas pelo teste. Imagine uma situação em que duas questões semelhantes A e B produzem escores com baixa correlação. Qual a explicação subjacente para tal? Será que os examinandos estão utilizando diferentes processos para responder cada uma das questões? Ambos os processos são válidos e integram o conceito avaliado?

A validade de construto é tipicamente subdividida em **validade convergente** e **validade discriminante**. A primeira busca testar a hipótese de que o teste realmente mede o que ele se destina a medir por meio da correlação do seu escore com o escore de outro instrumento que mede o mesmo construto ou variáveis diretamente relacionadas a ele. A validade discriminante, por sua vez, verifica se a medida em questão não está relacionada indevidamente com indicadores de outros construtos. Ou seja, com variáveis das quais o teste deveria diferir. Ambas as validades podem ser aferidas com o uso da matriz multi métodos (Multi-Trait Multi-Method - MTMM) descrita em Trochim e Donnelly (2008).

Dentro da validade de construto, existe ainda o critério da **validade fatorial**, desenvolvido para identificar traços psicológicos comuns em testes. O principal objetivo da análise fatorial é reduzir o número de dimensões necessárias para se descrever dados derivados de um grande número de medidas. Ela permite comprovar se o teste realmente está medindo o que pretende medir, pois proporciona: a) nitidez sobre os aspectos subjacentes do teste; b) auxílio na definição das dimensões do teste, e; c) visões de como essas dimensões estão correlacionadas entre si. A análise fatorial pode ser exploratória (confirmatory factor analysis), quando se tem hipóteses a priori acerca da estrutura subjacente do teste, ou confirmatória (exploratory factor analysis), quando as hipóteses ainda não estão definidas (COHEN; MANION; MORRISON, 2007).

### 3.2.3 Validade de critério

Validade de critério é a capacidade que um teste possui de distinguir sujeitos que se comportam de maneira diferente em relação aos construtos em estudo. O desempenho

do sujeito torna-se, assim, o critério contra o qual a medida obtida pelo teste é avaliada. Por exemplo, um estudante que obtém boas notas em um teste de habilidades matemáticas deve conseguir boas notas na disciplina de matemática. É óbvio que, para aferir a validade de critério do teste de habilidades matemáticas, os escores do teste não podem ser utilizados para gerar as notas da disciplina (critério).

Costuma-se subdividir a validade de critério em dois tipos: **validade preditiva** e **validade concorrente**. No primeiro caso, o teste funciona como um preditor dos escores do critério, ou seja, a medida do teste prediz o desempenho concreto do sujeito. Por exemplo, um teste de algoritmos que prediz o desempenho de um acadêmico de Ciência da Computação na disciplina de algoritmos ao final do semestre letivo é um teste com boa validade preditiva. No caso da validade concorrente, o resultado do teste é comparado simultaneamente com outra medida que represente o mesmo construto, onde a alta correlação positiva indica a validade do teste.

A validade de critério pode ser aferida por meio da análise do coeficiente de correlação entre os escores do teste e do critério. O valor do coeficiente pode variar de -1.00 a +1.00. Um valor positivo indica que ambos os escores tendem a aumentar ou diminuir em conjunto, e um valor negativo indica que um dos escores tende a aumentar à medida que o outro diminui. O valor 0.00 indica que não há correlação entre os escores. É importante testar o nível de significância do coeficiente. Em geral, um valor $\alpha$ (alfa) de 0,05 é adequado.

## 3.3 Tipos de fidedignidade

A fidedignidade de um teste representa o seu grau de precisão (exatidão). Basicamente, a perspectiva que avalia a precisão de um teste está preocupada em saber o quanto os seus escores são consistentes (livres de erros aleatórios), sem se preocupar com o que o teste está medindo (validade). Basicamente, um teste fidedigno deve ser capaz de produzir os mesmos escores se aplicado sucessivamente sobre o mesmo sujeito.

É importante saber que alguns fatores relativos ao instrumento e ao sujeito podem afetar a fidedignidade de um teste em diferentes graus. Testes que possuem uma grande quantidade de itens homogêneos e de média dificuldade estão mais propensos a apresentar uma boa fidedignidade. Em relação ao sujeito, a motivação e compreensão dos enunciados são fatores que também contribuem para uma boa fidedignidade. O mesmo teste, quando aplicado a diferentes sujeitos, terá provavelmente coeficientes de fidedignidade diferentes. Assim, fatores relativos ao sujeito precisam ser considerados na interpretação da fidedignidade.

A fidedignidade de um teste é expressa pelo coeficiente de correlação entre dois ou mais escores do teste. Os escores podem ser obtidos por meio de métodos de mensuração repetida (teste-reteste, formas paralelas e interavaliadores) e não-repetida (consistência interna). A seleção do método adequado depende das características do teste e da investigação, conforme será explicado a seguir.

### 3.3.1 Teste-reteste (estabilidade)

O coeficiente de fidedignidade baseado no método teste-reteste é uma medida de estabilidade, pois se relaciona com a constância no tempo. É obtido por meio dos escores dos mesmos indivíduos em duas ocasiões distintas. A principal dificuldade do método é estabelecer um intervalo entre teste e reteste não muito curto, a ponto de os indivíduos lembrarem as respostas, e nem muito longo, a ponto de as respostas dos indivíduos serem afetadas por novos fatores pessoais e de meio ambiente.

Os escores podem ser afetados pelo comportamento dos indivíduos. Por exemplo, eles podem reagir de forma negativa contra ter que repetir um teste muito longo. Não é um bom método para avaliar testes de conhecimento quando os escores estão muito suscetíveis a vieses.

### 3.3.2 Formas paralelas (equivalência)

No método das formas paralelas ou das metades (*Split half*), duas formas equivalentes (bem parecidas) do teste quanto ao número de questões, conteúdo e tempo são aplicadas aos sujeitos. Os escores das duas formas são então utilizados para calcular o coeficiente de fidedignidade.

Uma vantagem do método das metades é que se as formas do teste forem curtas, elas podem ser aplicadas de forma sequencial. Caso elas sejam aplicadas em ocasiões distintas, o coeficiente irá combinar dois tipos de fidedignidade. A primeiro relativo à consistência dos itens e o segundo relativo à estabilidade. A principal dificuldade do método é a suposição de que as duas formas do teste são realmente equivalentes. Dependendo das características do teste, pode ser difícil dividir os seus itens de forma a atingir a equivalência adequada. O uso da taxonomia de Bloom, conforme apresentado na seção 3.2, pode auxiliar na construção de testes equivalentes.

### 3.3.3 Interavaliadores

Mensura a variação não-sistemática ocorrida em função de quem corrige o teste, geralmente é utilizada para reduzir a subjetividade da avaliação onde estão envolvidos diferentes avaliadores. A informação da fidedignidade interavaliadores é particularmente importante quando a correção do teste depende do julgamento de quem o corrige. O coeficiente é obtido pela correlação entre os escores dados pelos avaliadores. Para que os avaliadores não exerçam influência um sobre o outro, eles devem trabalhar de forma independente. Quando são obtidos escores de correlação baixos, recomenda-se a discussão e aprimoramento dos critérios de avaliação entre os avaliadores e a realização de novas rodadas de aprimoramento. A seção de artigos exemplos deste capítulo inclui um estudo que apresenta um caso de avaliação de fidedignidade interavaliadores.

### 3.3.4 Consistência interna

A consistência interna verifica a consistência entre os vários itens que compõem um teste com base na correlação entre eles. O método verifica a extensão com que todos os itens mensuram o mesmo construto. Um teste apresenta boa consistência interna ou homogeneidade quando todas as suas subpartes mensuram o mesmo construto.

Os itens de um teste devem se correlacionar ou ser complementares uns com os outros. Se um teste é formado por diferentes domínios que mensuram diferentes características, a consistência interna deve ser avaliada separadamente para cada domínio. O Alfa de Cronbach é medida estatística bastante utilizada para a investigação da consistência interna de testes.

A Figura 3 apresenta um exemplo de cálculo do Alfa de Combrach para análise da consistência interna de um teste composto de 7 questões (itens) e aplicado com 10 participantes (P). As questões/itens receberam um escore de 0 a 7. Portanto a nota máxima do teste seria 49 pontos. Inicialmente foram calculadas as variâncias dos escores de cada item e a variância do escore total. Em seguida, os valores das variâncias e do número de itens foram aplicados na fórmula. Verifica-se que o teste possui uma boa consistência interna, pois o valor do coeficiente alfa é muito próximo de 1.

| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 | Total |
|---|---|---|---|---|---|---|---|---|
| P 1 | 2 | 3 | 2 | 3 | 2 | 3 | 2 | 17 |
| P 2 | 7 | 7 | 5 | 7 | 7 | 7 | 7 | 47 |
| P 3 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 6 |
| P 4 | 5 | 4 | 6 | 5 | 4 | 5 | 6 | 35 |
| P 5 | 2 | 3 | 2 | 1 | 2 | 3 | 4 | 17 |
| P 6 | 3 | 2 | 3 | 2 | 2 | 3 | 2 | 17 |
| P 7 | 7 | 7 | 7 | 7 | 7 | 7 | 5 | 47 |
| P 8 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| P 9 | 5 | 4 | 5 | 6 | 5 | 4 | 6 | 35 |
| P 10 | 1 | 2 | 3 | 4 | 2 | 3 | 2 | 17 |
| Variância | **5,81** | **4,24** | **4,05** | **6,04** | **4,81** | **4,01** | **4,64** | **214,24** |

| | |
|---|---|
| Número de itens (k) | **7** |
| Soma das variâncias dos itens (Vi) | **33,6** |
| Variância total do teste (Vt) | **214,24** |
| Valor alfa (a) | **0,983694** |

$$\alpha = \left(\frac{K}{K-1}\right)\left(1-\frac{\sum V_i}{V_T}\right)$$

Figura 3: Exemplo de cálculo do Alfa de Cronbach

Conforme (Fraenkel e Wallen, 2009) o alpha de Kronbach é mais geral que a medida de fidedignidade K de Kuder-Richardson. Ele pode ser aplicado em qualquer tipo de escala numérica para itens de um teste. No caso de questões de escala binária (0- errado, 1- certo) ele também pode ser usado mas torna-se equivalente a medida de Kuder-Richardson. Conceitualmente o alpha de Kronbach representa a média entre todas as possibilidades de divisão ao meio (Split half) que poderiam ser feitas no teste (TROCHIM e DONNELLY, 2008).

Neste ponto, é importante destacarmos a sutil diferença entre a validade fatorial (estudada na Seção 3.2.2) e a consistência interna. Enquanto a primeira está preocupada com a estrutura subjacente do teste, a segunda verifica a consistência entre os itens que medem o mesmo conceito.

## 4 Testes padronizados

A construção e validação de testes é um tipo de investigação que normalmente consome boa parte do esforço de uma pesquisa. Isto pode representar um enorme problema em pesquisas de mestrado e doutorado, nas quais há um prazo determinado para a conclusão das atividades. Por sorte, testes para mensurar diferentes conceitos já foram criados e utilizados por diversos pesquisadores no passado.

No entanto, você deve ter alguns cuidados ao selecionar testes padronizados para utilização em sua pesquisa. O primeiro deles é verificar se o teste é aderente, ou seja, se a linguagem do teste é adequada para o seu público alvo e se os itens do teste refletem bem o domínio da sua pesquisa. Encontrar um teste cujos itens sejam uma amostra adequada de um domínio pode não ser uma tarefa fácil, especialmente quando o domínio possui muitas dimensões.

Neste ponto, você pode estar se perguntando: “e se eu reunir itens de diferentes testes padronizados para representar todas as dimensões do domínio?”. Neste caso, é natural que seja necessário conhecer a estrutura subjacente originada pela reunião dos itens, ou seja, o construto precisará ser validado. Outro cuidado que você deve ter é o de verificar se o teste é preciso. Caso não existam informações sobre a fidedignidade do teste, uma boa ideia é considerar a aplicação dos métodos da Seção 3.3.

Atualmente, uma lista abrangente de testes pode ser encontrada na base de dados Education Resources Information Center (ERIC) (https://eric.ed.gov/). Para tal, é necessário realizar uma consulta na página inicial e configurar o filtro *Publication Type* com o valor *Tests/Questionnaires* na página de resultados. É importante lembrar que os testes em inglês precisam ser validados após serem traduzidos para português. A seção de artigos exemplos deste capítulo inclui um estudo que descreve a aplicação da técnica Delphi (corpo de juízes) para a validação do conteúdo de um questionário traduzido.

## 5 Exemplo ilustrativo

No ano de 2015, um pesquisador conhecido como Beto iniciou o seu mestrado na área da Informática na Educação. Ele estava bastante otimista, pois iria pesquisar sobre o desenvolvimento do Pensamento Computacional (PC) na Educação Básica (EB), um tema em expansão e repleto de lacunas a serem preenchidas.

No início do curso, Beto realizou uma pesquisa exploratória em diversas bases de dados para adquirir familiaridade com o tema PC. A pesquisa explorou definições e buscou identificar estratégias que estavam sendo utilizadas para o desenvolvimento do PC com diferentes tipos de públicos. Beto identificou uma importante lacuna. Nenhuma

das pesquisas, até então, explorava a compreensão pelos alunos da EB sobre o poder e os limites da computação, uma importante característica do PC.

Muito animado, Beto realizou uma nova pesquisa exploratória na tentativa de descobrir alguma estratégia que permitisse explorar os temas Teoria da Computação e Computabilidade na EB. Foi daí que surgiu a ideia de construir um jogo de lógica inspirado nos autômatos e na máquina de Turing. Tanto a pesquisa sobre o PC como a construção do jogo renderam publicações científicas para Beto. Tudo estava indo bem, até que no segundo ano do mestrado, era chegada a hora da coleta de dados. Beto precisava de um instrumento para mensurar a compreensão pelos alunos da EB sobre os limites da computação. Foi aí que os problemas começaram a surgir.

A pesquisa era a primeira a explorar noções de computabilidade na EB, não existiam testes padronizados. Beto passou muitos dias traduzindo o domínio de pesquisa em um teste com 20 questões. As questões exploravam noções de computabilidade utilizando exemplos de situações encontradas no dia a dia dos alunos. Após várias rodadas de aplicação da técnica Delphi (na seção de artigos exemplos, é apresentado um artigo que descreve a aplicação desta técnica), o conteúdo foi validado.

Com os dados da aplicação piloto do teste, Beto utilizou a técnica da análise fatorial confirmatória (na seção de leituras recomentadas, é apresentado um livro sobre análise fatorial) para validar a sua hipótese sobre a estrutura subjacente do teste. O resultado foi desanimador. Em seguida, Beto utilizou a técnica da análise fatorial exploratória para compreender a estrutura do teste. No entanto, após várias tentativas, não foi possível estabelecer nenhuma estrutura com correlações aceitáveis.

O prazo para término do mestrado estava chegando. Após analisar os dados por vários dias, Beto percebeu que o teste poderia estar indicando indevidamente outros construtos. O problema não estava nas questões em si, mas no domínio que não estava bem delimitado. Beto decidiu pormenorizar sua hipótese de pesquisa. No lugar de "noções de computabilidade" ela passaria a investigar "a compreensão sobre modelos de computação".

Com base no design centrado em evidência, Beto construiu um novo teste em que os alunos deveriam responder às questões analisando diagramas de transições de autômatos e máquinas de Turing. O conteúdo do teste foi validado por um corpo de juízes. Em seguida, Beto aplicou o teste com um grupo de alunos da EB com o objetivo de validar a aparência do teste. Nenhum dos alunos relatou dificuldade em compreender e realizar o teste.

Finalmente, a poucos meses do término do mestrado, Beto coletou os dados de pesquisa. Após validar os fatores do teste, Beto utilizou o Alfa de Cronbach (na seção de leituras recomendadas, é apresentado um livro sobre métodos estatísticos) para validar também a consistência interna do teste. O coeficiente obtido foi satisfatório. Durante a banca de defesa, um dos avaliadores perguntou porque não foram utilizadas notas dos alunos para medir a capacidade do teste de distinguir sujeitos (validade de critério). Beto argumentou que não haveria como estabelecer uma relação nítida entre o domínio explorado na pesquisa e o domínio das disciplinas tradicionais.

Após um grande sufoco, a dissertação de mestrado foi aprovada e Beto levou uma grande lição para suas futuras pesquisas: o popular ditado “Menos é mais” também se aplica ao domínio de uma pesquisa.

## 6 Resumo

Neste capítulo foram abordados os principais conceitos relacionados com a mensuração de construtos teóricos através de testes. Para facilitar a memorização dos conceitos, criamos o mapa mental apresentado na Figura 4. A ideia é que neste ponto da leitura você seja capaz de: i) conhecer as principais metodologias de **construção de testes** e suas respectivas potencialidades; ii) conhecer as **medidas de qualidade de testes** e suas respectivas técnicas de aferição; iii) conhecer mecanismos e estratégias para busca e seleção de **testes padronizados**.


Figura 4: Mapa mental da mensuração através de testes

## 7 Leituras recomendadas

A seguir são sugeridas leituras para o aprofundamento no tema abordado neste capítulo. Você pode selecionar os textos de acordo com as suas necessidades de pesquisa. Por exemplo, se você precisa construir um novo teste, poderá priorizar as sugestões que

abordam metodologias para a construção de testes. Se você pretende aplicar um teste composto de itens retirados de outros testes, poderá selecionar apenas as sugestões que se aprofundam nos métodos para aferição da validade e fidedignidade.

- **Psicometria: Teoria dos Testes na Psicologia e na Educação** (PASQUALI, 2013). Este livro se aprofunda em diversos conceitos estudados ao longo deste capítulo. É uma leitura que pode auxiliar pesquisadores tanto na construção de testes como na medida de qualidade destes.
- **Evidence-Centered Assessment Design: Layers, Structures, and Terminology** (MISLEVY; RICONSCENTE, 2005). Neste livro você encontrará uma descrição detalhada da metodologia de construção de testes ECD, estudada na Seção 2.4. É uma leitura muito útil para pesquisadores que precisam construir testes para a mensuração de construtos complexos.
- **An Easy Guide to Factor Analysis** (KLINE, 1993). A validade fatorial, estudada na Seção 3.2.2, é um critério de qualidade de testes cuja aferição envolve métodos que podem ser pouco familiares aos pesquisadores da área da Informática na Educação. Neste livro, o autor explica a análise fatorial de forma clara e simples, incluindo um passo a passo. É uma leitura essencial para pesquisadores que desejam explorar com profundidade a validade de construto de seus testes.
- **Robust Correlation: Theory and Applications** (SHEVLYAKOV; OJA, 2016). Grande parte das medidas de qualidades de teste estudadas na Seção 3 são baseadas em coeficientes de correlação. Assim, este livro é uma leitura indicada para pesquisadores iniciantes que ainda possuem pouca afinidade com métodos estatísticos. O livro aborda os conceitos clássicos e os tipos de correlação através de exemplos com dados simulados.
- **How to design and evaluate research in education (**FRAENKEL e WALLEN, 2009**).** É uma obra que cobre muitos aspectos da pesquisa educacional e que é muito rica em exemplos de tipos de questões e escalas para instrumentos de mensuração.

## 8 Artigos exemplos

Nesta seção apresenta-se uma lista de artigos que detalham a aplicação dos métodos de validação de testes estudados neste capítulo. Pode parecer estranho que alguns artigos sejam das áreas da saúde e psicologia. Porém, é nestas áreas que a validação de instrumentos tem sido amplamente discutida nos últimos anos.

- **Desenvolvimento e validação de conteúdo da nova versão de um instrumento para classificação de pacientes** (PERROCA, 2013). Este artigo apresenta em detalhes a aplicação da técnica Delphi (corpo de juízes) para a validação de conteúdo (Seção 3.2.1).
- **Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test** (González, González e Fernández, 2017).

O artigo apresenta os resultados da avaliação de Validade de Critério de um teste construído para mensurar o Pensamento Computacional. (Seção 3.2.3).

- **Validez de Constructo y Confiabilidad del Cuestionario de Creencias Epistemológicas sobre la Matemática en Alumnos de Secundaria Básica** (ESCOBAR; MIER; CARDOSO, 2015). Este artigo apresenta um estudo sobre a estrutura subjacente de um questionário. É um exemplo detalhado de aplicação da análise fatorial para a validação de construto (Seção 3.2.2).
- **Estudo de fidedignidade do avaliador em provas de compreensão leitora e oral** (LUCIO; et al., 2016). Este artigo é um exemplo detalhado de avaliação de fidedignidade interavaliadores (Seção 3.3.3).
- **Um Instrumento para Diagnóstico do Pensamento Computacional.** (RAABE et al., 2017). Este artigo ilustra diversas das técnicas explicadas neste capítulo em especial a matriz de conteúdos (seção 2.1) e cálculo da fidedignidade (seção 3.3.4).

## 9 Checklist

Resumidamente, para coletar dados de pesquisa em Informática na Educação utilizando testes, você terá que realizar as seguintes atividades, também ilustradas na Figura 5.

- Adotar um teste válido e aceito e coletar os dados de pesquisa, ou;
- Construir um teste adotando uma metodologia de construção de testes;
- Aperfeiçoar o teste até que seu conteúdo seja considerado válido por uma banca de experts e a sua aparência seja considerada válida pelos indivíduos que realizam o teste;
- Coletar dados do teste para validações intra-teste (fatorial e fidedignidade);
- Coletar dados de critério para validações inter-teste (validades convergente, discriminante, preditiva e concorrente);
- Realizar os procedimentos de validação;
- Coletar os dados de pesquisa (exceto quando os dados de pesquisa forem os dados utilizados para as validações).

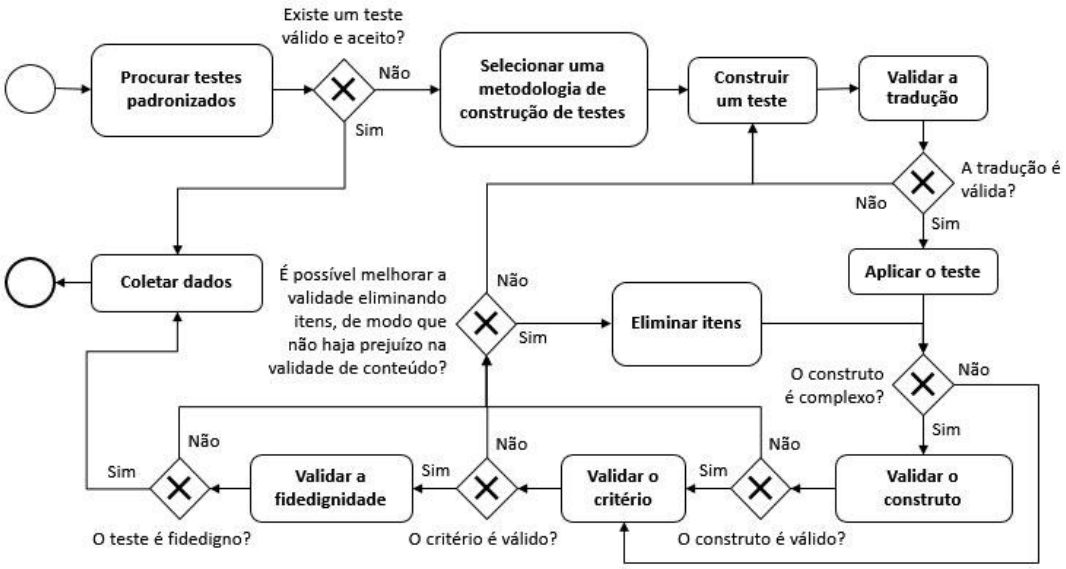



Figura 5: Fluxograma de atividades para coleta de dados de pesquisa utilizando testes

# 10 Referências


ANDERSON, LORIN; SOSNIAK, LAUREN A. **Bloom's Taxonomy: A Forty-Year Retrospective**. Ninety-Third Yearbook of the National Society for the Study of Education, University of Chicago Press, 1994.

ANDERSON, LORIN; KRATHWOHL, DAVID R. et al. **A Taxonomy for Learning, Teaching, and Assessing - A Revision of Bloom's Taxonomy of Educational Objectives**. Addison Wesley Longman, Inc.2001.

BLOOM, B. S.; ENGELHART, M. D.; FURST, E. J.; HILL, W. H.; KRATHWOHL, D. R. **Taxonomy of educational objectives: The classification of educational goals**. Handbook I: Cognitive domain. New York: David McKay Company, 1956.

COHEN, L.; MANION, L.; MORRISON, K. **Research methods in education**. 6. ed. London: Routledge, 2007.

VIZCAINO ESCOBAR, ANNIA ESTHER; MANZANO MIER, MAYRA; CASAS CARDOSO, GLADIS. **Validez de Constructo y Confiabilidad del Cuestionario de Creencias Epistemológicas sobre la Matemática en Alumnos de Secundaria Básica**. Rev. colomb. psicol., Bogotá , v. 24, n. 2, p. 301-316, jul. 2015

FERRAZ, Ana Paula do Carmo Marcheti; BELHOT, Renato Vairo. **Taxonomia de Bloom: revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais**. Gest. Prod., São Carlos , v. 17, n. 2, p. 421-431, 2010 .

FRAENKEL, J. R.; WALLEN, N. E. **How to design and evaluate research in education**. 7. ed. New York: McGraw-Hill, 2009.

GALL, M. D.; GALL, J. P.; BORG, W. R. **Educational research: An introduction**. 7. ed. Boston: Pearson, 2003.

GONZÁLEZ, MARCOS ROMÁN; GONZÁLEZ, JUAN CARLOS; FERNÁNDEZ, CARMEN JIMÉNEZ. **Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test**, In Computers in Human Behavior, Volume 72, 2017

MISLEVY, R. J.; ALMOND, R. G.; LUKAS, J. F. **A brief introduction to evidence-centered design**. Princeton, NJ: Educational Testing Service, 2003.

PERROCA, MARCI G. **Desenvolvimento e validação de conteúdo da nova versão de um instrumento para classificação de pacientes**. Rev. Latino-Am. Enfermagem, Ribeirão Preto , v. 19, n. 1, p. 58-66, Feb. 2011 .

RAABE, ANDRÉ; COUTO, NATÁLIA; GONÇALVES; FILIPE. **Um Instrumento para Diagnóstico do Pensamento Computacional**. Congresso Brasileiro de Informática na Educação 2017, Anais Workshop de Ensino em Pensamento Computacional Algoritmos e Programação, Recife, 2017.

TROCHIM, W. M. K.; DONNELLY, J. P. **The Research methods knowledge base**. 3. ed. Mason, Ohio: Atomic Dog/Cengage Learning, 2008.

## 11 Exercícios

Nesta seção apresentamos exercícios para ajudar você a avaliar a sua aprendizagem. As respostas comentadas aos exercícios são fornecidas ao final da seção.

**1.** Um pesquisador deseja testar a hipótese de que o uso de um determinado jogo de programar melhora o Pensamento Computacional (PC) dos alunos do Ensino Médio. No seu entender, quais as dificuldades que o pesquisador encontrará para construir e validar um teste capaz de mensurar o PC dos alunos? Que sugestão você daria ao pesquisador?

**2.** Obtenha os dados de teste de alguma pesquisa na área da Informática da Educação (de preferência sua) e realize seguintes tarefas:

a) Crie uma hipótese sobre das dimensões do teste, agrupando os itens que exploram os mesmos conceitos;

b) Utilize a técnica da análise fatorial confirmatória para testar a hipótese;

c) Caso não consiga validar a hipótese, utilize a técnica da análise fatorial exploratória para testar outras hipóteses e verificar a existência de uma estrutura válida.

**3.** O quadro abaixo apresenta os dados de um teste realizado por um pesquisador na área da Informática na Educação. Nas linhas temos as pessoas e nas colunas os itens (questões) onde a escala de resposta era de 1 a 5. Utilizando o alfa de Cronbach, responda às seguintes questões:

a) O teste apresenta uma boa consistência interna? Por quê?

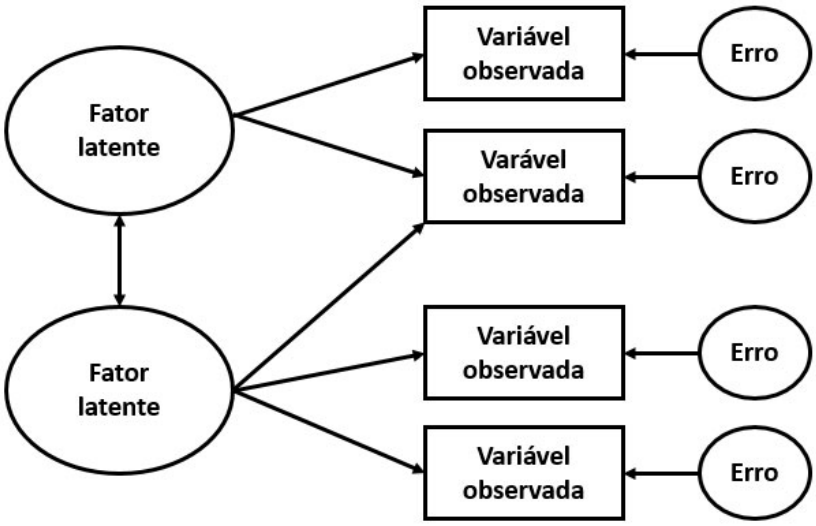
b) Para melhorar a consistência interna do teste, qual dos itens você eliminaria primeiro? Por quê?

| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 | Item 8 | Item 9 |
|---|---|---|---|---|---|---|---|---|---|
| P 1 | 1 | 2 | 3 | 3 | 1 | 3 | 4 | 3 | 2 |
| P 2 | 3 | 3 | 4 | 4 | 5 | 1 | 1 | 2 | 1 |
| P 3 | 4 | 4 | 1 | 3 | 3 | 4 | 4 | 4 | 4 |
| P 4 | 5 | 3 | 5 | 4 | 1 | 1 | 1 | 3 | 3 |
| P 5 | 3 | 1 | 1 | 3 | 4 | 5 | 5 | 1 | 4 |
| P 6 | 5 | 5 | 5 | 5 | 4 | 1 | 2 | 5 | 2 |
| P 7 | 5 | 2 | 3 | 2 | 1 | 4 | 1 | 1 | 5 |
| P 8 | 4 | 3 | 5 | 1 | 4 | 1 | 5 | 5 | 1 |
| P 9 | 1 | 1 | 4 | 4 | 1 | 5 | 5 | 2 | 4 |
| P 10 | 2 | 5 | 4 | 1 | 5 | 1 | 5 | 2 | 1 |
| P 11 | 5 | 2 | 3 | 4 | 1 | 4 | 5 | 5 | 5 |
| P 12 | 1 | 1 | 1 | 3 | 4 | 1 | 5 | 1 | 2 |
| P 13 | 5 | 5 | 5 | 5 | 3 | 5 | 3 | 5 | 5 |
| P 14 | 4 | 1 | 4 | 3 | 1 | 1 | 1 | 1 | 4 |
| P 15 | 1 | 4 | 1 | 3 | 5 | 3 | 4 | 4 | 2 |
| P 16 | 5 | 1 | 4 | 4 | 2 | 5 | 2 | 1 | 3 |
| P 17 | 3 | 4 | 1 | 1 | 3 | 1 | 5 | 5 | 5 |
| P 18 | 1 | 2 | 5 | 3 | 1 | 5 | 4 | 1 | 4 |
| P 19 | 3 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 2 |
| P 20 | 2 | 3 | 5 | 1 | 3 | 3 | 3 | 4 | 1 |

**Respostas comentadas**

**1.** O conceito a ser medido pelo teste é muito amplo. O PC é um processo de resolução de problemas que inclui diversas habilidades, tais como análise de dados, abstração, decomposição e paralelização. Mesmo que o pesquisador crie itens para cada uma dessas habilidades, pode ser muito difícil validar a estrutura subjacente do teste. Uma possível sugestão seria: delimitar melhor a variável em estudo, especificando quais as habilidades, dentro do domínio do PC, que o jogo pretende melhorar.

**2.** Para construir uma hipótese sobre a estrutura subjacente de um teste, é necessário definir quais são os supostos fatores latentes (dimensões do teste) e como eles supostamente se relacionam entre si e com as variáveis observadas (questões do teste), conforme ilustrado a seguir. Por meio de pesquisas na Internet, é possível encontrar softwares para construir e testar este tipo de hipótese. Com base no modelo e nos dados do teste, o software irá fornecer os coeficientes dos relacionamentos, o que permitirá confirmar ou refutar a hipótese.




**3.** Conforme o cálculo apresentado a seguir, verifica-se que o teste não apresenta uma boa consistência interna, pois *a*=0,298.

| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 | Item 8 | Item 9 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| P 1 | 1 | 2 | 3 | 3 | 1 | 3 | 4 | 3 | 2 | 22 |
| P 2 | 3 | 3 | 4 | 4 | 5 | 1 | 1 | 2 | 1 | 24 |
| P 3 | 4 | 4 | 1 | 3 | 3 | 4 | 4 | 4 | 4 | 31 |
| P 4 | 5 | 3 | 5 | 4 | 1 | 1 | 1 | 3 | 3 | 26 |
| P 5 | 3 | 1 | 1 | 3 | 4 | 5 | 5 | 1 | 4 | 27 |
| P 6 | 5 | 5 | 5 | 5 | 4 | 1 | 2 | 5 | 2 | 34 |
| P 7 | 5 | 2 | 3 | 2 | 1 | 4 | 1 | 1 | 5 | 24 |
| P 8 | 4 | 3 | 5 | 1 | 4 | 1 | 5 | 5 | 1 | 29 |
| P 9 | 1 | 1 | 4 | 4 | 1 | 5 | 5 | 2 | 4 | 27 |
| P 10 | 2 | 5 | 4 | 1 | 5 | 1 | 5 | 2 | 1 | 26 |
| P 11 | 5 | 2 | 3 | 4 | 1 | 4 | 5 | 5 | 5 | 34 |
| P 12 | 1 | 1 | 1 | 3 | 4 | 1 | 5 | 1 | 2 | 19 |
| P 13 | 5 | 5 | 5 | 5 | 3 | 5 | 3 | 5 | 5 | 41 |
| P 14 | 4 | 1 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 20 |
| P 15 | 1 | 4 | 1 | 3 | 5 | 3 | 4 | 4 | 2 | 27 |
| P 16 | 5 | 1 | 4 | 4 | 2 | 5 | 2 | 1 | 3 | 27 |
| P 17 | 3 | 4 | 1 | 1 | 3 | 1 | 5 | 5 | 5 | 28 |
| P 18 | 1 | 2 | 5 | 3 | 1 | 5 | 4 | 1 | 4 | 26 |
| P 19 | 3 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 2 | 18 |
| P 20 | 2 | 3 | 5 | 1 | 3 | 3 | 3 | 4 | 1 | 25 |
| Variância | **2,4275** | **2,0275** | **2,3275** | **1,5475** | **2,21** | **2,8875** | **2,4275** | **2,5275** | **2,1** | **27,8875** |

| | |
|---|---|
| Número de itens (k) | **9** |
| Soma das variâncias dos itens (Vi) | **20,4825** |
| Variância total do teste (Vt) | **27,8875** |
| Valor alfa (a) | **0,298723** |

$$\alpha = \left(\frac{K}{K-1}\right)\left(1-\frac{\sum V_i}{V_T}\right)$$

Refazendo o cálculo de *a* para cada item eliminado, verifica-se que a eliminação do item 5 é a que proporciona o maior ganho de consistência interna, conforme apresentado no quadro a seguir.

| Item eliminado | Valor *a* | Ganho |
|---|---|---|
| 1 | 0,223 | -0,075 |
| 2 | 0,119 | -0,180 |
| 3 | 0,367 | 0,068 |
| 4 | 0,232 | -0,067 |
| 5 | 0,398 | 0,100 |
| 6 | 0,294 | -0,004 |
| 7 | 0,383 | 0,084 |
| 8 | 0,062 | -0,236 |
| 9 | 0,278 | -0,021 |

## Sobre os autores


**André Luís Alice Raabe**

http://lattes.cnpq.br/3163271519013006

É Doutor em Informática na Educação pela UFRGS (2005) tendo realizado pós-doutoramento na universidade de Stanford (2016). É professor e pesquisador da UNIVALI (Universidade do Vale do Itajaí). Desenvolve pesquisas sobre Educação em Computação, Pensamento Computacional, Movimento Maker, Software Educacional e Ambientes de Aprendizagem Inteligentes.

**James Roberto Bombasar**

http://lattes.cnpq.br/4943270066505311

É Mestre em Computação Aplicada pela Universidade do Vale do Itajaí (2017). É pesquisador na UNIVALI (Universidade do Vale do Itajaí). Desenvolve pesquisas sobre Informática na Educação e Pensamento Computacional.