Artigo Original

TEXTO & CONTEXTO ENFERMAGEM
TEXT & CONTEXT NURSING TEXTO & CONTEXTO ENFERMERÍA

PROPRIEDADES PSICOMÉTRICAS DE INSTRUMENTOS DE MEDIDAS: BASES CONCEITUAIS E MÉTODOS DE AVALIAÇÃO – PARTE II

Maria Elena Echevarría-Guanilo[1]
Natália Gonçalves[1]
Priscila Juceli Romanoski[1]

[1]Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Enfermagem Florianópolis, Santa Catarina, Brasil.

RESUMO

Objetivo: apresentar e discutir bases conceituais e métodos de avaliação da validade de conteúdo, de construto e de critério dos instrumentos de medida autorrelatada.
Método: estudo teórico embasado nos conceitos do *Consensus-based Standards for the Selection of Health Measurement Instruments* e os avaliados no *Evaluating the Measurement of Patient-Reported Outcomes*, que contempla conceitos de avaliação de instrumentos para apreciação de resultados relatados pelo paciente.
Resultados: a validade é significante para a qualidade metodológica de um instrumento; entretanto, é um critério relativo, visto que depende da adequação do instrumento na qual se pretende medir. Há três diferentes propriedades de medição de validade descritas na literatura: a validade de conteúdo, de construto e de critério.
Conclusões: como a validade é uma importante propriedade, recomenda-se que seja verificada nos estudos que tiveram como objetivo desenvolver novas escalas e naqueles que adaptaram e validaram para outra cultura ou população.

DESCRITORES: Psicometria. Estudos de validação. Inquéritos e questionários. Reprodutibilidade dos testes.

COMO CITAR: Echevarria-Guanilo ME, Gonçalves N, Romaniski PJ. Propriedades psicométricas de instrumentos de medidas: bases conceituais e métodos de avaliação – parte II. Texto Contexto Enferm [Internet]. 2019 [acesso MÊS ANO DIA]; 28: e20170311 Disponível em: http://dx.doi.org/10.1590/1980-265X-tce-2017-0311

PSYCHOMETRIC PROPERTIES OF MEASUREMENT INSTRUMENTS: CONCEPTUAL BASIS AND EVALUATION METHODS - PART II

ABSTRACT

Objective: to present and discuss conceptual bases and methods for evaluating the content, construct and criterion validity of self-reported measuring instruments.
Method: theoretical study based on the concepts of the Consensus-based Standards for the Selection of Health Measurement Instruments and those evaluated in the Evaluating the Measurement of Patient-Reported Outcomes, which includes concepts of instrument assessment to assess patient-reported outcomes.
Results: validity is significant for the methodological quality of an instrument; however, it is a relative criterion, since it depends on the adequacy of the instrument to be measured. There are three different validity measurement properties described in the literature: content, construct and criterion validity.
Conclusions: as validity is an important property, it is recommended that it be verified in studies that aimed to develop new scales and in those that adapted and validated for another culture or population.

DESCRIPTORS: Psychometrics. Validation studies. Surveys and questionnaires. Reproducibility of results.

PROPIEDADES PSICOMÉTRICAS DE INSTRUMENTOS DE MEDIDAS: BASES CONCEPTUALES Y MÉTODOS DE EVALUACIÓN – PARTE II

RESUMEN

Objetivo: presentar y discutir bases conceptuales y métodos de evaluación de validez de contenido, de constructo y de criterio de instrumentos de medida autorrelatada.
Método: estúdio teórico basado en los conceptos del *Consensus-based Standars for the Slection of health Measurement Instrument*s y los evaluados en el *Evaluating the Measurement of Patient-Reported Outcomes*, que considera conceptos de evaluación de instrumentos para apreciación de resultados por el paciente.
Resultados: la validez es significante para la calidad metodológica de um instrumento; entretanto, es un criterio relativo, ya que depende de la adecuación del instrumento en el que se pretende medir. Hay tres diferentes propiedades de medición de validez descritas en la literatura: la validez de contenido, de constructo y de criterio.
Conclusiones: como la validez es una importante propriedad, se recomienda que sea verificada en los estudios que tuvieron como objetivo desarrollar nuevas escalas y en los que adaptaron y validaron para otra cultura o población.

DESCRITORES: Psicometría. Estudios de validación. Encuestas y cuestionarios. Reproducibilidad de los resultados.

INTRODUÇÃO

Nas últimas décadas, os pesquisadores em saúde têm se apropriado da utilização de instrumentos que descrevem o relato do indivíduo sobre sua condição de vida, seu estado de saúde e/ou sobre os determinantes sociais, ou seja, instrumentos que não avaliam diretamente o construto. Os instrumentos de medida válidos e confiáveis têm como vantagem a praticabilidade na aplicação; garantem indicadores confiáveis para a prática clínica, a avaliação em saúde e de pesquisas; exercem influência nas decisões acerca do cuidado, tratamento e/ou intervenções e formulações de programas e políticas de saúde, principalmente instrumentos que já estão disponíveis para a população a ser estudada e não necessitam de uma adaptação cultural.[1]

Na literatura, autores têm recomendado aos pesquisadores que realizem uma extensa busca literária sobre o assunto a ser estudado e os possíveis instrumentos utilizados na população de interesse antes de ser sugerido um novo instrumento, pois a elaboração de instrumentos de medida é dispendiosa não somente em relação a custos e prazos como também para que sejam validados.[2] Entretanto, o consenso sobre as propriedades de medida de instrumentos autorrelatados pelos pacientes – *Patient-Reported Outcome* (PRO) – é necessário, uma vez que essa forma de medida (PRO) é avaliada diretamente pelo paciente sem a interpretação do profissional de saúde; por isso, é amplamente empregada na área da saúde e associa-se à mensuração de estados subjetivos do paciente – por exemplo: como o paciente se sente – ou para medir resultados mais difíceis e de custo elevado – por exemplo: tabagismo, aspectos nutricionais, aspectos físicos, entre outros.[3–4]

Dessa maneira, recomenda-se a adaptação e validação de instrumentos de medidas de saúde ou doença autorrelatadas para uma determinada população;[2] mas é fundamental seguir um rigor metodológico, considerando a validação de três principais propriedades psicométricas: confiabilidade, validade e responsividade.[5]

A confiabilidade de um instrumento permite que se conheça o grau em que o instrumento reproduz de forma consistente os resultados aplicados em diferentes ocasiões; representa uma das principais propriedades de medida, a qual precisa ser avaliada quando se desenvolve uma nova medida, e oferece informações sobre a necessidade de aprimoramento de um instrumento já existente.[3,6] Ademais, a responsividade é outra importante propriedade quando se quer avaliar como o instrumento se comporta em estudos longitudinais ou com diferentes grupos e se o instrumento é capaz de detectar diferenças do construto medido ao longo do tempo.[3,5] Cumpre mencionar que aspectos relacionados à confiabilidade e responsividade são abordados em estudo anterior.[7]

Neste estudo, objetiva-se apresentar e discutir bases conceituais e métodos de avaliação da validade de conteúdo, de construto e de critério dos instrumentos de medida autorrelatada. Essa propriedade representa um importante aspecto para se conhecer o embasamento teórico que fundamenta um instrumento e se este apresenta coerência na medida do construto para o qual foi proposto.[3,5,8] Para clarificar os conceitos imbuídos na validade, apresenta-se os principais métodos de avaliação desta propriedade de medida, a partir da taxonomia do *Consensus-based Standards for the Selection of Health Measurement Instruments* (COSMIN)[4–5] e dos aspectos avaliados no *Evaluating the Measurement of Patient-Reported Outcomes* (EMPRO), com base nos conceitos do Comitê Consultivo Científico do *Medical Outcomes Trust* (MOT).[9]

VALIDADE

A validade refere-se ao grau em que um instrumento realmente mede o que se propõe a mensurar.[3,5,8–9] O estudo da validação de um instrumento implica obter um conjunto de informações provindas de diversas fontes, a fim de definir um julgamento avaliativo sobre o instrumento em questão. Isto é, o construto de interesse precisa ser cuidadosamente diferenciado de outros construtos estreitamente relacionados.[3] Essas informações permitem a identificação se o instrumento criado mede o que pretende medir ou, se quando adaptado, continua mensurando o mesmo construto.[6]

A partir da taxonomia contemplada no *checklist* do COSMIN[5] e dos conceitos contemplados no MOT[9] com base nos aspectos observados no EMPRO, é possível analisar a validade de um instrumento por meio da validade de conteúdo (*Content validity)*, de critério (*Criterion validity*) e de construto (*Construct validity*)[5] (Figura 1). Apesar de muitos investigadores da temática nomearem todos os tipos de validade como validade de construto,[10–11] outros não recomendam,[4] uma vez que, em nível de delineamento e método, essas três formas de validade são diferentes.

Assim, seguindo-se a proposta do COSMIN, estariam contempladas na validade de construto: a validade estrutural (*Structural validity*), a análise por hipóteses *(Hypotheses testing)* e a validade transcultural (*Cross-cultural validity*).[3–5] Destaca-se que a validade estrutural somente deve ser avaliada em instrumentos de saúde multi-item (composto por vários itens). Os aspectos restantes da validade de construto são necessários para todos os instrumentos de medidas em saúde. Deve-se avaliar a validade estrutural para determinar ou confirmar a existência e estrutura das subescalas que serão consideradas nas hipóteses testadas. Já a validade transcultural apenas deve ser avaliada no processo de tradução de um instrumento de medida em saúde.[3–5] Isto porque, o intuito seria avaliar a estrutura proposta (domínios) para análise do construto de interesse, cuja organização apresenta fundamento teórico; assim, quando se adaptam os instrumentos para outras línguas, estes devem ser estudados caso a estrutura de medida original tenha sido preservada.

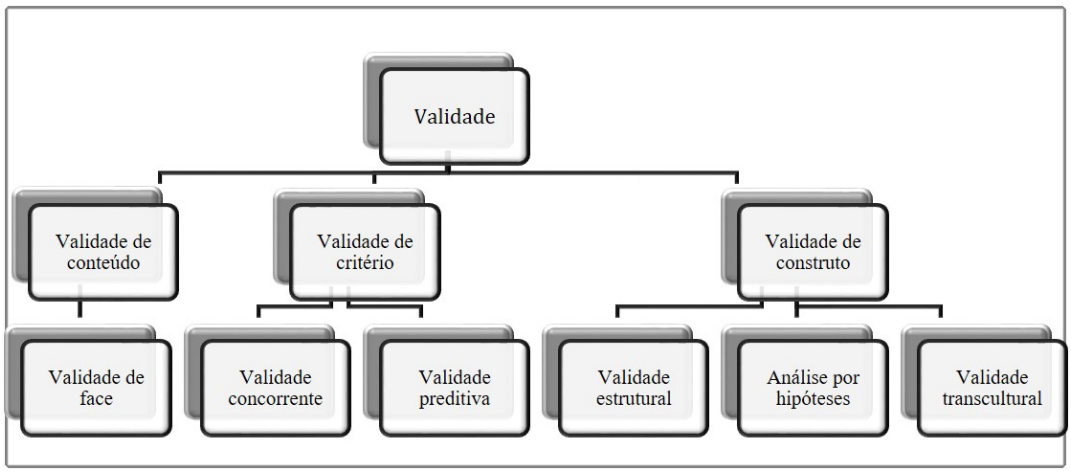

Figura 1 – Taxonomia das propriedades de medida para estudo da validade.

Validade de conteúdo *(Content validity)*

A validade de conteúdo analisa se os componentes do instrumento estão relacionados aos atributos a serem mensurados.[9] Esse tipo de validade avalia o rigor do método a partir do qual foi criado o instrumento e o objetivo da medida para o qual ele foi proposto, além de avaliar o número e relevância dos itens propostos.[12–14] Para sua verificação, é necessário que o instrumento seja submetido à avaliação de ao menos dois juízes, os quais avaliam a relevância de cada item em seus respectivos domínios.[6,14]

É uma propriedade bastante utilizada quando se desenvolve um novo instrumento e nos estudos de adaptação transcultural.[3] Considera-se que o conteúdo de instrumentos reflete com maior especificidade quando incluem no seu desenvolvimento a população para a qual o instrumento estaria sendo direcionado, uma vez que seria contemplada a experiência da condição de saúde que se pretende avaliar.[9]

Embora seja um critério de avaliação importante, a validade de conteúdo nem sempre recebe o merecido destaque no processo de validação, e isto pode ser atribuído ao fato de a validade de conteúdo envolver julgamentos, em sua maior parte subjetivos.[3–5] No entanto, cada vez se reconhece

que a avaliação e o aprimoramento de validade de conteúdo de uma medida representam um passo inicial crítico que requer muito critério,[3,13] porque pode influenciar a obtenção da validade de construto do instrumento. Nesse sentido, é importante ressaltar que uma medida pode apresentar boa validade de construto, sem necessariamente ser adequada em termos de sua validade de conteúdo.[3]

Como exemplo do processo da avaliação de validade de conteúdo, citamos o estudo do processo de adaptação transcultural da versão brasileira do *Caring Ability Inventory* (CAI),[15] no qual a validade de conteúdo (avaliação semântica, idiomática e cultural) foi realizada por um grupo de profissionais conhecedores da temática (*expertises*/comitê de especialistas) e de fundamentos teóricos que sustentam a construção do instrumento e da metodologia de adaptação. Essa constituição do comitê de especialistas contribuiu para resolver, por consenso, as poucas diferenças apresentadas na interpretação da tradução e retrotradução das afirmativas. Os itens que geraram controvérsia (13,5%) foram submetidos à nova rodada de avaliação, sendo incorporados à versão pré-final do CAI após consenso. A concordância final do comitê de especialistas sobre o conteúdo do CAI foi de 86,5%.[15]

É importante destacar que, assim como outras propriedades de medida, a validade de conteúdo não é uma propriedade de valor fixo, isto é, seu valor pode variar de uma população para outra. Entretanto, deve-se considerar a semelhança entre a população estudada e o contexto. Além disso, para o caso em que as evidências sobre a validade de conteúdo sejam consideradas frágeis, um novo estudo sobre as propriedades de medida pode ser necessário.[3]

O sucesso da avaliação da validade de conteúdo relaciona-se com os passos iniciais que contemplam o amplo e profundo conhecimento sobre o construto em questão, os possíveis fatores que poderiam influenciar na avaliação desse construto; e, quais seriam os aspectos a serem considerados que poderiam distinguir o construto analisado de outros existentes; a relevância, a integralidade e o equilíbrio do construto na medida e dos itens do instrumento; a validade de conteúdo por juízes/*experts*; a validade de conteúdo por estratégia qualitativa (exemplo: avaliação por grupos focais) e/ou por estratégia quantitativa (exemplo: pelo Índice de Validade de Conteúdo ou Coeficiente de Kappa) e a validade de conteúdo por meio das medidas derivadas da Teoria de Resposta ao Item (ITR).[3,13]

Como exemplo, pode-se citar o estudo que objetivou validar um instrumento para avaliação da habilidade dos graduandos de enfermagem para mensuração da pressão arterial. Participaram 27 enfermeiros como juízes do estudo que ocorreu em duas etapas: levantamento da literatura para a elaboração do instrumento e posterior validação de conteúdo mediante aplicação do Índice Kappa, que é um indicador de concordância que varia de “menos 1” a “mais 1”, aceitando-se o valor >0,61 (nível bom) e Índice de Validade de Conteúdo, que mede a concordância dos juízes quanto à representatividade dos itens em relação ao conteúdo em estudo, sendo este calculado dividindo-se o número de juízes que avaliaram o item como adequado/necessitando alterações pelo total de juízes que julgaram o item válido. Foi considerado como aceitável (IVC) >0,75, e os autores obtiveram IVC de 0,94 e de Kappa de 0,89.[16]

Validade de face *(Face validity)*

A validade de face ou aparente refere-se à percepção que paciente e/ou pesquisadores têm sobre a medida.[3] Com base na experiência, na área de interesse e/ou nos sujeitos da pesquisa, a validade de face consiste em medir se o instrumento avalia de maneira clara e sem ambiguidades o construto, além de identificar se o conceito medido é aquele pretendido pelo pesquisador.[5–6,17] Também se pode considerar este tipo de validade como uma forma de validade de conteúdo;[6,8,17] por isso, algumas das estratégias de avaliação qualitativa podem ser usadas para ambas. Embora se utilize amplamente a validade de face, ela é apontada como uma forma de avaliação casual e numericamente menos consistente.[18]

Salienta-se que um importante aspecto a considerar na avaliação da validade de face é definir qual a população-alvo, a condição de saúde e a(s) pessoa(s) encarregada(s) de realizarem as avaliações. Dessa forma, como esse instrumento é empregado por profissionais de saúde, os participantes no processo de avaliação da validade de face deverão contar com a participação desses profissionais.[3]

Pode-se citar, como exemplo, o processo de validação da versão brasileira da *Burn Specific Health Scale-Brief* (BSHS-B-Br) cuja validade de face e conteúdo foi determinada por consenso de uma equipe multiprofissional, submetida após as etapas do processo de tradução independentes e da retrotradução.[19]

Validade de critério *(Criterion validity)*

Validade de critério refere-se ao grau em que o instrumento produz resultados semelhantes aos de outros instrumentos/equipamentos já existentes e válidos (*Gold Standard*) para avaliar o mesmo construto.[4,8,20]

A validade de critério pode ser concorrente (*Concurrent validity*), quando a medida produzida pelo instrumento testado é similar ou pode substituir aquela considerada como padrão-ouro, quando a avaliação da medida pelos dois instrumentos ocorrem simultaneamente, ou preditiva *(Predictive validity)*, quando a medida produzida prediz algum evento futuro e a coleta de dados ocorre em momentos diferentes.[3] Outro grupo de autores acrescentam que a validade de critério preditiva é um aspecto também da validade de construto.[8]

A desvantagem da validade de critério é que as medidas padrão-ouro podem não ser fáceis de estabelecer ou, ainda, não se encontrarem disponíveis.[20] A falta de medidas de critério ou de referência restringe a avaliação dessa propriedade psicométrica quase que exclusivamente à realização de estudos de versões resumidas/abreviadas dos instrumentos, empregando a versão original como uma medida *Gold Standard* ou de critério.[14] Para ambas as formas de avaliação da validade de critério, embora as hipóteses possam ser raramente declaradas de maneira formal, é importante destacar que sempre há uma hipótese implícita.[3] Então, quanto maior a clareza na apresentação da hipótese (hipótese previa), maior será a clareza na interpretação dos dados obtidos.

A validade de critério pode ser estudada por meio da aplicação do coeficiente de correlação de Pearson entre duas medidas (medidas contínuas e critério contínuo/*continuous measure and continuous criterion*), aplicação de Teste de Regressão Múltipla, principalmente para a identificação de validade preditiva (medidas contínuas e critério contínuo/*Continuous measure and continuous criterion*), teste de sensibilidade e especificidade (medidas nominais e critério nominal) e por meio do Teste t de Student ou da área abaixo da curva *receiver operating characteristic* (ROC) (medida contínua e critério nominal/*Continuous measure and nominal criterion)*.[3,5,8]

Como exemplo de validade de critério, cita-se o estudo de validação e confiabilidade dos itens da escala de dor da versão chinesa *interRAI Community Health Assessment* (CHA)[21] – uma escala composta por quatro itens que avaliam frequência, intensidade, consistência e experiência da dor e, que, na sua aplicação, pode ser reduzida à aplicação de dois itens para avaliar dor (contemplando apenas frequência e intensidade). Para o estudo da validade de critério concorrente, foram aplicadas as versões Inventário Breve de Dor – Versão Chinesa BPI-C *(Brief Pain Inventory-Chinese version)* e Escala de avaliação verbal de cinco pontos – VRS *(Five-point verbal rating scale)* juntamente com a versão do *interRAI-CHA* (versão de Hong Kong). A análise de validade concorrente resultou na correlação entre a escala de dor, aos quatro itens de dor do *interRAI-CHA* e o BPI-C de 0,52 e 0,66, respectivamente ($p<0,05$); e correlações da escala de dor, os quatro itens de dor do *interRAI-CHA* e a VRS de cinco pontos de 0,47 e 0,67, respectivamente ($p<0,05$). Os resultados mostraram correlações significativas com níveis aceitáveis de validade concorrente.[21]

Considera-se importante destacar que, segundo o referencial teórico utilizado neste manuscrito, nem todas as medidas em saúde podem ser avaliadas por meio da validade de critério, principalmente aquelas autorrelatadas pelos pacientes (PRO), porque muitos atributos não são aparentes e medidas *gold standard* são inexistentes.[5,9] Medidas de construtos semelhantes podem mais facilmente ser identificadas, porém; deve-se esclarecer que medidas semelhantes ou relacionadas avaliam a validade de construto, não a validade de critério.

Validade de construto *(Construct validity)*

Validade de construto refere-se ao grau em que um instrumento está medindo o construto de interesse.[3,14] Ela examina a relação teórica dos itens do instrumento e os conceitos contidos na teoria e fornece evidências para a interpretação dos valores propostos, com base em relações hipotéticas de associação do construto no que concerne a outros construtos.[3,5,14] E, como mencionado anteriormente, a validade de construto contempla: a validade estrutural (S*tructural validity*), a análise por hipóteses *(Hypotheses testing)* e a validade transcultural (*Cross-cultural validit*y).[3–5]

Essa validade é a mais complexa e difícil de ser determinada, uma vez que estuda o grau em que as pontuações da medida se relacionam com outras pontuações de construtos conceitualmente relacionados,[3] isto é, esta propriedade está relacionada à habilidade do instrumento para confirmar as hipóteses esperadas,[22] as quais contemplam a relação numérica e se traduz em uma explicação conceitual.

Métodos comuns para obter a confirmação da validade de construto incluem: teste de correlação entre as medidas de instrumentos que avaliam construtos afins (validade convergente), ou pelo exame lógico das relações que deveriam existir com outras medidas e/ou padrões de valores para grupos que supostamente devem divergir dos valores relacionados ao construto (validade discriminante ou divergente).[3,14,18]

Validade estrutural *(Structural validity)*

Outro aspecto da validade de construto está relacionado com a avaliação da dimensionalidade do instrumento ou a validade estrutural *(Structural validity)*, a qual é definida na medida em que a estrutura de um instrumento multi-item reflete de forma adequada a multidimensionalidade da hipótese do construto que pretende ser medido[3] ou se todos os itens que compõem o instrumento avaliam uma ou mais variáveis latentes conforme a proposta original.

Especificamente para instrumentos compostos por vários itens, a análise fatorial é utilizada para avaliar a validade do construto, por meio da qual se identifica a estrutura de correlações entre os diferentes itens que compõem o instrumento. As equações resultantes desta análise podem ser interpretadas como agrupamentos de itens, os quais, são representados em um(a) ou mais fatores ou dimensões.[14] Em etapa inicial da proposta de um instrumento, o objetivo é evidenciar o número de construtos contidos no instrumento (unidimensional ou multidimensional), assim como avaliar a importância de manter ou retirar componentes (itens ou grupo de itens).[3–4]

É importante distinguir entre análise fatorial exploratória e confirmatória. Enquanto metodologia estatística é a mesma, as duas análises diferem na interpretação dos resultados. A análise fatorial exploratória refere-se à identificação de fatores potenciais contidos no instrumento e não requer conhecimento prévio sobre a estrutura postulada do instrumento. O objetivo principal é o de gerar hipóteses a serem testadas em estudos planejados para esse fim. A análise fatorial confirmatória refere-se ao teste das hipóteses geradas por análises exploratórias. Outra função dessa análise fatorial confirmatória é a de confirmar que os fatores e/ou a estrutura interna de um instrumento de medida em suas versões adaptadas para idiomas e culturas diferentes não tiveram a estrutura correlacional entre os itens modificada pela adaptação do instrumento.[12,14]

Embora se encontrem várias sugestões na literatura sobre o tamanho adequado de uma amostra para a análise fatorial, a maior parte dessas sugestões não se fundamenta em estudos teóricos.[14] Alguns estudos para situações específicas têm demonstrado que amostras maiores que 50 e menores que 100 podem ser suficientemente representativas e permitirem avaliar as propriedades métricas de instrumentos direcionados para avaliar construtos sociais.[23] No entanto, de modo geral, para obterem-se estimadores estáveis e um poder alto do teste estatístico, grandes amostras podem ser necessárias; logo, costuma-se sugerir que se coletem centenas de observações.[23]

Como exemplo, cita-se o estudo que verificou a validade de construto por meio da análise fatorial exploratória da versão reduzida da *Depression Anxiety Stress Scale-21* (EADS-21)[24] aplicada a adolescente na versão brasileira. A análise fatorial exploratória da EADS-21 efetuou-se a partir da estrutura tridimensional proposta pelo autor original; contudo, resultou em alguns itens com cargas semelhantes ou mais fortes em construtos centrais. Realizou-se análise ortogonal Varimax para dois fatores, e os itens correspondentes à ansiedade e ao estresse (carga fatorial variou entre 0,47 a 0,64) se agruparam em um único fator, e à depressão em um segundo fator (carga fatorial variou entre 0,52 a 0,77), o que resultou na melhor adequação de todos os 21 itens, com cargas fatoriais mais altas em seus respectivos construtos.[24]

É importante destacar que, quando se analisa a avaliação desta propriedade de medida, em instrumentos em processo de adaptação transcultural, esta etapa poderá definir a necessidade de retomar adequações que foram realizadas em etapas iniciais do processo, as quais resultaram, por exemplo, em grandes mudanças na organização estrutural do instrumento.[25]

Análise por hipóteses *(Hypothesis testing)*

Trata-se do estudo direcionado para avaliar o poder do teste psicológico em discriminar ou predizer um critério externo ao construto avaliado.[6]

A análise por hipóteses pode ser avaliada por meio da validade de construto convergente (*Convergent validity*), validade de construto divergente ou discriminante *(Discriminant validity)*, validade por grupos conhecidos (*Discriminative validity*) e abordagem Matriz Multitraço-multimétodo (*Multitrait-multimethod Matrix*).[3]

A validade de construto convergente *(Convergent validity)* refere-se à correlação linear do instrumento com o construto com o qual, conceitualmente, deveria estar correlacionado; todavia, a medida correlacionada não seria considerada uma medida padrão-ouro (*gold standard).*[3] Assim, a hipótese a ser testada seria a presença de correlações moderadas a altas ou muito altas, entre os construtos dos quais teoricamente se espera identificação de correlações e que, em delineamentos de estudos de coorte, por exemplo, poderiam explicar a variação da medida do construto estudado.

A validade de construto divergente ou discriminante *(Discriminant validity)* analisa a diferença entre o construto estudado e outro com o qual teoricamente não deveria apresentar correlação.[3,26–27] A hipótese a ser testada seria a ausência ou fraca correlação entre os construtos, o que sugeriria que as dimensões que compõem cada instrumento estariam medindo construtos diferentes[3] ou que o instrumento mediria aspectos distintos ou não relevantes para o que se pretende medir.[3,26–27] Por exemplo, se o pesquisador está desenvolvendo uma escala de percepção de função física, ao compará-la com uma escala psicossocial já validada, baseando-se nas teorias que fundamentam as duas escalas, ele espera que a correlação linear entre as duas escalas seja baixa e, dessa forma, estabeleça validade divergente.

É preciso ter cuidado durante a análise de validade divergente, pois é possível que se observe uma correlação linear alta quando ela não existe. Afinal, a correlação se apresenta em decorrência das duas escalas estarem relacionadas a um fator comum, o qual influencia a resposta de ambas as escalas (por exemplo, a idade da pessoa).[12]

Como formas de análises, tanto para a validade de construto convergente quanto divergente, comumente se utilizam parâmetros de coeficiente de correlação de Pearson e modelos de regressão múltipla.[3] É importante destacar a diferença entre validade convergente e validade concorrente, já que apenas nesta última se requer uma medida padrão-ouro *(gold standard)*. E ao se tratar desta, destaca-se que será necessário aplicá-la simultaneamente à medida em estudo.[3]

Para a avaliação de validades convergente e divergente, recomenda-se que as hipóteses sobre a relação entre as variáveis estudadas e as medidas de comparação sejam determinadas antes da

coleta de dados. Entre as diversas propostas de categorização dos coeficientes de correlação linear, destacam-se: muito baixa [0,0 a 0,25], baixa [0,26 a 0,49], moderada [0,50 a 0,69], alta [0,70 a 0,89] e muito alta [0,90 a 1,0][28] e <0,30 baixa;0,30 – 0,50: moderada; e >0,50: alta.[29] Essas classificações podem ser usadas na interpretação de correlações positivas e negativas (tomando-se o valor absoluto). Quanto maior o valor da correlação entre as medidas analisadas, maior será a indicação de validade convergente; quanto menor a correlação, maior será é a evidência de validade divergente.[12]

Considera-se importante que seja feita a definição prévia destes parâmetros de avaliação da força das correlações e a consideração, na escolha dos parâmetros e do tipo de variável ou construto que está sendo estudado, visto que variáveis sociais poderão apresentar correlações mais fracas e variáveis como dosagem de marcadores fisiológicos poderão apresentar correlações altas.

Como exemplo da avaliação de validade de construto convergente, pode-se citar o estudo de validação da versão brasileira do *Quality of Recovery-40 Item* (QoR-40) em pacientes submetidos à prostatectomia radical.[30] Para tanto, obteve-se o coeficiente de correlação de Pearson, verificando-se a correlação entre o QoR-40 com a Escala Visual Analógica (EVA) e o *36-Item Short-Form Health Survey Version 2.0* (SF-36), aplicados em três momentos (pré-operatório, primeiro retorno e segundo retorno). Exemplificando, identificaram-se correlações moderadas entre o domínio do estado emocional do QoR-40 com os domínios do SF-36: vitalidade (r=0,52; p<0,05), aspectos emocionais (r=0,50; p<0,05), aspectos sociais (r=54; p<0,05) e saúde mental (r=60; p<0,05) no pré-operatório. Correlações moderadas, no primeiro retorno, entre o domínio estado emocional do *QoR-40* e os domínios do *SF-36*: capacidade funcional (r=0,49; p<0,05), aspectos físicos (r=0,52; p<0,05), dor (r=0,45;p<0,05), estado geral da saúde (r=0,48; p<0,05), vitalidade (r=0,59; p<0,05), aspectos emocionais (r=0,50;p<0,05), aspectos sociais (r=0,61; p<0,05) e saúde mental (r=0,69; p<0,05) no pré-operatório. Mantiveram-se correlações moderadas em ambas as aplicações dos instrumentos no primeiro e segundo retorno. Ainda, foram identificadas correlações entre o QoR-40 e EVA fracas no pós-operatório (r=0,38; p<0,05) e fortes no primeiro (r=0,76; p<0,05) e segundo retornos (r=0,85; p<0,05).[30]

Destaca-se que a força da correlação é mais importante ao se avaliar a validade de construto de um instrumento do que o sentido da correlação entre a medida do instrumento que está sendo adaptado e do instrumento escolhido para testar a hipótese.

A abordagem denominada de Matriz Multitraço-multimétodo *(Multitrait-multimethod Matrix Approach)* foi proposta para avaliar, simultaneamente, a validação convergente e discriminante do instrumento.[3,12] Nesta técnica, dois ou mais métodos (instrumentos diferentes), ou dois ou mais traços diferentes, não relacionados, usualmente, serão avaliados simultaneamente por dois ou mais métodos. A matriz é construída com as subescalas de cada instrumento apresentadas tanto nas colunas como nas linhas, e as correlações lineares entre subescalas são apresentadas em cada célula da matriz. Dessa forma, correlações relacionadas à validade convergente e à divergente são facilmente identificadas. Este tipo de matriz pode, também, ser usado para apresentar correlações entre subescalas de um mesmo instrumento aplicado em dois períodos diferentes com a finalidade de estudar a confiabilidade do instrumento.[12]

A validade convergente, analisada por meio da aplicação, é satisfeita se a correlação entre um item e a dimensão a que pertence for superior a 0,30 e em estudos finais superiores a 0,40.[12] A validade discriminante, com a utilização da Matriz Multitraço-multimétodo, verifica a porcentagem de vezes em que a correlação de um item com uma dimensão, à qual o item pertence, é estatisticamente maior do que sua correlação com a dimensão à qual não pertence (ajuste). Assim, valores de ajuste próximos a 100% constatam a validade discriminante do instrumento.

Em estudo que teve como objetivo validar o Módulo Fibrose Cística para crianças e adolescentes (versão *self*), do instrumento de mensuração de qualidade de vida relacionada à saúde DISABKIDS® para brasileiros,[31] realizou-se a validade de construto a partir da validade convergente e discriminante.

Assim, para a análise das correlações entre os itens e as dimensões, foi utilizada a Matriz Multitraço-Multimétodo (MTMM), a qual permitiu obter informações sobre a alocação dos itens na escala e a porcentagem de ajuste para cada um dos itens *(scale fit)*. Para a análise de validade convergente, este estudo constatou correlações entre cada item e sua respectiva dimensão, a qual, na maioria das vezes, foi superior a 0,40, e apenas os itens 5 (r=0,26) e 6 (r=0,37) apresentaram correlações mais baixa. Entretanto, o item 6 apresentou valor considerado satisfatório. Em conclusão, os autores descrevem que a validade de construto foi satisfatória em razão de os valores de validade convergente e divergente também serem satisfatórios (ajuste de 100%).[31]

A validade discriminativa entre grupos conhecidos *(Known-groups)*, também chamada de validade de contraste (*Contrast validity*), é uma forma de validade que tem por objetivo identificar diferenças entre grupos nos quais teoricamente espera-se constatar essas diferenças, isto é, toma-se como base a hipótese de que grupos de indivíduos que são entendidos como diferentes em relação ao construto a ser medido produzem valores diferentes quando se aplica o instrumento.[3,6,32] O objetivo é avaliar se o instrumento testado discrimina as diferenças entre os grupos, por exemplo, doentes e não doentes, sintomáticos e não sintomáticos.[32] É importante lembrar que esse tipo de validade avalia a presença de diferença nas medidas obtidas entre os grupos, e não se a medida realmente mede o construto pretendido.

Como exemplo, cita-se o estudo da validade discriminativa por grupos conhecidos da *The Older Persons and Informal Caregivers Survey Minimum Data Set* (TOPICS-MDS).[33] Foram identificadas médias maiores em pessoas sem demência e depressão, e sem presença de tontura com quedas respectivamente. A partir de análises de correlação linear, foi possível constatar hipóteses de diferenças entre as médias, sendo maior para as pessoas casadas e que viviam de forma independente e com educação de nível universitário ($p<0,05$), ajustados para idade e sexo. A partir dos resultados, os autores referem que o TOPICS-MDS apresenta propriedade discriminativa entre grupos conhecidos, constituindo-se como um instrumento com grande potencial para utilização em estudos de intervenção e que pretendam estudar diferenças entre subgrupos da população-alvo.[33]

Validação transcultural *(Cross-Cultural Validity)*

Pesquisas em saúde têm se tornado cada vez mais multiculturais e de âmbito internacional, o que tem gerado grande preocupação entre os pesquisadores, no intuito de preservar a originalidade da medida dos instrumentos, assim como garantir a qualidade para uso em diversas culturas.[3]

Nesse sentido, a taxonomia do COSMIN[4] e os conceitos avaliados no EMPRO[7–8] abordam os passos requeridos na tradução e adaptação transcultural, de forma que se garanta a adequação e equivalência (individual e coletiva), em relação à versão original.[3]

Assim, os autores recomendam os seguintes passos:

a) equivalência conceitual e técnica (*Conceptual and Technical equivalence*) – representa o primeiro passo para optar pela adaptação. Esse passo envolve o conhecimento amplo do instrumento de interesse e a análise criteriosa da equivalência conceitual e aplicabilidade na prática,[3] isto é, se o construto medido seria um construto relevante para a cultura para o qual se pretende adaptar o instrumento. Para tanto, é possível adotar como estratégias a opinião de especialistas, as análises a partir de revisão de literatura e a apreciação da população-alvo. Autores apontam que muitos pesquisadores deixam de fazer esta avaliação inicial passando a contar com avaliações de equivalência conceitual apenas após a tradução ter sido concluída.[34] Esta etapa é relevante por permitir que os pesquisadores possam identificar, ainda em etapa inicial, a semelhança com a versão original ou a necessidade de promover mudanças, seja por adequações na tradução, pela necessidade de remoção de itens[3–4] ou pela não aplicabilidade do instrumento na realidade pretendida;

b) equivalência semântica – é o processo de tradução/adaptação (*Semantic equivalence*). Passo que contempla a tradução propriamente dita do idioma original para o idioma-alvo, composto

por quatro fases: tradução (*Forward translation*), síntese (*Synthesis*), retrotradução (*Back-translation*) e consenso (*Reconciliation*);[4]

c) pré-teste da versão pré-final (*Pretesting*)[3–4] – trata-se da avaliação relacionada à compreensão que o público-alvo tem sobre as partes que compõem o instrumento, ou seja, a avaliação da equivalência semântica e conceitual do instrumento;[35]

d) ensaio da versão final do instrumento *(Field testing of the final instrument)* – etapa que requer o delineamento de pesquisa que é adequado ao tipo de medida, visando a alcançar dois objetivos principais: avaliar em que medida as propriedades de mensuração da nova escala atendem a padrões de qualidade habitual para a aplicação pretendida, conforme a proposta do instrumento original; estudar outros aspectos importantes que contribuam com a constatação da equivalência da versão em idioma distinto do original.[3–4]

Logo, compreende-se que o desenvolvimento da validação transcultural requer profunda discussão em decorrência das particularidades de cada etapa, as quais poderão ser discutidas em um estudo futuro.

CONCLUSÃO

Neste estudo percebeu-se que testar as diferentes formas de validade de um instrumento de medida consiste em um processo metodológico rigoroso que permite identificar evidências não somente em relação ao que realmente está sendo medido mas também ao que pesquisador busca avaliar. Para tanto, o consenso sobre as propriedades de medida de instrumentos que incorporam a perspectiva do paciente PRO torna-se necessário.

As validades de face e conteúdo são avaliações de natureza mais qualitativas do que quantitativas, visto que estas estão embasadas, sobretudo, em julgamentos empíricos, porque não há métodos objetivos que garantam que um instrumento avalie adequadamente o construto para o qual foi construído.

Considerando que técnicas estatísticas têm sido desenvolvidas para constatar hipóteses de forma mais quantitativa para a validade de construto e de critério dos instrumentos de medidas de resultados percebidos pelos indivíduos, cabe aos pesquisadores a procura do domínio de conceitos teóricos metodológicos que permitam um delineamento de pesquisa adequado para constatar propriedades de medida mais adequadas para o instrumento de interesse.

REFERÊNCIAS

1. Coluci MZO, Alexandre NMC, Milani D. Construção de instrumentos de medida na área da saúde. Ciênc Saúde Coletiva [Internet]. 2015 [acesso 2018 Fev 19];20(3):925-36. Disponível em: https://dx.doi.org/10.1590/1413-81232015203.04332013
2. Epstein J, Santo RM, Guillemin F. A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. J Clin Epidemiol [Internet]. 2015 [acesso 2017 Mar 03];68(4):435-41. Disponível em: https://dx.doi.org/10.1016/j.jclinepi.2014.11.021
3. Polit DF, Yang FM. Measurement and the measurement of change. Philadelphia(US): Wolters Kluwer; 2016.
4. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol [Internet]. 2010 [acesso 2017 Mar 01];63(7):737-45. Disponível em: https://dx.doi.org/10.1007/s11136-010-9606-8
5. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol, DL, et al. COSMIN checklist manual. [Internet]. 2012 [acesso 2017 Mar 01]. Disponível em: http://www.cosmin.nl/cosmin_checklist.html

6. Pasquali L. Psicometria. Teoria dos testes na psicologia e na educação. 5a ed. Petrópolis, RJ(BR): Editora Vozes; 2013.

7. Echevarria-Guanilo ME; Goncalves N; Romanoski, PJ. Psychometric properties of measurement instruments: Conceptual bases and evaluation methods - Part I. Texto Contexto Enferm [Internet]. 2017 [acesso 2017 Mar 01]; 26(4):e1600017. Disponível em: https://dx.doi.org/10.1590/0104-07072017001600017

8. Valderas JM, Ferrer M, Mendívil J, Garin O, Rajmil L, Herdman M, et al. Development of EMPRO: A tool for the standardized assessment of patient-reported outcome measures. Value Health [Internet]. 2008 [acesso 2017 Mar 01];11(4):700-8. Disponível em: https://dx.doi.org/10.1111/j.1524-4733.2007.00309

9. Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res [Internet]. 2002 [acesso 2017 Mar 01];11(3):193-205. Disponível em: https://www.ncbi.nlm.nih.gov/pubmed/12074258

10. Anastasi A. Psychological testing. New York, NY(US): Macmillan; 1988.

11. Messick S. Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. ETS Res Report Series [Internet]. 1994 [acesso 2017 Mar 01]; (2). Disponível em: https://dx.doi.org/10.1002/j.2333-8504.1994.tb01618.x

12. Fayer PM, Machin D. Quality of Life. Measurement in nursing and health research. 5a ed. New York, NY(US): Springer; 2007.

13. Strauss ME, Smith GT. Construct validity: Advances in theory and methodology. Annu Rev Clin Psychol [Internet]. 2009 [acesso 2017 Mar 01]; 5:1-25. Disponível em: https://dx.doi.org/10.1146/annurev.clinpsy.032408.153639

14. Waltz CF, Strickland OL, Lenz ER. Measurement in Nursing and Health Research. 5a ed. New York, NY(US): Springer; 2017.

15. Rosanelli CLSP, Silva LMGD, Gutiérrez MGDR. Adaptação transcultural do Caring Ability Inventory para a língua portuguesa. Acta Paul Enferm [Internet]. 2016 [acesso 2017 Mar 01]; 29(3):347-54. Disponível em: https://dx.doi.org/10.1590/1982- 0194201600048

16. Tibúrcio M P, Melo GDSM, Balduíno LSC, Costa IKF, Dias TYDAF, Torres GDV. Validation of an instrument for assessing the ability of blood pressure measurement. Rev Bras Enferm [Internet]. 2014. [acesso 2017 Mar 01];67(4):581-7. Disponível em: https://dx.doi.org/10.1590/0034-7167.2014670413

17. Bölenius K, Brulin C, Grankvist K, Lindkvist M, Söderberg J. A content validated questionnaire for assessment of self reported venous blood sampling practices. BMC Res notes [Internet]. 2012 [acesso 2017 Mar 01];5(1):39. Disponível em: https://dx.doi.org/10.1186/1756-0500-5-39

18. Bolarinwa AO. Principles and methods of validity and reliability testing of Questionnaires Used in Social and Health Science Researches. Niger Postgrad Med J [Internet]. 2015 [acesso 2017 Mar 01];22(4):195-201. Disponível em: https://dx.doi.org/10.4103/1117-1936.173959

19. Piccolo MS, Gragnani A, Daher RP, Tubino Scanavino M, Brito MJ, Ferreira LM. Validation of the Brazilian version of the Burn Specific Health Scale-Brief (BSHS-B-Br). Burns [Internet]. 2015 [acesso 2017 Mar 01];41(7):1579-86. Disponível em: https://dx.doi.org/10.1016/j.burns.2015.04.016

20. Engel RJ, Schutt RK. Measurement. The practice of research in social work. 3a ed. Thousand Oaks, CA(US): Sage; 2013.

21. Liu JY, Chi I, Chan KS, Lai CK, Leung AY. The reliability and validity of the pain items of the Hong Kong version interRAI community health assessment for community-dwelling elders in Hong Kong. J Clin Nurs [Internet]. 2015 [acesso 2017 Mar 01];15(24):2352-4. Disponível em: https://dx.doi.org/10.1111/jocn.12885

22. Wong KL, Ong SF, Kuek TY. Constructing a survey questionnaire to collect data on service quality of business academics. Eur J Soc Sci [Internet]. 2012 [acesso 2017 Mar 01]; 29:209-21. Disponível em: http://eprints.utar.edu.my/860/1/6343.pdf

23. Sapnas KG, Zeller RA. Minimizing sample size when using exploratory factor analysis for measurement. J Nurs Meas [Internet]. 2002 [acesso 2017 Mar 01];10(2):135-54. Disponível em: https://www.ncbi.nlm.nih.gov/pubmed/12619534

24. Silva HAD, Passos MHPD, Oliveira VMAD, Palmeira AC, Pitangui ACR, Araújo RCD. Short version of the Depression Anxiety Stress Scale-21: is it valid for Brazilian adolescents? Einstein (São Paulo) [Internet]. 2016 [acesso 2017 Mar 01];14(4):486-93. Disponível em: https://dx.doi.org/10.1590/S1679-45082016AO3732

25. Beaton DE, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine [Internet]. 2000 [acesso 2017 Mar 01];25(24):3186-91. Disponível em: https://dx.doi.org/10.1097/00007632-200012150-00014

26. Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. Am J Health Syst Pharm [Internet]. 2008 [acesso 2017 Mar 01];65(23):2276-84. Disponível em: https://dx.doi.org/10.2146/ajhp070364

27. Reichenheim ME; Hökerberg YHM; Moraes CL. Assessing construct structural validity of epidemiological measurement tools: a seven-step roadmap. Cad Saude Publica [Internet]. 2014 [acesso 2018 Mar 01];30(5):927-39. Disponível em: https://dx.doi.org/10.1590/0102-311x00143613

28. Plichta EB, Kelvin EA. Munro's Statistical methods for health care research. 6a ed. Philadelphia (US): Lippincott; 2013.

29. Ajzen I. Understanding attitudes and predicting social behavior. New Jersey, NJ(US): Prentice-Hall; 1998.

30. Eduardo AHA, Santos CB, Carvalho AMP, Carvalho ECD. Validation of the Brazilian version of the Quality of Recovery-40 Item questionnaire. Acta Paulista de Enferm [Internet]. 2016 [acesso 2017 Mar 01];29(3):253-9. Disponível em: https://dx.doi.org/10.1590/1982-0194201600036

31. Santos DMSS, Deon KC, Bullinger M, Santo CB. Validade do instrumento DISABKIDS® - Módulo Fibrose Cística para crianças e adolescentes brasileiros. Rev Latino-am Enfermagem [Internet]. 2014 [acesso 2017 Mar 01];22(6):819-25. Disponível em: https://dx.doi.org/10.1590/0104-1169.3450.2485

32. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. Protocol of the COSMIN study: Consensus-based Standards for the selection of health Measurement Instruments. BMC Med Res Methodol [Internet]. 2006 [acesso 2017 Mar 01];6:2. Disponível em: https://dx.doi.org/10.1186/1471-2288-6-2

33. Hofman CS, Lutomski JE, Boter H, Buurman BM, de Craen AJM, Donders R, et al. Examining the construct and known-group validity of a composite endpoint for The Older Persons and Informal Caregivers Survey Minimum Data Set (TOPICS-MDS); A large-scale data sharing initiative. PLoS ONE [Internet]. 2017 [acesso 2017 Mar 01];12(3):e0173081. Disponível em: https://dx.doi.org/10.1371/journal.pone.0173081

34. Herdman M, Fox-Rushby J, Badia X. "Equivalence" and the translation and adaptation of health-related quality of life questionnaires. Qual Life Res [Internet]. 1997 [acesso 2017 Mar 01];6: 237-47. Disponível em: https://www.ncbi.nlm.nih.gov/pubmed/9226981

35. Nápoles-Springer AM, Santoyo-Olsson J, O'Brien H, Stewart AL. Using cognitive interviews to develop surveys in diverse populations. Med Care [Internet]. 2006 [acesso 2017 Mar 01]; 44(11 Suppl 3):s21-s30. Disponível em: https://dx.doi.org/10.1097/01.mlr.0000245425.65905.1d

NOTAS

CONTRIBUIÇÃO DE AUTORIA
Concepção do estudo: Echevarria-Guanilo ME.
Coleta de dados: Echevarria-Guanilo ME, Gonçalves N.
Análise e interpretação dos dados: Echevarria-Guanilo ME, Gonçalves N, Romaniski PJ.
Discussão dos resultados: Echevarria-Guanilo ME, Gonçalves N, Romaniski PJ.
Redação e/ou revisão crítica do conteúdo: Echevarria-Guanilo ME, Gonçalves N.
Revisão e aprovação final da versão final: Echevarria-Guanilo ME, Gonçalves N, Romaniski PJ.

CONFLITO DE INTERESSES
Não há conflito de interesses.

HISTÓRICO
Recebido: 31 de março de 2019
Aprovado: 16 de abril de 2019

AUTOR CORRESPONDENTE
Maria Elena Echevarría-Guanilo
elena_meeg@hotmail.com