Parte I - Métodos em epidemiologia nutricional

13 - Desenvolvimento de instrumentos de aferição epidemiológicos

Michael Eduardo Reichenheim
Claudia Leite Moraes

REICHENHEIM, ME., and MORAES, CL. Desenvolvimento de instrumentos de aferição epidemiológicos. In: KAC, G., SICHIERI, R., and GIGANTE, DP., orgs. *Epidemiologia nutricional* [online]. Rio de Janeiro: Editora FIOCRUZ/Atheneu, 2007, pp. 227-243. ISBN 978-85-7541-320-3. Available from SciELO Books <http://books.scielo.org>.

13

Desenvolvimento de Instrumentos de Aferição Epidemiológicos

Michael Eduardo Reichenheim e Claudia Leite Moraes

Este capítulo trata do desenvolvimento de instrumentos de aferição, uma área de interesse metodológico que nitidamente vai além do domínio precípuo das ferramentas e técnicas voltadas para a aferição nutricional (e que estão bem contempladas em outros capítulos deste livro). Ainda assim, este tema não é de todo estranho à área temática da epidemiologia nutricional, pois muitos de seus programas de investigação concernem a estudos de objetos que transcendem as avaliações sobre o *status* nutricional em si, tais como contenção alimentar (*dietary restraint*) (Bond, McDowell & Wilkinson, 2001; Van Strien et al., 2006), avaliação de apetite (Wilson et al., 2005), comportamento alimentar (Burrows & Cooper, 2002; De Lauzon et al., 2004; Wright, Parkinson & Drewett, 2006) ou conhecimentos e educação nutricional (Vereecken, Van Damme & Maes, 2005; Whati et al., 2005; Zinn, Schofield & Wall, 2005). Ademais, muitos estudos da área se debruçam sobre causas e determinantes das disfunções nutricionais e, destarte, forçosamente abarcam os vários domínios afins oriundos da epidemiologia como um todo. Claramente, rigor e refinamento na incorporação de construtos e dimensões conexas também requerem rigor e refinamento na escolha e uso do respectivo instrumental de aferição.

Como o leitor perceberá, a exposição que se segue está mais fundamentada na tradição de pesquisa oriunda das áreas de psicologia e educação denominada 'conceptualização dimensional' em contraste à 'categórica'. Esta, por sua vez, é mais afeita à área médica tradicional, cuja preocupação se concentra prioritariamente em diagnósticos e tratamentos. Uma premissa estruturante da abordagem 'dimensional' é que, subjacente aos itens empíricos manifestos, existe um contínuo de intensidade e/ou gravidade do fenômeno de interesse. Assim, tendo-se identificado instrumentos de aferição acurados e confiáveis, fica subentendido ser possível 'posicionar' indivíduos (unidades de análise) ao longo do espectro latente e, do ponto de vista das relações de determinação entre fenômenos assim mensurados, ser possível também uma aproximação verossímil entre os nexos conceituais sob investigação. Além dos construtos afins à área da epidemiologia nutricional citados anteriormente, bons exemplos de variáveis latentes assim trabalhadas são o apoio social (Sherbourne & Stewart, 1991; Chor et al., 2001), a resiliência (Wagnild & Young, 1993; Pesce et al., 2005), a qualidade de vida (Guillemin, Bombardier & Beaton, 1993; Teixeira-Salmela et al., 2004), a violência entre parceiros íntimos (Krug et al., 2002; Moraes & Reichenheim, 2002) e a auto-estima (Schmitt & Allik, 2005).

Qualquer texto sobre instrumentos de aferição necessita perpassar pela questão da validade de estudos epidemiológicos como um todo, aliás, um tema de constante preocupação e que tem gerado contínuos debates entre pesquisadores. Certos autores salientam a necessidade de detalhamento sobre as possíveis fontes de erros

sistemáticos e aleatórios na tentativa de evitar ou minimizar vieses (Kleinbaum, Kupper & Morgenstern, 1982; Miettinen, 1985; Steineck & Ahlbom, 1992; Rothman & Greenland, 1998). Outros enfatizam a necessidade de embasamento teórico-conceitual no desenvolvimento e execução de estudos epidemiológicos (Krieger & Zierler, 1997; Pearl, 2000; Weed, 2001; Greenland & Brumback, 2002; Luiz & Struchiner, 2002; Rothman & Greenland, 2005). Na tentativa de unificar os vários aspectos que perpassam a qualidade de estudos epidemiológicos, Reichenheim e Moraes (1998) propuseram seis pilares para a apreciação de validade, percorrendo as questões conceituais, operacionais, de domínio do estudo, de comparação, de mensuração e de especificação dos modelos estatísticos empregados. Dois destes pilares – validade operacional e de mensuração – são de particular interesse quando se discutem os fatores que podem influenciar a qualidade das informações.

Para a compreensão da validade operacional, é preciso perceber que a epidemiologia, como outras áreas da ciência, opera nos campos teórico e empírico. A conexão entre ambos é mediada pela formulação de hipóteses que expressam as relações terminais de um modelo teórico, servindo como ponte entre este e a realidade (Almeida Filho, 1989; Krieger & Zierler, 1997). A partir da construção desse quadro, o pesquisador organiza suas idéias em relação ao processo que está investigando, o que torna possível identificar as hipóteses de pesquisa e os nexos entre os construtos e respectivas dimensões teóricas supostamente envolvidas. Estabelecido o modelo teórico-conceitual, definem-se os indicadores e variáveis para representar os conceitos subjacentes no nível empírico. Esta etapa requer máxima atenção e aprofundamento, pois o traslado de um quadro teórico – em si um recorte da realidade – para o plano empírico produz, inevitavelmente, ainda mais simplificações. O pouco cuidado no processo de redução dos conceitos às variáveis e indicadores pode fazer com que um 'falso' representante do conceito seja incorretamente incorporado em uma análise subseqüente. Sem dúvida, a utilização de instrumentos de aferição bem desenvolvidos pode em muito contribuir para a qualidade de um estudo.

Um segundo eixo de interesse aborda os aspectos relacionados à validade da informação (ou sua falta) conseqüente ao processo de mensuração. Problemas na aferição e o seu enfrentamento têm tido atenção especial no meio epidemiológico e bioestatístico (Dunn, 1989; Carroll, Ruppert & Stefanski, 1995; Streiner & Norman, 2003). A abordagem tradicional parte da classificação que separa a confiabilidade da validade de um instrumento (Nunnally & Bernstein, 1995; Streiner & Norman, 2003). No entanto, é central explicitar o que se tem em mente ao se confrontar esses dois conceitos. Ainda que a sucessiva demonstração de confiabilidade seja útil para recomendar um instrumento de aferição a médio ou longo prazo, *grosso modo*, a confiabilidade diz respeito à qualidade do processo de aferição precípuo de um estudo, não sendo, portanto, uma característica estrutural ou imanente do instrumento de aferição. Trata-se de algo conjuntural e específico do processo. Pode-se pensar a confiabilidade como elemento que conota a robustez da aferição (ou sua falta) em um estudo pontual, apreendendo as pressões exercidas pelo examinador e o examinado sobre o instrumento. Pelo caráter particular dessa interação, a confiabilidade precisa ser investigada em cada pesquisa, e seus resultados são, em princípio, intransferíveis (Armstrong, White & Saracci, 1995).

Em contraposição, diz-se que um instrumento é válido se mede o que se espera que meça em termos do objeto ou fenômeno em questão (McDowell & Newell, 1996). Desde que se tenha em mente certa constância do domínio de aplicação, a validade pode ser considerada uma propriedade do instrumento, havendo, pois, transposição para uma população externa àquela do estudo de validação do instrumento. Todavia, é imperioso distinguir entre a validade própria do instrumento e a da informação sobre o objeto-alvo, que é finalmente apreendida no estudo epidemiológico em questão. A validade da informação como um todo depende também da confiabilidade do processo de aferição (Streiner & Norman, 2003). Se um instrumento considerado válido *a priori* tem, circunstancialmente, precária estabilidade e replicabilidade em conseqüência de mau desempenho dos entrevistadores, pode haver inadequação da informação captada, a despeito do potencial positivo do instrumento utilizado (Nunnally & Bernstein, 1995; Pett, Lackey & Sullivan, 2003).

Existe uma distinção entre o processo de construção de variáveis representativas de construtos/dimensões teóricas – algo estritamente relacionado à validade operacional – e o processo de aferição (mensuração) de indivíduos em si. É possível conceber uma situação em que ocorra um problema de classificação devido ao uso de uma escala (variável) construída com base em itens (indicadores) inadequados, mesmo não havendo qualquer problema na aferição. Em contrapartida, mesmo diante de uma escala satisfatoriamente concebida e desenvolvida, nada impede que haja um problema de mensuração, levando a um problema na ordenação de indivíduos que potencialmente seriam escalonados de forma acertada. Ambas as situações levam a má classificação dos sujeitos estudados, o que afeta a validade do estudo. Chama-se atenção para a necessidade de explicitação destes dois importantes aspectos – a qualidade do instrumental e de sua aplicação –, não só para garantir a validade interna de um estudo epidemiológico, mas também para permitir a comparação do próprio estudo com achados obtidos em outras pesquisas.

Cabe ressaltar que tanto questões relacionadas à validade operacional como à validade de mensuração têm sido pouco enfatizadas na prática e até, de certa forma, encaradas com descaso por muitos pesquisadores. Freqüentemente, ênfase é dada aos problemas relacionados aos desenhos de estudo e à análise de dados. Vale indagar, no entanto, para que servem um delineamento de estudo adequado e um tratamento de dados que utilize modelagem estatística sofisticada, se a qualidade das informações colhidas deixa a desejar. Esse quadro claramente merece reversão. É central que as estratégias de coleta de informação sejam planejadas cuidadosamente e se baseiem em premissas sólidas, envolvendo tanto as nuanças relacionadas à redução de conceitos a variáveis e indicadores como as inerentes ao processo de aferição.

Sublinhando a importância que merece ser dada à 'validade operacional' em estudos epidemiológicos, alguns dos pontos que permeiam essas reflexões são visitados a seguir. Este capítulo se concentra especificamente nas questões relacionadas ao desenvolvimento de novos instrumentos de aferição. Um outro componente central no âmbito do desenvolvimento e consolidação de ferramentas de aferição, no entanto, concerne ao processo de adaptação transcultural de instrumentos propostos e estabelecidos em outros contextos lingüístico-socioculturais. Para obter mais informações sobre as diversas abordagens teóricas e operacionais, o leitor interessado pode consultar Guillemin, Bombardier & Beaton (1993), Herdman, Fox-Hushby & Badia (1998), Perneger, Leplège & Etter (1999), Behling & Law (2000), Beaton et al. (2000) e Reichenheim & Moraes (2002). Um componente adicional para assegurar a qualidade de informação envolve as questões sobre a mensuração em si. Detalhes podem ser encontrados em Moser & Kalton (1984), Bowling (1997), Reichenheim & Moraes (2002) e Streiner & Norman (2003).

Inicialmente são abordados, aqui, alguns pontos gerais relativos ao instrumental de aferição e que visam a situar o leitor quanto à necessidade de investir em uma adaptação transcultural ou, alternativamente, partir para o desenvolvimento de um novo instrumento, o tema central do presente capítulo. Em seguida, discutem-se as etapas mais relevantes para a construção desses instrumentos.

Lidando com o Instrumental de Aferição

Estudos epidemiológicos com pretensões explicativas (determinantes, fatores de risco ou proteção, fatores etiológicos etc.), a rigor, tendem a utilizar questionários. Comumente, estes são compostos por diferentes módulos, abarcando um ou mais construtos (dimensões)[1] de um modelo teórico a ser testado. Nesse sentido, cada construto implica um instrumento epidemiológico que necessita ser incorporado ao questionário.[2] O primeiro passo para a construção de um questionário multitemático consiste em uma detalhada revisão bibliográfica envolvendo o escrutínio dos instrumentos disponíveis sobre cada um dos construtos de interesse. A compilação do histórico de cada instrumento candidato deve conter uma apreciação sobre o grau de utilização prévia e, principalmente, uma avaliação do estágio de desenvolvimento. Para isso, é crucial examinar as evidências de adequação e suficiência da

trajetória psicométrica[3] existente até então. Essa etapa serve para indicar ao pesquisador se realmente há ou não instrumentos satisfatórios para captar o objeto em pauta e, em se tratando daqueles desenvolvidos e consolidados fora da cultura em questão, se já passaram por um processo formal de adaptação transcultural. Por contraposição, a etapa também permite sugerir que se invista em um instrumental totalmente novo.

Mediante essa primeira e laboriosa etapa, o pesquisador pode decidir se, para um determinado construto, vale a pena admitir incondicionalmente um instrumento, se é preciso iniciar um programa de investigação ancilar de adaptação transcultural, ou, no extremo dos cenários, se há necessidade de partir para a construção de um novo instrumento. Em relação à última possibilidade, não deve passar ao largo o alerta de Streiner e Norman (2003) sobre a pletora de novos instrumentos, sempre considerados 'melhores' do que os antecedentes pelos seus proponentes. Sensatamente, os autores recomendam que o desenvolvimento de um instrumento original seja sempre a última opção, dando-se prioridade aos já existentes. Tempo 'perdido' com uma boa revisão bibliográfica é tempo 'ganho', por evitar que seja preciso investir no desenvolvimento de um novo instrumento que, como o leitor poderá perceber, não é uma tarefa trivial.

Alerta à parte, há ocasiões em que a insuficiência de instrumentos de aferição pertinentes a um ou mais construtos é genuína. Se efetivamente é necessário investir na construção de um novo instrumento, é fundamental que o processo seja o mais rigoroso possível. Como detalhado a seguir, trata-se de um processo longo e trabalhoso que requer diversas etapas, envolvendo os próprios pesquisadores, especialistas e membros da população entre a qual o instrumento será aplicado (Streiner & Norman, 2003).

Desenvolvimento de um Instrumento Novo

As diversas etapas do processo são sucintamente apresentadas no Quadro 1. O processo se inicia com a avaliação dos conceitos que subjazem às dimensões componentes do construto de interesse. Adaptando-se a terminologia cunhada por Wilson (2005), esta etapa do processo poderia ser chamada de 'especificação do mapa do construto'. No entanto, diferentemente do referido autor, que limita o mapa do construto a apenas uma dimensão a cada vez, sugere-se um alargamento de limites para permitir que o processo não só procure delinear o gradiente de intensidade do objeto teórico dentro de uma dimensão precípua, mas também possibilite mapear as possíveis dimensões formadoras do conteúdo do construto como um todo. Assim, faz parte desta etapa entender, debater e demarcar o que Streiner e Norman (2003) chamam de espaço de conteúdo. Efetivamente, no momento dessa primeira apresentação de um perfil dimensional, ainda se trata de uma postulada validade de face (Streiner & Norman, 2003), cuja corroboração ou refutação terá de ainda ser estabelecida mediante evidências psicométricas em fases posteriores do processo.

Uma vez mapeado o construto, passa-se para especificação e construção de seus indicadores manifestos, isto é, dos itens que comporão o instrumento. A esta etapa Wilson (2005) chama de "desenho de itens". Mesmo se tratando do desenvolvimento de uma ferramenta original, é boa prática que o processo retome a busca bibliográfica pela qual se julgou insuficiente o histórico dos instrumentos. A crítica aos já existentes permite evitar a repetição dos mesmos erros identificados no conjunto disponível, interessando identificar o que pode ser aproveitado das experiências anteriores. Contudo, não se trata de simplesmente enxertar itens antigos. Merece ser lembrado que estes não têm um significado nominal, mas servem para representar espaços de conteúdo do construto (dimensão) subjacente. Por conseguinte, não podem ser interpretados de forma isolada. Aproveitá-los dessa forma pode acarretar problemas de validade (Nunnally & Bernstein, 1995).

Na fase inicial de busca de itens, é profícuo investir em estudos qualitativos, como, entre outros, os métodos de consenso pela técnica Delphi, o processo de grupos nominais ou o de grupos focais (Dawson, Manderson & Tallo, 1992; Denzin & Lincoln, 1994; Krueger, 1994; Bowling, 1997). Nas situações em que nada ou pouco se sabe sobre como certo construto é percebido pela população-alvo, pode-se afirmar que estudos qualitativos são

obrigatórios. A meta é identificar os itens que melhor representem os conceitos de interesse. Várias opções devem ser propostas para que uma crítica subseqüente avalie e selecione os mais interessantes. O principal desafio é especificar um conjunto que seja suficientemente completo para garantir a validade de conteúdo, mas não tão extenso a ponto de dificultar a aceitabilidade e aplicabilidade do instrumento.

Quadro 1 – Etapas envolvidas na elaboração de um novo instrumento

Etapas[a]		Estratégia de execução	
Especificação do mapa do construto	Explicitação dos conceitos, identificando-se os construtos e respectivas dimensões a considerar	Revisão bibliográfica Apreciação do modelo teórico do estudo Identificação das dimensões que compõem o construto Identificação do possível gradiente de intensidade do objeto teórico	
Especificação do desenho de itens	Proposição de itens que representem as dimensões a estudar Seleção dos itens que comporão as primeiras edições do instrumento (protótipos) Redação das perguntas	Revisão bibliográfica Discussão envolvendo pesquisadores, outros especialistas e indivíduos da população-alvo Discussão envolvendo pesquisadores e outros especialistas Pesquisadores Discussão envolvendo pesquisadores e indivíduos da população-alvo	Pré-teste Aplicação dos protótipos a indivíduos da população-alvo visando a avaliar aceitabilidade, compreensão e impacto emocional.
Especificação do espaço de desfecho	Discussão do sistema de escores/opções de respostas	Discussão envolvendo pesquisadores e outros especialistas	
Especificação do modelo de medida	Avaliação das características psicométricas dos protótipos	Avaliação de validade dimensional e adequação de itens componentes Avaliação de confiabilidade (consistência interna, estabilidade temporal etc.) Avaliação de validade de construto e de critério	
Decisão	Seleção do instrumento final Estudos de corroboração	Discussão envolvendo pesquisadores e outros especialistas Utilização do instrumento em outros contextos de pesquisa	

a - Modelo e nomenclatura adaptados de Wilson (2005).

Tratado esse importante aspecto, passa-se ao aprimoramento e adequação semântica dos itens, estabelecendo-se uma ou mais alternativas de perguntas a serem testadas em seguida. Aqui interessa alcançar uma redação objetiva, clara, simples e curta, evitando-se frases ambíguas e com múltipla significação (Moser & Kalton, 1984; Converse & Presser, 1986; Streiner & Norman, 2003). A literatura recomenda que a escolha dos termos considere as particularidades da população-alvo à qual o instrumento se dirige, com destaque para os de fácil compreensão, harmônicos com a cultura em questão e sem erudição supérflua. Também tem-se enfatizado que um bom texto deve evitar assertivas 'positivas' e 'negativas' inseridas no mesmo item, jargão profissional (por exemplo, médico) e coloquialismo (gírias) indevido. Quanto à seqüência de itens, recomenda-se que os mais delicados ou constrangedores sejam colocados no final do instrumento, ainda que exceções possam ser encontradas em certos casos. Por exemplo, no desenvolvimento do instrumento *Revised Conflict Tactics Scales*, usado para avaliar violência entre parceiros íntimos, chegou-se à conclusão de que intercalar itens de diversas intensidades (gravidades) seria a melhor forma de apresentá-los aos respondentes (Straus et al., 1996).

O passo seguinte consiste em especificar o "espaço do desfecho" (Wilson, 2005), isto é, cuidar da escalonabilidade de cada item. Para atribuir o *status* de validade aos instrumentos, é fundamental que estes sejam capazes de posicionar as unidades de aferição (células, indivíduos, municípios etc.) dentro do espaço de conteúdo do construto (dimen-

são) e lhes atribuir valores e/ou categorias que permitam a demarcação de distâncias e importância. Nesse sentido, vale inicialmente sintonizar a metria interna de cada item com o que estipula o 'mapa do construto' subjacente delineado em etapas anteriores. A literatura sobre o assunto está repleta de técnicas e estratégias com vista à definição de opções de resposta (por exemplo, escalas visuais analógicas, adjetivais, *Likert*, diferenciais semânticas). Evidentemente, um aprofundamento está além do escopo deste texto, mas o leitor pode encontrar valiosos subsídios em Moser & Kalton (1984), Converse & Presser (1986), Streiner & Norman (2003) e Wilson (2005).

Conforme indica o Quadro 1, as etapas de 'desenho de itens' e de especificação do 'espaço do desfecho' contemplam uma primeira visita ao campo, para que os primeiros lotes de protótipos (propostas alternativas) sejam submetidos a uma intensa avaliação de aceitabilidade, compreensão e impacto emocional. Uma técnica interessante no pré-teste é solicitar aos respondentes que parafraseiem cada item, devendo o entrevistador anotar em uma questão adicional se houve ou não compreensão de seus termos. Essa é também uma boa oportunidade para avaliar se as opções de resposta dos itens se adéquam ou não à população-alvo. Tantas 'séries' de n (por exemplo, 30) entrevistas são realizadas até que um percentual preestabelecido de ajustamento (entendimento) em todos os itens seja alcançado (por exemplo, ≥ 90%). Essas avaliações interinas podem ser realizadas pela própria equipe de pesquisa ou, melhor ainda, por um grupo de especialistas no assunto convocados para tal. Com base nas evidências encontradas nesse pré-teste, são escolhidos os protótipos mais promissores, que são postos à prova subseqüentemente.

Parte-se, então, para a consolidação da escala, o que Wilson (2005) chama de "modelo de mensuração". Também reconhecida sob a designação de modelagem psicométrica, esta etapa visa a avaliar os instrumentos-candidatos em diferentes perspectivas. Primeiro, quanto à pertinência dos itens em relação ao construto e às dimensões componentes. É aqui que a validade de face do espaço de conteúdo postulada durante o mapeamento do construto é ou não corroborada. Cada item é testado, não só para avaliar seu peso na formação de uma escala dimensional, mas também se e o quanto contribui de forma exclusiva a apenas uma das escalas formadoras do construto (dimensão). Para além da métrica interna de cada item, também é nesta etapa do processo que se testa e se consolida o escore composto da escala. Nesse passo, procura-se estabelecer e garantir a escalonabilidade do conjunto de itens, independentemente de se a escala é constituída por um escore calculado diretamente com base nas análises multivariadas que subjazem ao processo; por um escore obtido por meio do somatório simples ou ponderado da pontuação dos itens componentes; ou ainda por transformações desses escores, tais como percentis, escores-padrão, escores padronizados ou escores normalizados (Streiner & Norman, 2003). Também é parte integral da psicometria a avaliação da confiabilidade potencial e da validade de construto e/ou de critério de cada escala em teste.

Uma síntese dos procedimentos envolvidos nas análises psicométricas está exposta nos Quadros 2 a 5. Devido a restrições editoriais, o conteúdo é forçosamente restritivo e não exaustivo. Entretanto, pode servir como roteiro de aplicação, não só em relação aos objetivos e métodos de análise disponíveis, mas também quanto a uma possível seqüência de procedimentos. Claramente, não há como se apresentar e discutir os prós e os contras de cada método/técnica, mas o leitor poderá notar que estes, assim como alguns outros pontos importantes, podem ser encontrados na bibliografia.

No âmbito do desenvolvimento de instrumentos de 'conceptualização dimensional', a seqüência de quadros procura, esquematicamente, apresentar três enfoques psicométricos. Tão logo se encerra a etapa de especificação do espaço de conteúdo dos itens, a primeira tarefa consiste em corroborar a validade dimensional do instrumento e a adequação dos itens componentes. O Quadro 2 oferece alguns requisitos para que se possam julgar satisfatórias as escalas (e respectivos itens) de um instrumento. Métodos multivariados estão no âmago do processo. Este se inicia com uma Análise de Fatores Exploratória (AFE) (Gorsuch, 1983; Kline, 1994; Pett, Lackey & Sullivan, 2003; Loehlin, 2004; Skrondal & Rabe-Hesketh, 2004), ainda que, no contexto do desenvolvimento de instrumentos, já se tenha alguma estrutura postulada *a priori* quanto à dimensionalidade e aos itens participantes.

Mesmo que a conotação de exploração seja um tanto nebulosa aqui, para que se possa implementar uma Análise de Fatores Confirmatória (AFC) (Maruyama, 1998; Loehlin, 2004; Skrondal & Rabe-Hesketh, 2004; Kline, 2005) com bases firmes, é boa prática realizar uma AFE prévia. Primeiro, para explorar se efetivamente existe a estrutura multidimensional conjeturada, e segundo, para explorar o comportamento dos itens. Evidenciada uma inadequação, nada impede que já nesse ponto da seqüência se tenha de voltar para a 'prancheta', isto é, para fases anteriores com vista ao encontro de novos e melhores itens. O processo iterativo de todo o desenvolvimento é bem nítido.

Quadro 2 – Enfoque psicométrico I. Avaliação de validade dimensional e adequação de itens componentes

Objetivos	Métodos e/ou estimadores
• Estabelecer a dimensionalidade (uni ou multi) postulada na etapa de formulação do mapa do construto, corroborando ou refutando a validade de face postulada quanto aos espaços de conteúdo do construto.	• Análise de Fatores Exploratória (AFE) usando, por exemplo, o método de fatoração por eixos principais com rotação ortogonal do tipo Varimax ou oblíqua do tipo Oblimin (Gorsuch, 1983; Rummel, 1988; Comrey & Lee, 1992; Kline, 1994; Pett, Lackey & Sullivan, 2003; Loehlin, 2004; Skrondal & Rabe-Hesketh, 2004). Questões centrais a observar:
• Identificar os itens mais profícuos em cada uma das escalas dimensionais, escrutinando suas propriedades psicométricas e decidindo pela sua manutenção ou retirada da composição escalar.	- Número de fatores extraídos. - Magnitude das cargas *(loadings)* de cada item nos fatores (isto é, correlação entre itens e fatores). Diversos pontos de corte podem ser utilizados, por exemplo, 0,40. Veja Comrey & Lee (1992) para detalhes.
• Reconhecer e estabelecer o espaço do desfecho de cada escala, propondo uma métrica à consolidação do escore final.	- Presença ou não de cargas cruzadas (*cross-loading*), o que, a princípio, deve ser evitado. Estratégias de decisão podem ser encontradas em Pett, Lackey & Sullivan (2003).
• Apresentar uma ou mais escalas alternativas para cada dimensão do construto, visando à testagem subseqüente (confiabilidade e validade de construto/critério).	• Análise de Fatores Confirmatória (AFC) implementada no âmbito dos modelos de equações estruturais (Bollen, 1989; Maruyama, 1998; Loehlin, 2004; Skrondal & Rabe-Hesketh, 2004; Kline, 2005). Questões centrais a observar: - Corroboração de ausência de cargas cruzadas. - Grau de ajustes de modelo. - Padrão de dimensionalidade, que pode ser de quatro tipos: estrita, forte, intermediária e fraca (veja Skrondal & Rabe-Hesketh, 2004).
	• Análises via modelos de Teoria de Resposta ao Item (TRI) para o caso de escalas formadas por itens dicótomos ou ordinais (Hambleton, Swaminathan & Rogers, 1991; Mellenbergh, 1994; Van der Linden & Hambleton, 1996; Cella & Chang, 2000; Embretson & Reise, 2000; Sijtsma & Molenaar, 2002; Streiner & Norman, 2003; De Boeck & Wilson, 2004; Skrondal & Rabe-Hesketh, 2004; Wilson, 2005). Questões centrais a observar em cada escala dimensional: - Corroboração de escalonabilidade dos itens. - Capacidade discriminante dos itens. - Posicionamento absoluto e relativo dos itens ao longo do contínuo da variável latente (dimensão) subjacente a que a escala do instrumento aspira captar, visando a identificar a presença (indesejável) ou não (desejável) de lacunas de informação ao longo do espectro. - Grau de informatividade coberto pelos itens ao longo da escala. - Precisão de informação ao longo do espectro (contínuo) da variável latente.

Ainda que não explícito no Quadro 2, o método de Teoria de Resposta ao Item (TRI) (Hambleton, Swaminathan & Rogers, 1991; Van der Linden & Hambleton, 1996; Cella & Chang, 2000; Embretson & Reise, 2000; Sijtsma & Molenaar, 2002; Streiner & Norman, 2003; De Boeck & Wilson, 2004; Skrondal & Rabe-Hesketh, 2004; Wilson, 2005) é, de fato, um tipo de AFC baseado em modelos não lineares, apropriado para escalas formadas por itens dicótomos ou ordinais. Além de se alcançar uma melhor especificação do modelo estatístico, uma análise via TRI permite também apreciar algumas propriedades psicométricas atraentes e proveitosas para uma escolha conscienciosa de itens (Reichenheim, Klein & Moraes, 2007). Como indicado no Quadro 2, a TRI permite corroborar a presença de escalonabilidade conjunta dos itens; a capacidade discriminante de cada item; o posicionamento absoluto e relativo dos itens ao longo do contínuo da variável latente (dimensão) subjacente; a abrangência da informatividade dos itens ao longo da escala e a precisão da informação ao longo do espectro (contínuo) da variável latente.

Por mais que uma análise via TRI deva ser encorajada quando se está diante de itens binários ou ordinais, vale comentar que existe alternativa para acomodá-los em análises de fatores (AFE ou AFC), que, a rigor, utilizam matrizes de correlações que assumem distribuições gaussianas. Uma opção para contornar o real problema da má especificação de modelo ao se aplicar análises de fatores 'tradicionais' a dados discretos (Gorsuch, 1983; Rummel, 1988; Jöreskog & Sörbom, 1996) é utilizar matrizes de correlações tetracóricas ou policóricas obtidas por transformações prévias à submissão à análise (Divgi, 1979; Uebersax, 2006). Essas transformações necessitam ser ativamente implementadas em alguns *software* como, por exemplo, [R] (Fox, 2006) e Stata (StataCorp, 2005; Kolenikov, 2006), ou já são usadas como *default* em outros como Lisrel 8 (Jöreskog & Sörbom, 2004).

O segundo enfoque psicométrico envolve avaliações formais de confiabilidade das escalas obtidas após a 'depuração' dos itens e satisfatória evidência de dimensionalidade (Quadro 3). O objetivo é avaliar em que medida os escores de um instrumento (isto é, das escalas componentes) estão livres de erro aleatório (Pedhazur & Schmelkin, 1991), o que, como dito anteriormente, serve não apenas para robustecer a qualidade do estudo relacionado ao desenvolvimento do instrumento em si, mas como uma instância de adequação processual. A longo prazo, uma série de estudos usando certo instrumento e revelando consistentemente uma boa confiabilidade da mensuração (informação) acaba também atestando sua qualidade. Evidências como essas acrescentam ao histórico do instrumento e podem ser benéficas à decisão sobre qual instrumento utilizar em uma pesquisa epidemiológica.

O Quadro 3 oferece várias referências que o leitor poderá consultar para obter mais detalhes sobre a finalidade, o mérito e os procedimentos concernentes a cada tipo de confiabilidade (consistência interna; estabilidade/reprodutibilidade intra[4] ou interobservador; equivalência de formulários). Cabe aqui um comentário sobre a Teoria da Generalização (TG) desenvolvida por Cronbach e colaboradores (1972), cujo objetivo principal é oferecer uma elaborada sistemática para a redução das fontes de erros aleatórios de mensuração. No caso específico de estudos de desenvolvimento de instrumentos em que diferentes tipos de confiabilidade devem ser buscados, é possível obter uma análise 'unificada', na qual os componentes de erros são decompostos e cada aspecto ("faceta", no jargão da TG) é avaliado à luz da contribuição dos outros (Cronbach et al., 1972; Shavelson & Webb, 1991; Nunnally & Bernstein, 1995). Por extensão, é também possível obter um coeficiente de generalização que resume a fração de erro decorrente do conjunto de abordagens.

Mesmo que tenha sido possível identificar dimensionalidade, adequação de itens (em termos de variância compartilhada, como requer a análise de fatores) e confiabilidade, a validade de uma escala precisa ser avaliada explicitamente. Afinal, se um pesquisador visando a centralmente captar um construto C_1 (por exemplo, 'apoio social') inadvertidamente arrolar uma gama de itens consistentemente atinada a um outro construto C_2 (por exemplo, 'resiliência'), é bem plausível que os resultados obtidos nas análises psicométricas descritas anteriormente sejam bastante satisfatórios. Mas nem por isso o instrumento traz 'embutida' automaticamente a validade sobre o construto C_1 em foco. Ainda que as situações no dia-a-dia das pesquisas epidemiológicas sejam bem menos claras, o exemplo lembra que escrutinar a validade de um instrumento vai além das avaliações dos componentes

'internos' de variância, requerendo um escrutínio adicional das covariações das escalas (dimensões) com outros elementos pertencentes ao quadro teórico subjacente. Como já mencionado, assumir validade de face (dos itens) importa nas fases inicias do programa de investigação para guiar as discussões e decisões de escolha dos protótipos de instrumentos a serem mais trabalhados. Mas, diferentemente do que muitos crêem, a validade de face não é suficiente, sendo necessários estudos aprofundados para corroborá-la.

Quadro 3 – Enfoque psicométrico II. Avaliação de confiabilidade

Objetivos	Métodos e/ou estimadores
• Avaliar a consistência interna das escalas identificadas anteriormente.	• Análise via coeficiente α para o caso de variáveis contínuas (Cronbach, 1951; Nunnally & Bernstein, 1995; Osburn, 2000) ou coeficiente de Kuder Richardson, Fórmula 20 no caso de variáveis discretas (Kuder & Richardson, 1937; Streiner & Norman, 2003). Estimadores alternativos são descritos em Osburn (2000). Pontos de corte de decisão (adequação) são discutidos em Nunnally & Bernstein (1995). • Correlação entre cada item e o escore total sem o mesmo – item-resto (Nunnally & Bernstein, 1995). • Percentual de aumento ou redução do coeficiente α ou kr-20 à retirada de cada item da escala – p. ex., 10% (Reichenheim & Moraes, 2006).
• Avaliar a estabilidade temporal (reprodutibilidade intra-observador e teste-reteste) das escalas identificadas anteriormente. • Avaliar a estabilidade (reprodutibilidade interobservador) das escalas identificadas anteriormente.	• Para o caso de variáveis contínuas: análise via correlações intraclasse (Shrout & Fleiss, 1979; Shrout, 1998; Streiner & Norman, 2003), sendo a correlação de Pearson e o coeficiente de concordância de Lin (1989) tipos especiais; ou ainda, o método de Bland e Altman (1986). • Para o caso de variáveis discretas (dicótomas ou policótomas): análises de concordância via estimador *kappa* simples ou ponderado (Cohen, 1960,1968; Fleiss, 1981; Donner & Eliasziw, 1992); ou, alternativamente, coeficiente *kappa* ajustado para viés e prevalência (Byrt, Bishop & Carlin, 1993). Pontos de corte de decisão (adequação) são discutidos em Landis & Koch (1977) e Shrout (1998). • Estimadores alternativos são descritos em Cicchetti & Feinstein (1990) e em uma revisão de Elmore e Feinstein (1992).
• Avaliar a equivalência (de formas) das escalas identificadas anteriormente.	• Análise pelo método de *half-split*, que consiste em estimar de forma sistemática (exaustiva) as correlações entre escores de pares de subescalas (formas paralelas) formadas pela metade dos itens constituintes da escala sob escrutínio (Pett, Lackey & Sullivan, 2003; Streiner & Norman, 2003).

Vários outros tipos além da validade de face têm sido definidos, propostos, utilizados e, até certo ponto, criticados (Streiner & Norman, 2003). Entretanto, no âmbito do desenvolvimento de instrumentos que buscam conceptualizações dimensionais, talvez seja de interesse enfatizar a perspectiva dada por Streiner & Norman (2003), na qual estabelecer a validade de um instrumento, em última instância, é estabelecer a adequação da teoria que a suporta. Estudar a validade de um instrumento é estudar a própria teoria que a embasa, em ciclos de conjecturas e refutações/corroborações. É um processo continuado pelo qual se determina o grau de credibilidade a ser atribuído a uma inferência com base na 'leitura' de uma escala (Landy, 1986; Streiner & Norman, 2003).

Os Quadros 4 e 5 (página seguinte) explicitamente discernem duas situações. A primeira, exposta no Quadro 4, concerne aos objetos de pesquisa em que não há consenso sobre o que seria a referência (ou padrão-ouro) de aferição para o fenômeno de interesse ou quando não é possível defini-la de forma inequívoca. Construtos como 'auto-estima' e 'resiliência' são bons exemplos. Nessa situação, é preciso acessar a validade do construto. Avaliam-se as relações entre as dimensões supostamente captadas pelas diferentes escalas do instrumento, bem como as relações com outros conceitos, atributos e características ligadas à teoria geral na qual se insere o construto sob escrutínio. O encontro de associações previstas ou afinadas com evidências pregressas corrobora e reforça a validade do instrumento. Avaliar o inverso também é relevante, pois constatar a inexistência de relações entre os conceitos teóricos manifestos pelas escalas em pauta e certos construtos (escalas) reconhecidamente fora do escopo

da teoria geral envolvendo o fenômeno de interesse também fortalece a idéia de validade do instrumento. Pode-se constatar que a validade de construto é a epítome de validade teórica.

Quadro 4 – Enfoque psicométrico III-a. Avaliação de validade de construto

Objetivos	Métodos e/ou estimadores
Avaliar a validade de construto quando não há instrumento de referência (padrão-ouro) para o contraste.	• Análise exploratória de associações via tabulações envolvendo duas ou três variáveis (estratificada) e usando razão de risco/prevalência ou razão de produtos cruzados (*odds-ratio*) como estimador; ou associações via coeficiente de correlações de Pearson para variáveis contínuas (Armitage & Berry, 1994) ou coeficientes de correlação não paramétricos (teste de posição de Spearman ou*tau-b* de kendall) para variáveis ordinais (Blalock Jr., 1985). • Análise epidemiológica multivariável complexa encerrando o quadro teórico-conceitual do qual faz parte o construto (e suas respectivas dimensões) sob escrutínio (Kleinbaum, Kupper & Morgenstern, 1982; Rothman & Greenland, 1998; Skrondal & Rabe-Hesketh, 2004). Questões centrais a observar (Cronbach & Meehl, 1955; Streiner & Norman, 2003): - Se e como os conceitos teóricos manifestos pelas escalas dimensionais do construto se relacionam entre si. - Se e como os conceitos teóricos manifestos pelas escalas dimensionais do construto em pauta se relacionam com os outros conceitos prescritos ou postulados pela teoria (validade convergente). - Se os conceitos teóricos manifestos pelas escalas dimensionais do construto em pauta apropriadamente 'não' se relacionam a conceitos que a teoria da qual fazem parte 'não' prescreve ou postula (validade divergente).

Ainda que não seja impeditivo buscar a validade de construto, quando existe um instrumento, exame ou teste de referência para contrastar o 'novo' instrumento em desenvolvimento, é próprio avaliar a validade de critério. Streiner e Norman (2003) distinguem a validade concorrente da preditiva. A classificação se baseia na finalidade da proposta e depende da cronologia de realização dos testes. A validade concorrente é admissível quando já se tem o resultado de um instrumento de referência na ocasião da aplicação do instrumento em teste e permite a apreciação da validade paralelamente à sua aplicação. A validade preditiva só é possível quando as informações auferidas por meio do instrumento de referência são obtidas tempos depois da aplicação do instrumento em teste.

Comumente, estudos de validade de critério são bastante utilizados quando é de interesse maximizar custo-benefício, prever e planejar ações sanitárias, seja reduzindo o próprio instrumento considerado de referência ou propondo um completamente diferente, mas que ainda permita reter a capacidade de classificação original. Contudo, vistas na ótica precípua do contexto de desenvolvimento de um instrumento de conceptualização dimensional (e que não necessariamente pretenda ser uma redução de outro maior, nem uma ferramenta de finalidade pragmática), avaliações da capacidade discriminante das escalas de um instrumento podem ser extremamente esclarecedoras. Saber que um instrumento de aplicação em estudos epidemiológicos não só capta o contínuo da variável latente subjacente, mas também está substantivamente 'colado' ao que um exame ou instrumento de referência encontraria é claramente profícuo e atraente. Os procedimentos apresentados no Quadro 5 são exemplos a serem contemplados.

Quadro 5 – Enfoque psicométrico III-b. Avaliação de validade de critério

Objetivos	Métodos e/ou estimadores
Avaliar a validade de critério (concorrente e preditiva) quando há um instrumento de referência (padrão-ouro) para o contraste.	• Para o caso de se testar uma escala de conceptualização dimensional usando-se o escore completo em relação a um instrumento ou exame de referência de metria contínua: - Análise via correlações intraclasse (Bartko, 1976; Shrout & Fleiss, 1979; Shrout, 1998; Streiner & Norman, 2003), entendendo-se que se está avaliando o grau de concordância do instrumento 'novo' sob escrutínio com uma medida 'infalível' de referência. - Análises via correlação de Pearson também têm sido implementadas em avaliações de concordância, mas seu uso nesse contexto requer alguma reserva. Veja Bartko (1976), Bland & Altman (1986) ou Streiner & Norman (2003) para detalhes. No contexto da epidemiologia nutricional, veja Willett & Lenart (1998) para uma discussão sobre este tópico. • Para o caso de se testar a escala de conceptualização dimensional usando-se o escore completo em relação a um instrumento ou exame de referência de metria dicótoma: - Análise via curvas ROC (*Receiver Operating Characteristic analysis*) (Tanner & Swets, 1954; Hanley & McNeil, 1982; Streiner & Norman, 2003), observando-se em particular a área abaixo da curva ROC, que indica o grau de discriminação da escala em teste em relação ao instrumento de referência. • Para o caso de se testar uma escala de conceptualização dimensional usando-se o escore completo em relação a um instrumento ou exame de referência de metria em mais de dois níveis: - Análise via curvas ROC (Tanner & Swets, 1954; Hanley & McNeil, 1982; Streiner & Norman, 2003) entre níveis crescentes (em gradação, por exemplo), nível 1 *vs* 2 + 3 e 1 + 2 *vs* 3) do instrumento de referência, observando-se o grau de discriminação em relação a cada ponto de corte através da área abaixo da curva ROC. - Uma vez identificado um grau de discriminação satisfatório do conjunto, identificam-se os pontos de corte de máxima discriminação do instrumento em teste, respeitando-se o número de categorias do instrumento de referência. Cria-se uma variável policótoma de tantos níveis quantos os da variável de referência e prossegue-se com as análises via índices de sensibilidade e especificidade (Sackett et al., 1991; Choi, 1992; Fletcher & Fletcher, 2006) obtidas com base em tabelas 2´2, formadas pelas tabulações de variáveis derivadas, tal como se procedeu nas análises via curvas ROC descritas anteriormente. • Para o caso de análises via índices de sensibilidade e especificidade segundo subgrupos/estratos populacionais, pode-se usar modelagem multivariável (Coughlin et al., 1992). - Alternativamente aos índices de sensibilidade e especificidade simples, podem-se usar (a) análises via índices de sensibilidade e especificidade corrigidos por concordância aleatória (Coughlin & Pickle, 1992); (b) o método de razão de verossimilhança (Sackett et al., 1991) ou (c) o coeficiente Phi de concordância (Streiner & Norman, 2003), novamente entendendo-se que se está avaliando o grau de concordância entre uma medida 'infalível' de referência e o instrumento 'novo' sob escrutínio.

Por fim, vale lembrar que o processo de avaliação da qualidade de um novo instrumento não se esgota no primeiro estudo que o utiliza. Mesmo que as evidências iniciais tenham sugerido validade, é capital que se conheça seu desempenho em outros contextos. Uma primeira edição necessita ser continuamente posta à prova pelos profissionais interessados. A vasta gama de detalhes e opções, muitas intrinsecamente subjetivas, demanda que o aprimoramento do novo instrumento dependa de debates e negociações contínuas entre pares.

Notas

[1] Distingue-se, aqui, 'construto' de 'dimensão', entendendo-se que um construto pode ser composto de várias dimensões. Por extensão, entende-se que uma dimensão tem na escala o seu representante empírico que, por sua vez, tem no escore a ordenação numérica subjacente.

[2] Da mesma forma, distingue-se 'instrumento' de 'questionário', convencionando chamar de questionário o conjunto de instrumentos específicos que, por sua vez, abarcam construtos/dimensões específicos.

[3] Entende-se pelo termo 'psicométrica (psicométrico/psicometria)' um conjunto de avaliações quantitativas visando ao escrutínio das propriedades de mensuração de um instrumento. Apesar de ter sido inicialmente proposto e usado no contexto da psicologia e psiquiatria, o termo tem sido largamente utilizado fora dessas áreas.

[4] No contexto de instrumentos de autopreenchimento ou laboratoriais, a confiabilidade intra-observador tem sido denominada 'teste-reteste'.

Referências

ALMEIDA FILHO, N. *Epidemiologia sem Números: uma introdução crítica à ciência epidemiológica*. Rio de Janeiro: Campus, 1989.

ARMITAGE, P. & BERRY, G. *Statistical Methods in Medical Research*. 3. ed. London: Blackwell Scientific Publications, 1994.

BARTKO, J. J. On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 83: 762-765, 1976.

BEATON, D. E. et al. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25: 3.186-3.191, 2000.

BEHLING, O. & LAW, K. S. *Translating Questionnaires and Other Research Instruments*. Thousand Oaks: Sage Publications, 2000. v. 133.

BLALOCK JR., H. M. *Social Statistics*. 2. ed. London: McGraw-HillBook Company, 1985.

BLAND, J. M. & ALTMAN, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 8: 307-310, 1986.

BOLLEN, K. A. *Structural Equations with Latent Variables*. New York: John Wiley & Sons, 1989.

BOND, M. J.; MCDOWELL, A. J. & WILKINSON, J. Y. The measurement of dietary restraint, disinhibition and hunger: an examination of the factor structure of the Three Factor Eating Questionnaire (TFEQ). *International Journal of Obesity and Related Metabolic Disorders*, 25: 900-906, 2001.

BOWLING, A. *Research Methods in Health: investigating health and health services*. Buckingham: Open University Press, 1997.

BURROWS, A. & COOPER, M. Possible risk factors in the development of eating disorders in overweight pre-adolescent girls. *International Journal of Obesity and Related Metabolic Disorders*, 26: 1.268-1.273, 2002.

BYRT, T.; BISHOP, J. & CARLIN, J. B. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46: 423-429, 1993.

CARROLL, R. J.; RUPPERT, D. & STEFANSKI, L. A. *Measurement Errors in Nonlinear Models*. London: Chapman and Hall, 1995.

CELLA, D. & CHANG, C. H. A discussion of item response theory and its application in health status assessment. *Medical Care*, 38, suppl. II: 66-72, 2000.

CHOI, B. C. K. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *Journal of Clinical Epidemiology*, 45: 581-586, 1992.

CHOR, D. et al. Medidas de rede e apoio social no Estudo Pró-Saúde: pré-testes e estudo piloto. *Cadernos de Saúde Pública*, 17: 887-896, 2001.

CICCHETTI, D. V. & FEINSTEIN, A. R. High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43: 551-558, 1990.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20: 37-46, 1960.

COHEN, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70: 213-220, 1968.

COMREY, A. L. & LEE, H. B. *A First Course in Factor Analysis*. Hillsdale: Lawrence Erlbaum, 1992.

CONVERSE, J. M. & PRESSER, S. *Survey Questions: handcrafting the standardized questionnaire*. London: Sage Publication, 1986. v. 63.

COUGHLIN, S. S. & PICKLE, L. W. Sensitivity and specificity-like measures of the validity of a diagnostic test that are corrected for chance agreement. *Epidemiology*, 3: 178-181, 1992.

COUGHLIN, S. S. et al. The logistic modeling of sensitivity, specificity and predictive value of a diagnostic test. *Journal of Clinical Epidemiology*, 45: 1-7, 1992.

CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16: 297-334, 1951.

CRONBACH, L. J. & MEEHL, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 52: 281-302, 1955.

CRONBACH, L. J. et al. *The Dependability of Behavioral Measurement: theory of generalizability for scores and profiles*. New York: Wiley & Sons, 1972.

DAWSON, S.; MANDERSON, L. & TALLO, V. *Social and Economic Research (SER): the Focus Group Manual*. Geneva: World Health Organization, 1992.

DE BOECK, P. & WILSON, M. *Explanatory Item Response Models: a generalized linear and nonlinear approach*. New York: Springer, 2004.

DE LAUZON, B. et al. The Three-Factor Eating Questionnaire-R18 is able to distinguish among different eating patterns in a general population. *Journal of Nutrition*, 134: 2.372-2.380, 2004.

DENZIN, N. K. & LINCOLN, Y. S. *Handbook of Qualitative Research*. London: Sage Publication, 1994.

DIVGI, D. R. Calculation of the tetrachoric correlation coefficient. *Psychometrika*, 44: 169-172, 1979.

DONNER, A. & ELIASZIW, M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation [see comments]. *Statistics in Medicine*, 11: 1.511-1.519, 1992.

DUNN, G. *Design and Analysis of Reliability Studies: the statistical evaluation of measurement errors*. New York: Oxford University Press, 1989.

ELMORE, J. G. & FEINSTEIN, A. R. A bibliography of publications on observer variability (final installment). *Journal of Clinical Epidemiology*, 45: 567-580, 1992.

EMBRETSON, S. E. & REISE, S. P. *Item Response Theory for Psychologists*. Maheah: Lawrence Erlbaum Associates Publishers, 2000.

FLEISS, J. L. *Statistical Methods for Rates and Proportions*. 2. ed. New York: John Wiley & Sons, 1981.

FLETCHER, R. H. & FLETCHER, S. W. *Epidemiologia Clínica: elementos essenciais*. 4. ed. São Paulo: Artmed, 2006.

FOX, J. Polycor: polychoric and polyserial correlations, function for [R] (0.7-2). CRAN (R-project). 19 Oct. Disponível em: <www.cran.r-project.org/src/contrib/Descriptions/polycor.html>. Acesso em: 19 nov. 2006.

GORSUCH, R. L. *Factor Analysis*. 2. ed. Hillsdal: Lawrence Erlbaum, 1983.

GREENLAND, S. & BRUMBACK, B. An overview of relations among causal modelling methods. *International Journal of Epidemiology*, 31: 1.030-1.037, 2002.

GUILLEMIN, F.; BOMBARDIER, C. & BEATON, D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *Journal of Clinical Epidemiology*, 46: 1.417-1.432, 1993.

HAMBLETON, R. K.; SWAMINATHAN, H. & ROGERS, H. J. *Fundamentals of Item Response Theory*. Newbury Park: Sage, 1991.

HANLEY, J. A. & MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143: 29-36, 1982.

HERDMAN, M.; FOX-RUSHBY, J. & BADIA, X. A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. *Quality of Life Research*, 7: 323-335, 1998.

JÖRESKOG, K. G. & SÖRBOM, D. *LISREL 8 User's Reference Guide*. Chicago: Scientific Software International, 1996.

JÖRESKOG, K. G. & SÖRBOM, D. *Interactive LISREL* 8.7. Chicago: Scientific Software International, 2004.

KLEINBAUM, D. G.; KUPPER, L. L. & MORGENSTERN, H. *Epidemiologic Research: principles and quantitative methods*. New York: Van Nostrand Reinhold Company, 1982.

KLINE, P. *An Easy Guide to Factor Analysis*. 2. ed. New York: Routledge, 1994.

KLINE, R. B. *Principles and Practice of Structural Equation Modeling*. 2. ed. London: The Guilford Press, 2005.

KOLENIKOV, S. Polychoric: the polychoric correlation package for Stata Statistical Software. *Release*, 8(1.4), 2006. Disponível em: <www.unc.edu/~skolenik/stata>. Acesso em: 10 nov. 2006.

KRIEGER, N. & ZIERLER, S. The need for epidemiologic theory. *Epidemiology*, 8: 212-213, 1997.

KRUEGER, R. *Focus Groups: a practical guide for applied research*. 2. ed. London: Sage Publications, 1994.

KRUG, E. G. et al. *World Report on Violence and Health*. Geneva: World Health Organization, 2002.

KUDER, G. F. & RICHARDSON, M. W. The theory of estimation of test reliability. *Psychometrika*, 2: 151-160, 1937.

LANDIS, J. R. & KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 33: 159-174, 1977.

LANDY, F. J. Stamp collection versus science. *American Psychologist*, 35: 1.012-1.027, 1986.

LIN, L. I. K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45: 255-268, 1989.

LOEHLIN, J. C. *Latent Variable Models. An Introduction to Fator, Path and Structural Equation Analysis*. 4. ed. Mahwah: Lawrence Erlbaum Associates Publishers, 2004.

LUIZ, R. R. & STRUCHINER, C. J. *Infererência Causal em Epidemiologia*. Rio de Janeiro: Editora Fiocruz, 2002.

MARUYAMA, G. M. *Basics of Structural Equation Modeling*. Thousand Oaks: Sage Publications, 1998.

MCDOWELL, I. & NEWELL, C. *Measuring Health: a guide to rating scales and questionnaires*. 2. ed. New York: Oxford University Press, 1996.

MELLENBERGH, G. J. Generalized linear item response theory. *Psychological Bulletin*, 115: 300-307, 1994.

MIETTINEN, O. *Theoretical Epidemiology: principles of occurrence research in medicine*. New York: John Wiley & Sons, 1985.

MORAES, C. L. & REICHENHEIM, M. E. Cross-cultural measurement equivalence of the Revised Conflict Tactics Scales (CTS2) Portuguese version used to identify violence within couples. *Cadernos de Saúde Pública*, 18: 783-796, 2002.

MOSER, C. A. & KALTON, G. *Survey Methods in Social Investigation*. 2. ed. London: Heinemann, 1984.

NUNNALLY, J. C. J. & BERNSTEIN, I. *Psychometric Theory*. 2. ed. New York: McGraw-Hill, 1995.

OSBURN, H. G. Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, 5: 343-355, 2000.

PEARL, J. *Causality: models, reasoning, and inference*. Cambridge: Cup, 2000.

PEDHAZUR, E. J. & SCHMELKIN, L. P. *Measurement, Design, and Analysis: an integrated approach*. Illsdale: Lawrence Erlbaum, 1991.

PERNEGER, T. V.; LEPLÈGE, A. & ETTER, J. F. Cross-cultural adaptation of a psychometric instrument: two methods compared. *Journal of Clinical Epidemiology*, 52: 1.037-1.046, 1999.

PESCE, R. P. et al. Adaptação transcultural, confiabilidade e validade da Escala de Resiliência. *Cadernos de Saúde Pública*, 21: 436-448, 2005.

PETT, M. A.; LACKEY, N. R. & SULLIVAN, J. J. *Making Sense of Factor Analysis: the use of factor analysis for instrument development in health care research*. Thousand Oaks: Sage Publication, 2003.

REICHENHEIM, M. E. & MORAES, C. L. Alguns pilares para a apreciação da validade de estudos epidemiológicos. *Revista Brasileira de Epidemiologia*, 1: 131-148, 1998.

REICHENHEIM, M. E. & MORAES, C. L. Buscando a qualidade das informações em pesquisas epidemiológicas. In: MINAYO, M. C. S. & DESLANDES, S. F. (Orgs.) *Caminhos do Pensamento: epistemologia e método*. Rio de Janeiro: Editora Fiocruz, 2002.

REICHENHEIM, M. E. & MORAES, C. L. Psychometric properties of the Portuguese version of the Conflict Tactics Scales: Parent-child Version (CTSPC) used to identify child abuse. *Cadernos de Saúde Pública*, 22: 503-515, 2006.

REICHENHEIM, M. E.; KLEIN, R. & MORAES, C. L. Assessing the physical violence component of the Revised Conflict Tactics Scales when used in heterosexual couples: an item response theory analysis. *Cadernos de Saúde Pública*, 23: 53-62, 2007.

ROTHMAN, K. J. & GREENLAND, S. *Modern Epidemiology*. 2. ed. Philadelphia: Lippincott-Raven Publishers, 1998.

ROTHMAN, K. J. & GREENLAND, S. Causation and causal inference in epidemiology. *American Journal of Public Health*, 95, suppl. 1: S144-S150, 2005.

RUMMEL, R. J. *Applied Factor Analysis*. 4. ed. Evanston: Northwest University Press, 1988.

SACKETT, D. L. et al. *Clinical Epidemiology: a basic science for clinical medicine*. 2. ed. Boston: Little, Brown & Co, 1991.

SCHMITT, D. P. & ALLIK, J. Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, 89: 623-642, 2005.

SHAVELSON, R. J. & WEBB, N. M. *Generalizability Theory: a primer*. Newbury Park: Sage Publications, 1991.

SHERBOURNE, C. D. & STEWART, A. L. The MOS social support survey. *Social Science and Medicine*, 32: 705-714, 1991.

SHROUT, P. E. Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, 7: 301-317, 1998.

SHROUT, P. E. & FLEISS, J. L. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86: 420-428, 1979.

SIJTSMA, K. & MOLENAAR, I. W. *Introduction to Nonparametric Item Response Theory*. Thousand Oaks: Sage Publications, 2002.

SKRONDAL, A. & RABE-HESKETH, S. *Generalized Latent Variable Modeling: multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall, CRC, 2004.

STATACORP. tetrachoric. Tetrachoric correlations for binary variables. Program in Stata Statistical Software, Release 9. College Station (TX): Stata Corporation, 2005.

STEINECK, G. & AHLBOM, A. A definition of bias founded on the concept of the study base. *Epidemiology*, 3: 477-482, 1992.

STRAUS, M. A. et al. The revised Conflict Tactics Scales (CTS2): Development and preliminary psychometric data. *Journal of Family Issues*, 17: 283-316, 1996.

STREINER, D. L. & NORMAN, G. R. *Health Measurement Scales. A Practical Guide to Their Development and Use*. 3. ed. Oxford: Oxford University Press, 2003.

TANNER, W. P. J. & SWETS, J. A. A decision making theory of visual detection. *Psychological Review*, 61: 401-409, 1954.

TEIXEIRA-SALMELA, L. F. et al. Adaptação do Perfil de Saúde de Nottingham: um instrumento simples de avaliação da qualidade de vida. *Cadernos de Saúde Pública*, 20: 905-914, 2004.

UEBERSAX, J. S. The tetrachoric and polychoric correlation coefficients. Statistical Methods for Rater Agreement web site. Disponível em: <http://ourworld.compuserve.com/homepages/jsuebersax/tetra.htm>. Acesso em: 11 nov. 2006.

VAN DER LINDEN, W. J. & HAMBLETON, R. K. Handbook of modern item response theory. New York: Springer, 1996.

VAN STRIEN, T. et al. The validity of dietary restraint scales: comment on Stice et al. (2004). *Psychological Assessment*, 18: 89-94, 2006.

VEREECKEN, C. A.; VAN DAMME, W. & MAES, L. Measuring attitudes, self-efficacy, and social and environmental influences on fruit and vegetable consumption of 11- and 12-year-old children: reliability and validity. *Journal of the American Dietetic Association*, 105: 257-261, 2005.

WAGNILD, G. M. & YOUNG, H. M. Development and psychometric evaluation of the Resilience Scale. *Journal of Nursing Measurement*, 1: 165-178, 1993.

WEED, D. L. Theory and practice in epidemiology. *Annals of the New York Academy of Sciences*, 954: 52-62, 2001.

WHATI, L. H. et al. Development of a reliable and valid nutritional knowledge questionnaire for urban South African adolescents. *Nutrition*, 21: 76-85, 2005.

WILLETT, W. & LENART, E. Reproducibility and validity of food-frequency questionnaires. In: WILLETT, W. (Ed.) *Nutritional Epidemiology*. 2. ed. New York: Oxford University Press, 1998.

WILSON, M. Constructing measures. *An Item Response Modeling Approach*. Mahwah: Lawrence Erlbaum Associates Publishers, 2005.

WILSON, M. M. et al. Appetite assessment: simple appetite questionnaire predicts weight loss in community-dwelling adults and nursing home residents. *American Journal of Clinical Nutrition*, 82: 1.074-1.081, 2005.

WRIGHT, C. M.; PARKINSON, K. N. & DREWETT, R. F. How does maternal and child feeding behavior relate to weight gain and failure to thrive? Data from a prospective birth cohort. *Pediatrics*, 117: 1.262-1.269, 2006.

ZINN, C.; SCHOFIELD, G. & WALL, C. Development of a psychometrically valid and reliable sports nutrition knowledge questionnaire. *Journal of Science and Medicine in Sport*, 8: 346-351, 2005.