COUTINHO, Clara P. (2000). Instrumentos na investigação em Tecnologia Educativa: escolha e avaliação. In A. BARCA & M. PERRALHO (Eds). *Revista Galego-Portuguesa de Psicologia e Educación, Vol. 6(4), Actas do V Congresso Galaico-Português de Psicopedagogia*, 154-166.

Instrumentos na investigação em Tecnologia Educativa: escolha e avaliação.

Clara Coutinho – Universidade do Minho, Portugal - ccoutinho@iep.uminho.pt

RESUMO: Num momento em que tanto se questiona a relevância social e científica da investigação educativa em geral e, em particular, na Tecnologia Educativa, área científica que, com passado ainda recente conta com mais de meio século de actividade investigativa, a questão dos instrumentos utilizados para medida e recolha de dados assume redobrado interesse. De facto, acredita-se cada vez mais que a qualidade dos resultados de uma investigação estejam directamente relacionados com a utilização de instrumentos que, por um lado, sejam válidos e fiáveis, e, por outro, adaptados ao tipo de variável em estudo bem como aos objectivos da investigação.
Pretende-se com esta comunicação auxiliar os investigadores em Tecnologia Educativa nesta tarefa crucial sobre a qual tão pouco se tem escrito e discutido: Que tipo de instrumento se adequa melhor ás variáveis do meu estudo? Devo construir um instrumento ou usar um existente? Onde posso encontrar instrumentos de boa qualidade? Como identifico se um instrumento é válido e fiável? Questões aparentemente simples, mas que colocam desafios a todos nós, docentes e investigadores responsáveis pela produção e avaliação do conhecimento na área.

Introdução

A questão da qualidade dos instrumentos na investigação em Tecnologia Educativa tem vindo a merecer crescente interesse desde que começaram a surgir críticas á investigação desenvolvida na área, de que é um exemplo a célebre série de artigos publicados em torno da questão *"Does media influence learning?"* e que tanto inflamou a comunidade científica da TE na década passada (Clark, 1994; Kozma, 1994; Morrison, 1994, Reiser, 1994, Schrok, 1994). Claro que a questão não é exclusiva da investigação nesta área, expressa-se num sentimento geral de desencanto geral perante os resultados da investigação em educação acusada de "fragmentada", "desajustada à prática" e, sobretudo, demasiado "irrelevante" (Gage, 1991; Kaestle, 1993). Importa inverter esta situação, afastar de uma vez por todas o fantasma da "pseudociência" que ainda a persegue (Tuijman, 1995) e de que só se conseguirá libertar produzindo *investigação socialmente relevante*, capaz de competir na emergente comunidade científica global onde o conhecimento é imediatamente discutido e avaliado.

São muitos os factores que influenciam e determinam a qualidade da investigação educativa - fundamentação teórica robusta, amostras representativas, análise estatística apropriada-, mas o papel dos instrumentos de medida e recolha de dados assume particular destaque (Meltzoff, 1998; Katzer, 1998). No caso particular da TE, esta questão coloca-se com mais pertinência ainda: a incidência da investigação em planos *baseados na medição* (planos experimentais, quase-experimentais, survey e ex post facto) ou seja, planos em que a *qualidade dos instrumentos é condição sine qua non para a relevância e credibilidade dos resultados obtidos*. O princípio de base é bem simples: *a forma como os dados são recolhidos, determina a sua capacidade informativa e, esta, a qualidade da investigação em si*. Se juntarmos a este facto a percepção que todos temos da *falta de instrumentos de medida e avaliação aferidos para a área da TE*, pensamos estar justificada a pertinência desta questão.

Construir um instrumento ou usar um existente?

É a primeira questão com que se defrontam quase todos os investigadores em TE, e para a qual não há, de facto, respostas objectivas. Caberá ao investigador decidir, mas aconselhamos a ponderação dos aspectos que seguem:

1. É sempre possível conceber instrumentos originais de boa qualidade mas isso requer um trabalho considerável. Se a variável é complexa e multidimensional[1] é preciso muito tempo, trabalho e recursos para obtermos um instrumento válido e fiável

2. Se usarmos um instrumento já existente contribuímos para o conhecimento das suas propriedades e valor, sobretudo quando se trata de uma variável central para o domínio de estudo em causa, em que muitos investigam e em que é importante comparar e confrontar resultados para se obterem instrumentos cada vez mais fiáveis e válidos

3. Usar um instrumento já existente mas cujos itens não esgotam a variável em análise, não vale de todo a pena, devendo o investigador avançar para o desenvolvimento de um instrumento novo

Ponderado cada caso, temos duas situações distintas á partida: ou o investigador utiliza um instrumento já existente e estandardizado, ou então adapta ou constrói o seu próprio instrumento.

Instrumentos Estandardizados

Um instrumento é estandardizado se inclui procedimentos uniformes e consistentes para a administração, avaliação e interpretação de resultados (Moore, 1983; Wiersma, 1995). Os testes estandardizados podem dividir-se em dois grupos, referenciados á *norma*[2] ou ao *critério*[3], e aparecem na literatura em variadíssimos formatos classificados em função do conceito ou constructo que pretendem medir, sendo os mais comuns e divulgados os testes de inteligência e aptidões cognitivas (Escala de Inteligência de Wechsler, o Cognitive Abilities Test ou CAT), os testes de personalidade (como o Teste de Rorschach), os inventários de atitudes, de autoconceito, etc.

No caso da investigação em TE, embora não existam muitos testes deste tipo à disposição dos investigadores, poderão eventualmente ser encontrados testes estandardizados para algumas das variáveis em análise e por isso é sempre recomendada a consulta a revistas especializadas para a área da educação (Journal of Educational Measurement) ou a bases de dados internacionais como o ERIC ou o EUDISED[4].

Sempre que um investigador decida utilizar um teste estandardizado, a tarefa essencial é escolher um teste que "meça o que ele próprio pretende também medir" (Wiersma, 1995: 316). Isto requer uma criteriosa revisão de literatura que vá ás fontes originais em que se baseou o construtor do teste, bem como uma verificação da adequação do seu grupo ás normas

[1] Há variáveis simples de operacionalizar (p.e. tempo de resposta a uma questão) mas a grande maioria das variáveis educativas são complexas e multidimensionais. Imaginemos o caso da variável ansiedade: o que é? Como detecto a sua manifestação? Uma variável só está operacionalizada se conseguir captar a totalidade das suas manifestações no mundo real (Smithson, 2000)

[2] Neste caso o desempenho de cada sujeito no teste é comparado com todos os outros que já realizaram a mesma prova e que são o *grupo norma (ou normativo)* que serve de aferição para os desempenhos considerados normais (trata-se de um grupo com características bem definidas e semelhantes a nível de idade, sexo, nível instrução, região geográfica, etc.). São testes de realização máxima, em que se comparam e seriam os sujeitos em função das respostas consideradas "boas" ou "más" às questões do teste. (Moore, 1983)

[3] Um teste referenciado a um critério descreve o desempenho do sujeito com base num dado *critério* predefinido, sem uma comparação inter-individual própriamente dita. O critério que pode ter natureza objectiva - tempo de realização da prova (completar a tarefa sem erro em 20 segundos) é definido por um valor - o "valor do critério"[3]- e o sujeito é classificado num de dois grupos mestria/não mestria ou passar/reprovar. (Black, 1999).

[4]Mesmo no caso de planos de tipo survey, em que o instrumento toma que sempre a forma de inquérito ou questionários em que se obtém a informação perguntando aos sujeitos, é sempre útil procurar na literatura instrumentos já elaborados que podem revelar-se extremamente úteis como base de orientação inicial.

estabelecidas no manual que acompanha o teste, bem como a implementação de um estudo piloto de aferição (Punch, 1998)

Instrumentos Não Estandardizados

São todos os outros que não cabem na categoria anterior e que são construídos/adaptados pelos investigadores podendo tomar formatos diversificados desde o questionário á entrevista, passando pelas escalas e testes de aptidão/aproveitamento.

Prévia à construção ou escolha de um qualquer tipo de instrumento de medida é fundamental que o investigador defina de forma clara e operacional a variável dependente que pretende medir: *a natureza da variável determina o tipo de dados a obter e estes o instrumento que melhor se lhes adequa*. A tabela que apresentamos abaixo, ilustra a relação que existente entre o tipo de instrumento e a natureza da variável a medir:

Tipo de Instrumento **Objectivo do instrumento** (tipo de dados a obter)	**Inquéritos (recolha directa factual)**	**Questionários (escolhas, escalas)**	**Observações entrevistas**	**Testes**
Classificação/ordenação (variávies nominais/ordinais)				
Background pessoal (idade, sexo, nível instrução)	♦			
Classe social (indicadores)	♦		♦	
Tipo de organização(escola)	♦		♦	
Preferências (políticas, disciplinares, etc)	♦		♦	
Traços quantificáveis que originam dados de natureza contínua (variáveis intervalares)				
Atitudes, percepções, opiniões, grau empenhamento		♦	♦	
Valores		♦	♦	
Aproveitamento			♦	♦
Aptidões			♦	♦
Inteligência, criatividade			♦	♦

Adaptado de Black, (1999:191),

O quadro permite uma dupla leitura: a horizontal, informa-nos qual o tipo de instrumento que deveremos usar em função do indicador/variável que pretendemos medir:

- se pretendemos uma recolha directa factual de dados relacionados com o background dos sujeitos, indicadores da classe social, preferências, a técnica será "perguntar" aos sujeitos seja através de um inquérito escrito ou durante uma entrevista
- se o objectivo é, por exemplo, medir atitudes, opiniões, percepções e valores, então o tipo de instrumento a utilizar será, preferencialmente, o questionário de tipo escala que pode ser (ou não) complementado com observações e/ou entrevistas
- pelo contrário, se pretendermos medir o aproveitamento ou as aptidões cognitivas então deveremos optar por um teste de conhecimentos/aptidões

A leitura vertical repete a informação na inversa, ou seja, para cada tipo de instrumento informa quais os indicadores/variáveis que ele está apto a medir:

- Inquéritos[5]: dados do background dos sujeitos, indicadores de classe social, tipo de organização, tipo de organização, preferências: surveys, estudos correlacionais

[5] Sobre construção de questionários consultar: BABBIE, Earl (1997) Survey Research Methods. (2ª Ed) Belmont, California: Wadsworth Publishing Company.
FOWLER, Floyd J. (1993) Survey Research Methods (2ª Ed) . Newbury Park: SAGE Publications. Applied Social Research Methods Series, Vol 1

- Questionários tipo escala[6]: atitudes, opiniões, valores (survey de atitudes, estudos ex post facto)

- Testes[7]: aptidões, aproveitamento, inteligência, etc (estudos de tipo experimental e quase experimental)

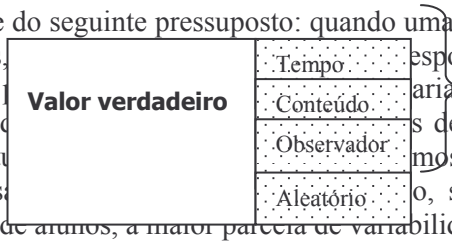
Mas regressemos de novo ao cerne da questão: se encontrarmos na literatura, ou numa investigação afim, um qualquer instrumento para a recolha de dados, como verificar da sua qualidade para a investigação que vamos desenvolver? Por outro lado, se construirmos um instrumento, que garantias me dá de obter dados com qualidade informativa?

Para ambas as questões, a resposta está nos dois critérios técnicos da *fidelidade,* e *validade* também conhecidas por *características psicométricas de um instrumento.*

Fidelidade

Um instrumento ter alta *fidelidade significa que se medirmos hoje com ele ou noutra ocasião (dentro de 15 minutos, depois de amanhã, daqui a um mês) obteremos os mesmos resultados (Black, 1993)*. Na prática é um conceito estatístico baseado na associação entre dois grupos de observações resultantes de duas medições obtidas por um (ou dois) instrumento(s) a um grupo de sujeitos (Wiersma, 1995).

A fidelidade parte do seguinte pressuposto: quando uma série de medições são efectuadas num grupo de sujeitos, [illegible] espondam exactamente o mesmo e terminem com a igual [illegible] ariabilidade nas observações, o que matemáticamente se tra[illegible]s d[illegible] essa variância? Uma parcela será fruto da natu[illegible]mo[illegible] tra parcela é fruto de *erro do instrumento* us[illegible]o, [illegible] um teste de aptidão matemática a um grupo de alunos, a maior parcela de variabilidade (felizmente...!) será fruto das diferenças individuais, mas há sempre uma parcela que é atribuível ao instrumento, ou seja, à forma como as questões são redigidas, como são interpretadas, como foram seleccionadas. Temos assim dois valores: os observados e avaliados com o instrumento (*observed scores*) e os valores verdadeiros, reais (*true scores*) que incluem o *erro instrumental.*



O cálculo da fidelidade obtém-se pela razão entre as duas variâncias (real e de erro), partindo da estimativa das três fontes de erro possíveis:

- *consistência no conteúdo*
- *estabilidade no tempo*
- *consistência entre observadores*

O quadro abaixo sintetiza formas de obter «pares» de medições que vão possibilitar o cálculo dos diferentes tipos de fidelidade de um instrumento de medida:

Que duas medidas?	Fonte de variabilidade	Tipo de fidelidade
Dois instrumentos separados no tempo	Tempo	Estabilidade
Dois instrumentos simultâneos	Conteúdo, interpretação	Consistência

---

OPPENHEIM, A N (1992) Questionnaire Design, Interviewing and Attitude Measurement (2ª Ed) London: Pinter

[6] Sobre construção de escalas consultar:
DeVELLIS, Robert F. (1991) Scale Development: Theory and Applications. Applied Social Sciences Methods Series Vol 26. Newbury Park: Sage.
GABLE, R. M. (1986) Instrument Development in the Affective Domain. Boston: Kluwer-Nijhoff Publishing.

[7] Sobre construção de testes consultar:
GRONLUND, N. E. (1976) Measurement and evaluation in teaching. New York MacMillan
KLINE, Paul (1986) Handbook of Test Construction: Introduction to Psychometric Design. London: Routledge
HAMBLETON, Ronald K.; ZAAL, Jac N. (Ed) (1998) Advances in Educational and Psychological Testing. Boston: Kluwer Academic Publishers

Duas metades do mesmo instrumento	Conteúdo, interpretação	Consistência, estabilidade
Mesmo instrumento administrado 2 vezes em tempos diferentes	Tempo	Estabilidade
Mesmo instrumento administrado pelo mesmo avaliador em tempos diferentes	Tempo	Consistência
Mesmo instrumento administrado por 2 observadores diferentes	Observador	Consistência

Adaptado de Black (1999:197)

Coeficientes de fidelidade

Há vários indicadores da fidelidade de um instrumento devendo cada investigador optar pelo que melhor se adapta ao seu caso específico:

1. *Coeficiente de estabilidade*, para a medida da estabilidade de um único instrumento administrado mais do que uma vez (teste-reteste)

2. *Coeficiente de equivalência*, para a medida da equivalência de dois instrumentos diferentes para o mesmo domínio (formas paralelas)

3. *Coeficientes para medida da consistência interna*: a) coeficiente de bipartição (split-half); b) Alpha de Cronbach (α) e c) Coeficiente de Kuder-Richardson

4. *Coeficiente de fidelidade do acordo de observadores*. a) inter-observadores para observadores distintos, ou b) intra observadores, para um mesmo observador em dois ou mais momentos separados no tempo

Coeficiente de Estabilidade (Test-Reteste)

Procura averiguar da estabilidade do instrumento no tempo, ou seja, se esta dá resultados idênticos quando administrado em dois momentos diferentes (Moore, 1983). Os resultados das duas aplicações são correlacionados pelo coeficiente de Pearson produto/momento, e o valor obtido indica a estabilidade do teste no tempo. Se for superior a .80, significa que os sujeitos tendem a obter scores semelhantes (não iguais) nas duas aplicações logo o instrumento é estável.

Este coeficiente deve calcular-se sempre em planos com pré-pós teste, e em planos de tipo série temporal. Convém no entanto não esquecer que há que interpretar este coeficiente com cuidado uma vez que, a)a repetição do mesmo instrumento aos mesmos sujeitos pode influenciar a segunda aplicação (efeito conhecimento do teste), b)se o intervalo entre as duas aplicações fôr muito curto (dias a uma semana) os sujeitos podem relembrar as respostas e invalidar estimativas; se demasiado longo (um ano ou mais) pode não reflectir a estabilidade fruto da maturação dos sujeitos.

Coeficiente de Equivalência

Mede a consistência do conteúdo de um instrumento, administrando aos mesmos sujeitos, e de preferência no mesmo dia, duas versões equivalentes de um teste para o mesmo domínio/constructo a medir (formas paralelas). Especialmente indicado para testes de aproveitamento que cobrem determinado conteúdo/área curricular ou ainda para alguns questionários de atitudes, não se aplica a testes que abarquem dimensões de tipo psicológico. O seu cálculo é feito como anteriormente pelo coeficiente de correlação de Pearson produto/momento. (Black, 1999; Moore, 1983; Almeida & Freire, 1997).

Consistência Interna

É o verdadeiro indicador da *homogeneidade* das questões num teste ou questionário, ou seja, do grau relativo com que as respostas a itens individuais se correlacionam com o valor total obtido no teste, sendo a única medida possível de obter quando temos um *único teste que é administrado uma única vez* (DeVellis, 1991; Black, 1999; Punch, 1998).

A estimativa da consistência interna de um instrumento pode fazer-se por duas vias:

- a partir de duas medições obtidas no mesmo grupo de sujeitos em duas versões gémeas de um mesmo instrumento, ou seja, pelo *coeficiente de bipartição do teste (split half)*

- a partir de uma medição única feita com o instrumento tendo em conta a) a média das correlações entre todos os seus itens ou partes e b) o número de itens ou partes, como o fazem os coeficientes alpha de Cronbach α) ou a fórmula de Kuder Richardson,

Coeficiente de Bipartição (split half)

Ideal se o teste foi construído como sendo formado por duas metades paralelas, em que cada questão tem uma gémea algures no teste (pergunta o mesmo por outras palavras), indicando este coficiente a consistência com que os sujeitos tendem a responder a pares de questões. O teste é tratado matemáticamente como se de dois testes se tratasse (duas médias e dois desvios padrão), e o coeficiente de correlação de Pearson é aplicado com uma leve adaptação[8] derivada do facto de termos dois meios testes, mais curtos do que o original (a fidelidade de um teste é sensível ao número de questões).

É o indicador mais apropriado para testes cognitivos (aproveitamento) e para testes afectivos ou de atitudes. (Black, 1999)

Alpha de Cronbach

É o indicador mais aconselhado para o cálculo da consistência interna de instrumentos de tipo escala de Likert ou rating. Em termos de procedimento, exige uma única aplicação do teste. Em termos matemáticos, procura avaliar em que grau a variância geral dos resultados da prova se associa ao somatório da variância item a item (Almeida & Freire, 1997). De certa forma, este coeficiente estima a média de todas as possíveis bipartições do instrumento, pelo que não deve ser aplicado em instrumentos em que a bipartição é possível (caso anterior) porque é menos preciso e rigoroso. O seu cálculo tem em conta as médias das correlações inter itens bem como o número de questões do teste.

[8] A aplicação da fórmula de correcção visa atenuar o efeito negativo da diminuição do número de itens, ou seja, estima o coeficiente esperado se o mesmo fosse calculado com o tamanho do teste na sua globalidade (Almeida & Freire, 1997)

Interpretar o Alpha de Cronbach requer alguns cuidados. Por vezes, um valor de α moderado pode não significar baixa fidelidade do teste em si, mas apenas de falta de homogeneidade em alguns dos seus itens, que por vezes há que retirar. A melhor forma de avaliar itens individuais é calcular uma matriz de correlações que relacione cada item com a correlação total da prova (Black, 1999). Podemos então comparar a resposta de um sujeito a um item com o valor da sua resposta ao instrumento como um todo, e verificar quão consistentemente um item mede o mesmo que o instrumento total, contribuindo para a sua consistência interna. Um item com correlação baixa (ou negativa) mostra que não induziu a respostas consistentes com o instrumento no seu todo. Este valor é conhecido como o poder discriminativo[9] do item.

Coeficiente de Kuder-Richardson

Nem sempre os testes podem ser concebidos como sendo constituídos por duas metades emparelhadas, pelo que nestes casos o coeficiente de bipartição não se aplica.

Quando se trata de respostas objectivas de tipo certo/errado ou escolha múltipla o coeficiente de Kuder-Richardson[10], uma adaptação do α de Cronbach é o mais apropriado.

Paralelamente ao cálculo deste coeficiente é aconselhável neste tipo de testes de resposta objectiva que o investigador proceda a uma análise individual de itens, para aferir do seu *indice de dificuldade*[11] e do *poder discriminativo* a que já nos referimos anteriormente,, aspectos fundamentais em testes que requerem uma resposta objectiva (testes de aproveitamento e/ou desempenho) (Black, 1999)

Fidelidade de acordo entre observadores

Uma outra forma de estimar a fidelidade de um teste é a sua capacidade para atribuir uma cotação correcta a um sujeito, o que por vezes depende de critérios que tem a ver com o observador/avaliador (até que ponto o observador não foi demasiado exigente/condescendente na atribuição de cotações aos sujeitos?)

Esta questão coloca-se sempre que o desempenho de um sujeito num teste seja susceptível de múltiplas interpretações, situações essas em que é conveniente calcular a fidelidade entre observadores ou acordo entre observadores (Moore, 1983; Black, 1999): um grupo de observadores/avaliadores/juízes atribui pontuações a um determinado número de testes (por exemplo 4 avaliadores pontuam 10 testes). Calculam-se coeficientes de correlação de Pearson para todos os pares de observações possíveis, e obtém-se o valor do *acordo entre observadores*, medida da fidelidade a este nível.

---

[9] Para mais informação consultar Almeida & Freire (1997:129).

[10] A fórmula pode encontrar-se em Black (1999:282) ou em Almeida & Freire (1997: 150)

[11] O índice de dificuldade traduz a proporção de sujeitos que respondem correctamente a um item, e calcula-se dividindo o nº de respostas correctas pelo nº total de sujeitos que realizaram a prova, o que nos dá uma relação inversa, ou seja quanto maior o valor de ID mais fácil a questão. O leque de resultados possíveis varia entre 0 e 1.00, sendo que no primeiro caso ninguém responde correctamente e no segundo todos respondem correctamente. O valor médio ID= 0.50 corresponde ao valor que mais discrimina os sujeitos em termos da dificuldade por proporcionar maior variância nos resultados. A escolha do nível de dificuldade dos itens a incluir na prova depende dos objectivos da prova: nos testes normativos que visam a seriação é aconselhável optar por mais itens de dificuldade média (0.5); se o teste for referenciado a um critério, já fará mais sentido incluir predominantemente itens com índices de dificuldade mais elevados. Em testes de tipo escala de Likert, a questão do índice de dificuldade dos itens põe-se de forma diferente, uma vez que se avaliam atitudes/valores e não conhecimentos objectivos de tipo certo/errado, não havendo pois itens difíceis e fáceis. No entanto, importa ao investigador seleccionar itens geradores de dispersão nos resultados, já que no final se pretende obter uma pontuação dos sujeitos na escala. Por isso devem seleccionar-se os itens em que os sujeitos se distribuem mais pelos vários pontos da escala, já que são esses os detém maior potencial de variabilidade diferenciando melhor os sujeitos na prova. (Almeida & Freire, 1997)

Também aparece na literatura a chamada *fidelidade intra-observadores*, ou seja, a consistência da avaliação por um mesmo observador ao longo de um período no tempo: por exemplo, será que o observador pontua igual hoje e daqui a seis meses uma grelha de observação de registos video?. Neste caso, calcula-se a correlação de Pearson de pares de observações de um mesmo avaliador. Como forma de estimar a fidelidade intra observadores. (Black, 1999)

Coeficientes de fidelidade: síntese e interpretação

Foram vários os tipos de coeficientes de fidelidade que apresentámos, variando consoante o tipo de instrumento que se usa e os objectivos que tem o investigador.

O quadro abaixo, sintetiza a informação mais relevante que temos vindo a desenvolver:

Tipo de consistência	**Estabilidade (teste-reteste)**	**Equivalência (formas paralelas**	**Interna**			**Acordo observadores**	
						Inter	**Intra**
Coeficiente	**Pearson**	**Pearson**	**Bipartição (Spearman-Brown)**	**α de Cronbach**	**Kuder-Richardson**	**Pearson e α de Cronbach**	**Pearson**
Desempenho/ Aproveitamento							
Objectivo (escolha múltipla, V/F; matching)	♦	♦	♦		♦		
Resposta curta			♦	♦		♦	♦
Resposta longa (ensaio)			♦	♦		♦	♦
Grelha de observações			♦	♦		♦	♦
Atitudes/opiniões							
Questionário (likert, rating)	♦	♦	♦	♦			
Observação			♦	♦		♦	♦
Entrevista estruturada			♦	♦		♦	♦

Teóricamente os coeficientes de fidelidade podem tomar qualquer valor entre 0 e 1.0: se for zero, não haveria nenhuma componente "verdadeira" no resultado obtido com o instrumento, ou seja tudo consistiria em erro. Pelo contrário, um valor de 1.0 significa que o resultado obtido com o teste estaria isento de erro, ou seja seria totalmente "verdadeiro e real".

Almeida & Freire (1997) apontam valores superiores a .85 para os coeficientes de consistência interna e superiores a .75 para os coeficientes assentes na estabilidade. Wiersma (1995) considera valores superiores a .90 como um indicador de boa consistência interna, embora admita que esse valor dependa do tipo de variável a medir, sendo que testes de desempenho e aproveitamento tendem a exigir valores dos coeficientes de fidelidade mais elevados do que os de atitudes e interesses (para estes testes um valor superior a .70 é muitas vezes apontado como limiar aceitável).

Quando num teste o valor do coeficiente de fidelidade é baixo, uma forma prática de o fazer aumentar para o limiar desejável consiste no aumento do número de itens do teste, havendo fórmulas para o respectivo cálculo (Almeida & Freire, 1997:152)

No caso do coeficiente de acordo entre observadores o patamar de >0.90 é o mais aceite se forem tidos em conta a totalidade dos registos, embora possam aceitar-se valores de >0.75 se não se contabilizarem as situações de ausência de registo (Almeida & Freire, 1997)

Convém relembrar que o cálculo da fidelidade se processa sempre depois de um instrumento estar concluído, na fase do chamado estudo piloto que deve anteceder a sua aplicação definitiva e em que são ainda possíveis rectificações ao mesmo. No entanto, a investigação confirma que há cuidados que tidos em conta durante a construção de um instrumento, podem ser determinantes para aumentar a fidelidade. São eles:

- Número suficiente de itens capazes de cobrirem o universo operacional da variável: instrumentos longos são geralmente mais fiáveis por reduzirem a variabilidade derivada do erro
- Qualidade da redacção dos itens, em termos de objectividade e factualidade (evitar palavras difíceis de interpretar, ambíguas e/ou indutoras de resposta)
- Tempo permitido/necessário: urge uma adequação lógica
- Homogeneidade do grupo para o traço a medir pode diminuir a fidelidade já que a variância dos resultados verdadeiros será menor

Validade de um instrumento

A validade indica se um teste, de facto, "mede aquilo que acreditamos (ou queremos) que ele meça " (Punch, 1998:100). Quer isto dizer que o instrumento, tal como a definição operacional, tem de ser consistente e cobrir todos os aspectos do conceito abstracto (variável) que se está a estudar (Black, 1993).

Na literatura é feita referência a três tipos de validade: de *conteúdo*, de *critério* e de *constructo*.

Validade de Conteúdo

Também chamada validade lógica (Almeida & Freire, 1997) ou ainda de "face" aplica-se á validação do conteúdo de testes de aptidão cognitiva e de aproveitamento (Black, 1995)

O objectivo aqui é investigar se o conteúdo dos itens da prova *cobrem os aspectos mais relevantes do constructo/conceito que o instrumento pretende medir*. De natureza subjectiva, não é possível obter um valor numérico para este indicador como acontecia com o coeficiente de fidelidade que referimos atrás. Aqui o que se costuma fazer é submeter o teste á opinião de peritos e especialistas que se vão pronunciar sobre uma *tabela de especificações* onde o investigador operacionalizou as definições dos construtos que os itens devem abranger. (Almeida & Freire, 1997).

Moore (1983) sugere que metade dos especialistas analisem o teste em si, pronunciando-se sobre aquilo que ele pretende medir, enquanto a outra metade a quem se fornece o teste e a tabela de especificações deve avaliar se os itens cobrem a totalidade do constructo; do confronto de ambos resulta uma estimativa da validade de conteúdo do teste. Para este autor, a validade de conteúdo faz sentido em instrumentos que avaliam inteligência, atitudes, conhecimentos e destrezas dos sujeitos, mas pode não se adequar a testes de personalidade e outras variáveis psicológicas (Moore, 1983).

Validade de Critério

Até há bem pouco tempo era a validade mais divulgada na investigação educativa, embora na actualidade tenha caído em desuso e sendo de certa forma englobada na validade de constructo de que falaremos em seguida.

Averiguar o nível deste indicador exige comparar os resultados obtidos no instrumento com outro já existente (que constitui o *critério* externo), calculando-se as correlações dos resultados dos sujeitos nos dois testes (Black, 1999).

Mas fará algum sentido justificar a validade de um instrumento novo com base nos resultados obtidos pelos sujeitos noutro cuja validade não é da nossa responsabilidade? Quem nos garante que haja uma associação entre a minha prova e a prova critério?

Para a maioria dos autores que consultámos, o cálculo da validade por referência a um critério externo pode servir como um referencial para a validade do instrumento, mas não deve ser nunca a *única justificação* para a sua validade (Black, 1999; Moore, 1983; Wiersma, 1995; Smithson, 2000)

Validade de conceito ou de constructo

Para Moore (1983) é o sentido de validade mais ampla e actual englobando as anteriormente referidas. Para Almeida & Freire (1997:159) " o que está em causa neste tipo de validade é o grau de consonância entre os resultados obtidos no teste, a teoria e a prática a propósito das dimensões em avaliação e daí a pertinência da expressão também usada de validade hipotético-dedutiva "

A validade de constructo deve acompanhar todo o processo da construção de um instrumento e não se expressa sob a simples forma de um coeficiente de correlação. Quer isto dizer que a metodologia usada para a apreciação da validade de conceito ou constructo de um teste é diversificada e, segundo Black (1999) deve comportar a combinação e ponderação da informação proveniente de três abordagens distintas: a *lógica*, *estatística* e *empírica*

A *abordagem lógica* à elaboração do constructo (com base numa teoria ou modelo) implica a análise lógica da consistência entre a definição do constructo e a sua especificação e operacionalização em termos dos itens contidos no instrumento. É nesta fase ainda que se deve proceder à revisão da redacção dos itens, sobretudo quando se trata de testes de resposta não objectiva caso de atitudes, opiniões e valores. Black (1999) aponta alguns dos possíveis desvios derivados de padrões de resposta e/ou má redacção dos itens:

a) dissimulação, se o sujeito responde em função do que pensa irá causar boa impressão no avaliador, como sucede em testes de personalidade em que a intenção da pergunta é transparente

b) socialmente desejável, se os sujeitos respondem não o que pensam mas o que consideram ser socialmente aceite: por ex. homens de meia idade não devem gostar de música metálica

c) distorção, caso das escalas de Likert em que os sujeitos se inclinam para um extremo ou o centro pode revelar incapacidade de induzir uma tomada de decisão (a prevalência de respostas nos valores centrais pode indiciar questões mal redigidas, incapazes de estimular resposta)

d) má interpretação das questões, fruto de mau uso do vocabulário, caso de palavras difíceis que não fazem parte do universo dos inquiridos

e) respostas aleatórias em testes de escolha múltipla como acontece quando os inquiridos não forma motivados para a tarefa ("voluntários" pedidos pelo professor)

f) questões indutoras de uma resposta que conduzem a resultados inválidos

A *abordagem estatística*, passa pela aplicação de um método denominado análise factorial aos itens e resultados (Kline, 1994): parte-se da intercorrelação entre os itens de um teste para se identificarem as componentes gerais e/ou diferenciadas que possam explicar a variância comum neles encontrada, sendo que os itens mais válidos e a incluir no instrumento

são os que explicam a maior proporção de variância e por isso se agrupam em clusters explicativos do constructo (Black, 1999).

A *abordagem empírica*, implica a aplicação do instrumento a um grupo piloto identificado como contendo o conceito/traço em análise e sobre o qual versa o instrumento, que será confrontado com outro grupo que, sabemos, não contém esse mesmo traço: se o grupo que possui o traço responder consistentemente melhor ao instrumento do que o grupo sem traço, isso sugere alta validade do instrumento para classificar os sujeitos no traço (Black, 1999)

Conclusão

Introduzimos alguns princípios que estão na base da concepção e escolha de instrumentos para medida e recolha de dados na investigação em Tecnologia Educativa. Definimos formalmente o conceito de fidelidade bem como diferentes formas de calcular o seu coeficiente, em função da natureza do instrumento e da forma como se vão recolher os dados. Analisámos também o conceito de validade, com particular destaque para a validade de constructo, identificando fontes de invalidade bem como formas de a superar e aumentar.

Se lograrmos ter sensibilizado a comunidade de investigadores em Tecnologia Educativa para estas questões, das quais depende a credibilidade e afirmação da nossa área científica em circuitos nacionais e internacionais, terão os nossos objectivos sido atingidos em pleno.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, Leandro; FREIRE, Teresa (1997) Metodologia da Investigação em Psicologia e Educação. Coimbra: APPORT.

BLACK, Thomas R. (1999) Doing quantitative research in the social sciences: na integrated approach to research design, measurement and statistics. London: Sage Publications

CLARK, Richard (1994) Media will never influence learning. Educational Technology Research and Development, Vol 42, nº 2, 20- 28

DeVELLIS S, Robert F. (1991) Scale Development: Theory and Applications. Applied Social Sciences Methods Series Vol 26. Newbury Park: Sage.

GAGE, N. L. (1991) The obviousness of social and educational results. Educational Researcher, nº 20, 10-16

KAESTLE, C. F. (1993) The awful reputation of educational research. Educational Researcher, nº 23, 23-31

KATZER, Jeffrey; COOK, Kenneth; CROUCH, Wayne (1998) Evaluating Information: a Guide for Users of Social Science Research. Boston. Massachusetts: McGraw Hill. 4ª Ed

KLINE, Paul (1994) An Easy Guide to Factor Analysis. London: Routledge.

KOZMA, Robert (1994) Will Media Influence Learning? Reframing the debate. Educational Technology Research and Development, Vol 42, nº 2, 7-19

MELTZOFF, Julian (1998) Critical Thinking About Research- Psychology and Related Fields. Washington DC: American Psychology Association

MOORE, Gary W. (1983) Developing and Evaluating Educational Research. NY: HarperCollins Publishers

MORRISON, Gary (1994) The media effects question: "Unresolvable" or asking th right question?. Educational Technology Research and Development, 42 (2), 41-44

PUNCH, Keith, (1998) Introduction to Social Research: quantitative & qualitative approaches. London: SAGE Publications.

REISER, Robert (1994) Clark`s invitation to the dance: na instructional´s designer response. Educational Technology Research and Development, Vol 42, nº 2, 45-54

SCHUTT, Russell K. (1999) Investigating the Social World: The Process and Practice of Research. (2ª Ed). Thousand Oaks: Pine Forge Press.

SHROCK, S A (1994) The media influences debate: read the fine print, but don´t loose sight of the big picture. Educational Technology Research and Development, 42 (2), 49-53

SMITHSON, Michael (2000) Statistics with confidence. London: SAGE Publications.

TUIJMAN, A (1995) The Futures of Educational Research. In CAMPOS, Bártolo (Org) (1995) A investigação educacional em Portugal. Lisboa: IIE.

WIERSMA, William (1995) Research Methods in Education: an introduction, 6th Ed. Boston: Allyn and Bacon.