

Adaptação para a língua portuguesa e validação do *Lunney Scoring Method for Rating Accuracy of Nursing Diagnoses**

ADAPTATION TO THE PORTUGUESE LANGUAGE AND VALIDATION OF THE *LUNNEY SCORING METHOD FOR RATING ACCURACY OF NURSING DIAGNOSES*

ADAPTACIÓN A LA LENGUA PORTUGUESA Y VALIDACIÓN DEL *LUNNEY SCORING METHOD FOR RATING ACCURACY OF NURSING DIAGNOSES*

Diná de Almeida Lopes Monteiro da Cruz¹, Cassiana Mendes Bertencello Fontes², Cristiane Giffoni Braga³, Márcia Paschoalina Volpato⁴, Suely Lopes de Azevedo⁵

RESUMO

O *Lunney Scoring Method for Rating Accuracy of Nursing Diagnoses* (LSM) é uma escala de diferencial semântico que foi desenvolvida por Lunney para estimar a acurácia dos diagnósticos de enfermagem. O objetivo deste estudo foi adaptar o LSM para a língua portuguesa e avaliar as suas propriedades psicométricas. A escala original foi traduzida para o português, revertida para o inglês e as duas versões em inglês foram comparadas para ajustar a versão em português que passou a ser denominada Escala de Acurácia de Diagnóstico de Enfermagem de Lunney – EADE. Quatro enfermeiras foram orientadas sobre a EADE e a aplicaram em 159 diagnósticos formulados para 26 pacientes de três estudos primários com base nos registros de entrevista e exame físico de cada paciente. Os índices *Kappa* de Cohen mostraram ausência de concordância entre as avaliadoras, o que indica que o instrumento adaptado não tem confiabilidade satisfatória. Em virtude desse resultado, não foi realizada estimativa de validade.

DESCRIPTORES

Diagnóstico de enfermagem.
Avaliação em enfermagem.
Psicometria.

ABSTRACT

The *Lunney Scoring Method for Rating Accuracy of Nursing Diagnoses* (LSM) is a semantic differential scale developed by Lunney to rate the accuracy of nursing diagnoses. The objective of this study was to adapt the LSM to the Portuguese language and to estimate its psychometric properties. The original scale was translated into Portuguese, back-translated into English, and the English versions were compared in order to adjust the Portuguese one (*Escala de Acurácia de Diagnóstico de Enfermagem de Lunney – EADE*). Four nurses were trained on the EADE and applied it on 159 diagnoses made for 26 patients of three primary studies, based on the records of patients' interviews and physical examinations. Cohen's *Kappa* estimates produced unacceptable inter-observer agreement rates, showing that the adapted tool has not acceptable reliability. Because of this result, validity tests were not conducted.

KEY WORDS

Nursing diagnosis.
Nursing assessment.
Psychometrics.

RESUMEN

O *Lunney Scoring Method for Rating Accuracy of Nursing Diagnoses* (LSM) es una escala de diferencial semántico desarrollada por Lunney para estimar la perfección de los diagnósticos de enfermería. El objetivo de este estudio fue adaptar el LSM a la lengua portuguesa y evaluar sus propiedades psicométricas. La escala original fue traducida al portugués, revertida al inglés y las dos versiones en inglés fueron comparadas para ajustar la versión en portugués que pasó a ser denominada Escala de Perfeccionamiento de Diagnóstico de Enfermería de Lunney – EPDE. Cuatro enfermeras fueron orientadas sobre la EPDE y la aplicaron a 159 diagnósticos formulados para 26 pacientes de 3 estudios primarios con base en los registros de entrevista y examen físico de cada paciente. Los índices *Kappa* de Cohen mostraron ausencia de concordancia entre las evaluadoras, lo que indica que el instrumento adaptado no tiene confiabilidad satisfactoria. En virtud de ese resultado, no se realizó la estimación de validez.

DESCRIPTORES

Diagnóstico de enfermería.
Evaluación en enfermería.
Psicometria.

* Extraído do Projeto de Pesquisa financiado pela FAPESP (02-02345-0).

- 1 Enfermeira. Professora Associada da Escola de Enfermagem da Universidade de São Paulo (EEUSP). mtmlf@usp.br
- 2 Enfermeira. Doutoranda do Programa de Pós-Graduação em Enfermagem na Saúde do Adulto da EEUSP. pccf@terra.com.br
- 3 Enfermeira. Doutoranda do Programa de Pós-Graduação em Enfermagem da EEUSP. cristianegbraga@uol.com.br
- 4 Enfermeira. Mestre em Enfermagem pela EEUSP. Professora Assistente da Universidade Estadual de Londrina. volpato@uel.br
- 5 Enfermeira. Mestre em Enfermagem pela EEAN. Professora Assistente da Universidade Federal Fluminense. sulazjr@uol.com.br



INTRODUÇÃO

Dispor de instrumento que permita julgar a acurácia dos diagnósticos formulados pelas enfermeiras trará benefícios para diversos ramos de estudos sobre o diagnóstico de enfermagem e também para a sua aplicação na assistência.

Da acurácia do diagnóstico depende a adequação das intervenções escolhidas e, portanto, a qualidade da assistência de enfermagem. A confiabilidade dos resultados de pesquisas que se propõem a descrever perfis de diagnósticos de populações específicas depende da acurácia dos diagnósticos identificados, o que serve de exemplo da importância desse assunto para a pesquisa. O potencial que esse tipo de pesquisa tem de se caracterizar como pesquisa epidemiológica na clínica de enfermagem depende do desenvolvimento de métodos que assegurem a confiabilidade dos diagnósticos identificados. Desenvolver nos alunos de graduação a noção de acurácia diagnóstica permitirá a eles desenvolver um senso de maior flexibilidade no julgamento clínico.

Estudos sobre os elementos que garantem a acurácia na identificação de determinado diagnóstico trarão informações valiosas para o refinamento do próprio diagnóstico. Esses são exemplos de estudos que poderão ser facilitados se dispusermos de um instrumento capaz de mensurar a acurácia diagnóstica.

Esta pesquisa teve como objetivo analisar as propriedades psicométricas do instrumento adaptado do Lunney Scoring Method for Rating Accuracy of Nursing Diagnoses, desenvolvido por Lunney⁽¹⁾ que passamos a designar Lunney Scoring Method (LSM). Trata-se de uma escala com sete níveis criada para pontuar a acurácia de diagnósticos. É norteada pelo princípio de suficiência dos dados para a afirmação diagnóstica e de pertinência do diagnóstico aos dados e ao contexto em que eles ocorrem⁽²⁾.

A definição conceitual de acurácia do diagnóstico de enfermagem foi baseada no pressuposto de que há ampla variedade de diagnósticos possíveis nas situações clínicas, além dos que são altamente acurados⁽²⁾. Isto é, além de um diagnóstico altamente acurado, outros diagnósticos são aceitáveis com base nos dados existentes. Assim a acurácia de um diagnóstico não é uma qualidade dicotômica - a acurácia é um *continuum*⁽²⁾. A acurácia envolve o esforço de diferenciar um diagnóstico dos outros diagnósticos possíveis para chegar ao diagnóstico que melhor expressa os dados disponíveis em dada situação⁽³⁾.

A autora⁽²⁾ descreve cinco características da acurácia dos diagnósticos de enfermagem. A primeira característica é

que a acurácia é julgada frente ao conjunto das informações presentes em dada situação e não só frente aos dados selecionados como características definidoras de um determinado diagnóstico. Por isso, quanto maior o número e variedade de elementos representados na situação do cliente, mais difícil é alcançar acurácia.

A segunda característica apontada é que algumas combinações de dados são inerentemente mais difíceis de interpretar, independentemente da experiência do diagnosticador ou do envolvimento do cliente na validação dos significados dos dados. Alguns dados são insuficientemente estudados para nortear o raciocínio diagnóstico. Por exemplo: tocar a campainha para chamar a atenção da enfermeira é um dado que não foi estudado. Por isso os enfermeiros não conhecem a probabilidade de essa pista indicar um comportamento exigente ou de indicar algo como sentimento de impotência⁽²⁾.

A terceira característica da acurácia é que ela inclui o uso de dados que sustentam o diagnóstico, bem como o uso de dados que com ele conflitam⁽²⁾. Como existe uma tendência em ignorar os dados que conflitam com uma hipótese diagnóstica, a acurácia depende da consideração dos dados que conflitam com a hipótese diagnóstica⁽⁴⁾.

A quarta característica é que altos graus de acurácia dos diagnósticos de enfermagem são resultado da avaliação da relevância dos dados disponíveis. Avaliar a relevância significa pesar a sua importância para a situação como um todo e para diagnósticos específicos. Os julgamentos sobre a relevância dos dados são feitos pelos diagnosticadores na interação com o cliente, mas também são feitos por quem quer que esteja julgando a acurácia do diagnóstico⁽²⁾.

A quinta característica é que os enfermeiros não são igualmente rigorosos na sua atenção quanto a acurácia dos diagnósticos que fazem. O contexto em que o diagnóstico está sendo realizado pode afetar a sua acurácia. São exemplos de contextos com diferentes exigências de acurácia: se o diagnóstico é realizado em situação real ou simulada, se o diagnóstico é realizado em situação em que o enfermeiro é totalmente responsável pelo cliente ou em situação em que compartilha com outros profissionais essa responsabilidade; se o enfermeiro está sendo avaliado ou não por pares ou se o aluno está ou não sendo avaliado por seus professores. Em síntese, essa característica leva em conta que a atenção que a enfermeira dá à acurácia depende do contexto em que a tarefa diagnóstica está ocorrendo⁽²⁾.

Com base nessas propriedades a autora⁽²⁾ define acurácia de um diagnóstico como o julgamento de um avaliador so-

A confiabilidade dos resultados de pesquisas que se propõem a descrever perfis de diagnósticos de populações específicas depende da acurácia dos diagnósticos identificados, o que serve de exemplo da importância desse assunto para a pesquisa

bre o grau com que uma afirmação diagnóstica corresponde aos dados da situação do paciente/cliente. O avaliador pode ser o próprio diagnosticador ou outro enfermeiro.

O LSM, fundamentado no conceito descrito de acurácia é uma escala ordinal de graus de acurácia de diagnóstico. Essa escala tem sete valores com seus respectivos critérios conforme apresentado na Tabela 1.

Tabela 1 - Escala ordinal para graus de acurácia⁽²⁾

Value	Criteria
+ 5	Diagnosis is consistent with all of the cues, supported by highly relevant cues, and precise.
+4	Diagnosis is consistent with most or all of the cues and supported by relevant cues but fails to reflect one or a few highly relevant cues.
+3	Diagnosis is consistent with many of the cues but fails to reflect the specificity of available cues.
+2	Diagnosis is indicated by some of the cues but there are insufficient cues relevant to the diagnosis and/or the diagnosis is a lower priority than other diagnoses.
+1	Diagnosis is only suggested by one or a few cues.
0	Diagnosis is not indicated by any of the cues.
	No diagnosis is stated when there are sufficient cues to state a diagnosis.
-1	Diagnosis is indicated by more than one cue but should be rejected based on the presence of at least two disconfirming cues

Essa proposta pode ser desenvolvida para preencher uma lacuna importante na pesquisa sobre diagnóstico de enfermagem⁽²⁾. Dispor de instrumento que balize o julgamento da acurácia diagnóstica poderá garantir maior confiabilidade nos resultados de estudos em que o estabelecimento de diagnósticos é uma das variáveis envolvidas. O conceito de acurácia geralmente não é explicitado nesses estudos e naqueles que já trataram da acurácia dos diagnósticos de enfermagem⁽⁵⁻⁶⁾ adotou-se concepção dicotômica de acurácia.

Do que se tem descrito sobre o desenvolvimento do LSM o método ainda não foi aplicado a diagnósticos formulados com base em registros de histórico e exame físico de enfermagem.

Numa versão preliminar do método, 3 descrições de situações clínicas hipotéticas, com 20 possíveis diagnósticos cada uma, foram apresentadas a quatro peritos em diagnóstico de enfermagem. Trabalhando independentemente e sem nenhum treinamento esses peritos obtiveram 59,6% de concordância nos escores para os 60 diagnósticos (20 de cada uma das 3 descrições)⁽¹⁾. As confiabilidades entre dois avaliadores estimadas pela correlação de Pearson foram iguais a 0,96, 0,97 e 0,97 nos testes sobre os escores atribuídos por dois avaliadores às respostas para três estudos de caso⁽¹⁾.

O método foi aplicado em estudo em que 62 enfermeiros avaliaram e diagnosticaram problemas psicossociais de 153 pacientes internados⁽⁸⁾. Tratava-se, portanto, de casos clínicos reais. Em cada situação, o mesmo paciente era avaliado por 2 peritos que, após a avaliação, aplicavam o LSM aos diagnósticos formulados pelos enfermeiros. A confiabilidade entre avaliadores das pontuações dos peritos, estimada pela correlação de Pearson foi 0,95⁽⁷⁾. No entanto, a publicação não deixa claro se essa estimativa foi obtida com os dados das 153 avaliações ou com base em descrições hipotéticas aplicadas antes da coleta dos dados como procedimento para indicação dos peritos.

Em um estudo preliminar aplicamos o método aos diagnósticos formulados por 89 enfermeiras e estudantes

de enfermagem a um paciente hipotético. Foram enunciados 388 diagnósticos no total e dois avaliadores aplicaram o LSM a esses enunciados⁽⁸⁾. Todos os níveis da escala apareceram pelo menos uma vez. O escore -1 foi atribuído três vezes por apenas um dos avaliadores. Excluindo-o, o coeficiente de Kappa foi igual a 0,51 ($p < 0,001$). Esses resultados nos estimularam a realizar os procedimentos formais de tradução e adaptação desse instrumento considerando a possibilidade de aplicá-lo a diagnósticos formulados com base em registros de histórico e exame físico de enfermagem.

Os objetivos deste estudo foram: adaptar o LSM para a língua portuguesa e verificar a sua confiabilidade. Havia o objetivo de testar a validade do LSM. No entanto, os resultados obtidos com as estimativas de confiabilidade contra-indicaram a análise de validade, o que será discutido oportunamente.

MÉTODO

Versão para a língua portuguesa

A autora do LSM autorizou a adaptação para a língua portuguesa. Os procedimentos para obter a escala em língua Portuguesa foram: versão, retro-tradução e comparação do original e retro-tradução. As duas primeiras etapas foram realizadas por tradutores independentes e a comparação das duas versões foi realizada pelos autores do estudo, que fizeram os ajustes necessários na versão em Português.

Terminando essa etapa o instrumento estava disponível em língua Portuguesa para ser testado quanto a confiabilidade e passou a ser denominado Escala de Acurácia de Diagnóstico de Enfermagem de Lunney (EADE).

Análise de confiabilidade

A confiabilidade da EADE foi estudada com base em análise secundária de dados de pesquisas realizadas pelos autores deste estudo.

A EADE foi aplicada por 4 enfermeiras, alunas de cursos de pós-graduação em enfermagem, com experiência no uso de diagnósticos de enfermagem na prática clínica e na pesquisa.

Para a análise da confiabilidade da EADE foram utilizados dados já obtidos em 3 estudos de descrição de perfil de diagnósticos, a saber: um estudo com 60 pacientes internados em unidade médico-cirúrgica⁽⁹⁾ (Médico-Cirúrgica), um estudo com 40 pacientes em tratamento ambulatorial por diabetes⁽¹⁰⁾ (Diabetes), e um estudo com 30 pacientes internados para cirurgias de reconstrução de lábio e palato⁽¹¹⁾ (Fissuras). Em cada um desses estudos os pacientes foram entrevistados e examinados e os dados obtidos foram registrados em formulários próprios. Nos estudos primários, os formulários com os registros de entrevista e exame físico foram distribuídos entre 4 enfermeiras que estabeleceram, independentemente, os diagnósticos de cada paciente. Uma das enfermeiras era a que havia realizado a entrevista e o exame físico dos pacientes. Todos os diagnósticos formulados pelos 4 enfermeiros foram analisados e foram aceitos para compor os resultados principais dos estudos primários apenas os diagnósticos que, por consenso, foram considerados suficientemente embasados nos dados disponíveis. Disponha-se assim dos seguintes dados desses estudos primários: formulário com registro dos dados obtidos em entrevista e exame físico de cada paciente, todos os diagnósticos indicados independentemente pelos quatro enfermeiros e os diagnósticos aceitos por consenso entre eles.

No presente estudo a EADE foi aplicada a uma amostra aleatória dos dados desses estudos, considerando-se o paciente como unidade amostral. De cada estudo foi sorteado 20% dos formulários. Os dados de exame físico e entrevista de cada um, bem como os diagnósticos formulados independentemente pelos 4 enfermeiros, foram apresentados a 4 enfermeiras (diferentes das 4 dos estudos primários), com experiência clínica e em pesquisa sobre diagnósticos de enfermagem, para que aplicassem a EADE. Dos 3 estudos primários foram sorteados: 12 casos do estudo de pacientes em unidade médico-cirúrgica, 8 do estudo de pacientes diabéticos e 6 do estudo de pacientes com fissuras lábio-palatais. Os totais de hipóteses diagnósticas para os pacientes sorteados de cada estudo foram 68, 51 e 40, respectivamente. Assim, o total de diagnósticos submetidos a EADE pelas 4 avaliadoras foi 159.

A EADE, como já descrito, tem como princípio a confiabilidade e suficiência dos dados para afirmar os diagnósticos estabelecidos. Portanto, os enfermeiros avaliadores neste estudo aplicaram o método comparando os diagnósticos estabelecidos com os dados disponíveis nos formulários de entrevista e exame físico de cada paciente.

O teste da hipótese de que os diagnósticos que foram aceitos nos consensos dos estudos primários teria maior proporção de escores mais altos na EADE que os diagnósticos que não foram incluídos no consenso indicaria a validade de construto do instrumento. Como já referido, essa etapa do estudo não foi realizada, em virtude de os resultados de confiabilidade não terem sido satisfatórios, como se verá mais adiante.

A confiabilidade da EADE foi testada pelo coeficiente de concordância entre avaliadores Kappa de Cohen⁽¹²⁾ analisando-se duas a duas as respostas das quatro avaliadoras e também pelo Kappa geral⁽¹²⁾.

A confiabilidade entre avaliadores é especialmente importante quando se trata de medidas observacionais ou de instrumentos subjetivos tais como respostas livres que requerem categorização, e estudos de caso⁽¹³⁾. Além desse método, pretendia-se aplicar a análise de cluster⁽¹⁴⁾, ao conjunto das respostas das quatro avaliadoras para verificar se havia consistência na conformação dos conglomerados de diagnósticos segundo os escores. Pela mesma razão porque não foi realizada a análise de validade, a análise de cluster também não foi aplicada.

Antes da aplicação da EADE, os 4 enfermeiros avaliadores participaram de uma sessão de treinamento de 4 horas em que se discutiu o conceito de acurácia subjacente ao instrumento e o próprio instrumento em língua portuguesa. Nessa oportunidade a EADE foi aplicada a diagnósticos que haviam sido formulados para 2 pacientes dos estudos primários com a finalidade de treinamento. Esses 2 casos não foram incluídos no sorteio para o estudo propriamente dito.

RESULTADOS

A Tabela 2 mostra a versão brasileira do Lunney Scoring Method for Rating Accuracy of Nursing Diagnosis, denominado, em português, de Escala de Acurácia de Diagnóstico de Enfermagem de Lunney - Versão Brasileira (EADE).

Tabela 2 - Escala de Acurácia de Diagnóstico de Enfermagem de Lunney - versão brasileira

Value	Criteria
+ 5	Diagnosis is consistent with all of the cues, supported by highly relevant cues, and precise.
+4	Diagnosis is consistent with most or all of the cues and supported by relevant cues but fails to reflect one or a few highly relevant cues.
+3	Diagnosis is consistent with many of the cues but fails to reflect the specificity of available cues.
+2	Diagnosis is indicated by some of the cues but there are insufficient cues relevant to the diagnosis and/or the diagnosis is a lower priority than other diagnoses.
+1	Diagnosis is only suggested by one or a few cues.
0	Diagnosis is not indicated by any of the cues.
	No diagnosis is stated when there are sufficient cues to state a diagnosis.
-1	Diagnosis is indicated by more than one cue but should be rejected based on the presence of at least two disconfirming cues

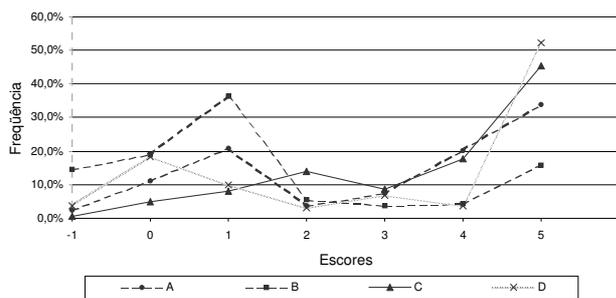


Figura 1 - Distribuição das respostas dos avaliadores segundo escores da EADE

A Figura 1 mostra as proporções de aplicação de cada um dos 7 escores da EADE pelas 4 avaliadoras aos 159 diagnósticos dos 26 pacientes sorteados dos 3 estudos primários.

A Figura 1 mostra que todas as avaliadoras aplicaram os 7 escores da EADE pelo menos uma vez. Ao observá-la pelo eixo dos escores verifica-se que o escore +3 é o que teve proporções mais semelhantes entre os quatro avaliadores e também que essas proporções são baixas, menores que 10%. O mesmo ocorre com o escore +2 e -1, excetuando-se os resultados das avaliadoras C e B, respectivamente. Nota-se também que os escores +1 e +5 foram os que tiveram proporções mais diferenciadas entre as avaliadoras. Ao observar a Figura 2 pelo eixo de frequência, verifica-se que os escores +1 e +5 foram os que tiveram proporções mais elevadas de aplicação.

A Tabela 3 mostra a estatística descritiva dos escores atribuídos segundo as avaliadoras.

Tabela 3 - Mediana, média e desvios-padrão dos escores na EADE segundo avaliadora - São Paulo - 2003

Avaliadora	Escore na EADE					
	Mínimo	Mediana	Máximo	Média	Desvio padrão	Total válido
A	-1	4	5	2,99	1,97	159
B	-1	1	5	1,42	1,94	158
C	-1	4	5	3,61	1,63	158
D	-1	5	5	3,15	2,21	156

Na Tabela 3 observa-se que a média mínima foi de 1,42 e a máxima de 3,61.

Considerando-se que estas avaliações foram feitas na mesma unidade de observação (cliente - diagnóstico), levou-se em consideração a correlação entre estes escores. Realizou-se, então, o teste não paramétrico de Friedman, cujo resultado mostrou que os escores atribuídos pelas 4 avaliadoras não se correlacionaram ($p < 0,001$).

Para verificar se essas diferenças foram efeito de características dos estudos primários (tipo de pacientes ou tipo de dados, por exemplo), analisaram-se os escores dos 4 avaliadores segundo os estudos primários. A Tabela 4 mostra a estatística descritiva dos escores na EADE segundo os estudos primários de origem dos pacientes.

Tabela 4 - Mediana, média e desvios-padrão dos escores na EADE segundo estudo primário de origem - São Paulo - 2003

Estudo primário	Avaliadora	Escore					n
		Mínimo	Mediana	Máximo	Média	Desvio padrão	
Fissuras	A	0	3	5	2,55	1,88	40
	B	-1	0,5	5	0,60	1,28	40
	C	0	4	5	3,20	1,60	40
	D	-1	5	5	3,39	2,33	38
Médico-Cirúrgica	A	-1	5	5	3,60	1,96	68
	B	-1	1	5	1,54	1,90	67
	C	-1	5	5	3,76	1,68	67
	D	-1	5	5	3,31	2,19	67
Diabetes	A	-1	3	5	2,51	1,85	51
	B	-1	1	5	1,90	2,23	51
	C	0	5	5	3,73	1,58	51
	D	-1	3	5	2,76	2,15	51

Os escores médios por conjunto de pacientes em situações semelhantes, isto é, que compuseram o mesmo estudo primário, mantiveram ampla variação segundo as avaliadoras (fissuras: entre 0,60 e 3,39; médico-cirúrgica: entre 1,54 e 3,76; diabetes: entre 1,90 e 3,73). Observa-se também, na

Tabela 2, que as variações dos escores médios, fixando-se a avaliadora, foram menores que fixando-se o estudo primário (avaliador A: entre 2,51 e 3,60; B: entre 0,60 e 1,90; C: entre 3,20 e 3,73; D: entre 2,76 e 3,39).

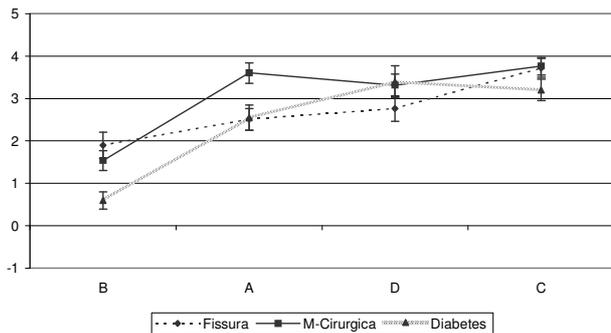


Figura 2 - Perfil médio dos escores na EADE segundo avaliadora e estudo primário (média + EMP)

A Figura 2 mostra os escores médios atribuídos pelas 4 avaliadoras em cada conjunto de pacientes segundo os estudos primários.

Nota-se que, no geral, a ordem crescente de magnitude dos escores de acurácia, segundo as avaliadoras é B, A, D, C (Figura 2).

Para avaliar essas diferenças e verificar se os escores de alguma avaliadora estavam influenciando as diferenças, calculou-se o *odds-ratio* e *odds-ratio* inverso combinando-se 2 a 2 os escores das quatro avaliadoras segundo os estudos primários. A Tabela 5 mostra os resultados.

Se houvesse consistência nas diferenças entre pares de avaliadores nos três estudos (*odds-ratio* invertido semelhante entre os 3 estudos para o mesmo par), poderíamos utilizar esse resultado como critério para analisar se os escores de alguma avaliadora específica exerceram papel importante na determinação das diferenças entre as 4. Os resultados na Tabela 5 mostram que as diferenças entre os pares de avaliadoras não foram consistentes entre os estudos primários, o que não permitiu indicar se uma ou outra estava influenciando as diferenças entre os escores que atribuíram aos 159 diagnósticos submetidos à EADE.

O índice Kappa é uma estimativa de concordância freqüentemente aplicada a pares de avaliadores, quando o nível de mensuração é nominal. Varia entre 0 e 1 e quanto mais próximo de 1 maior a concordância. Para a maioria das finalidades, os valores maiores ou iguais a 0,75 representam concordância excelente, valores entre 0,40 e 0,75 representam concordância razoável ou boa, e valores menores ou iguais a 0,40 representam baixa concordância⁽¹²⁾.

Calculou-se o Kappa entre pares de avaliadoras e o Kappa geral escores atribuídos pelas 4 avaliadoras para os 159 diagnósticos. A Tabela 6 mostra os resultados obtidos.

Tabela 5 - *Odds-ratio* e *odds-ratio* inverso dos escores de pares das avaliadoras, segundo estudos primários - São Paulo - 2003

Pares de avaliadoras e estudos primários	Odds	Odds Inverso
1. B x C		
Fissuras	0,682	1,47
Médico-Cirúrgica	0,614	1,63
Diabetes	0,500	2,00
2. A x C		
Fissuras	0,788	1,27
Médico-Cirúrgica	0,972	1,03
Diabetes	0,875	1,14
3. D x C		
Fissuras	0,832	1,20
Médico-Cirúrgica	0,924	1,08
Diabetes	1,030	0,97
4. B x D		
Fissuras	0,819	1,22
Médico-Cirúrgica	0,665	1,50
Diabetes	0,485	2,06
5. A x D		
Fissuras	0,947	1,06
Médico-Cirúrgica	1,052	0,95
Diabetes	0,849	1,18
6. B x A		
Fissuras	0,865	1,16
Médico-Cirúrgica	0,632	1,58
Diabetes	0,571	1,75

Tabela 6 - Resultados do Kappa de Cohen sobre os escores da EADE atribuídos a 159 diagnósticos por 4 avaliadoras - São Paulo - 2003

Variáveis	Resultados		
	Kappa	p	n
Pares de avaliadoras			
A x B	0,04	0,22	158
A x C	0,134	<0,001	158
A x D	0,084	0,022	158
B x C	0,039	0,152	157
B x D	0,073	0,021	155
C x D	0,133	<0,001	155
A, B, C, D (geral)	0,007	0,00001	155

Em todos os casos a concordância foi muito baixa. O Kappa geral foi o mais próximo de zero; os mais elevados foram os obtidos pela comparação entre os avaliadores A e C, e entre C e D. Mesmo assim ficaram ainda muito distantes de valores considerados razoáveis (entre 0,40 e 0,75)⁽¹²⁾.

DISCUSSÃO

Para ser útil, um instrumento de medida deve ser válido e confiável. A validade refere-se ao poder do instrumento para medir aquilo que se quer medir. No caso da EADE, a validade refere-se ao poder que ela tem para realmente medir a acurácia dos diagnósticos de enfermagem. A confiabilidade trata do poder que o instrumento tem para apresentar resultados muito semelhantes quando aplicado diversas vezes à mesma situação. No caso do EADE, a confiabilidade poderia ser indicada por uma variação mínima de escores quando aplicado ao mesmo diagnóstico na mesma situação. Neste estudo, a EADE teria boa confiabilidade se houvesse concordância entre as avaliadoras nos escores aplicados aos mesmos diagnósticos. Como todas avaliaram o mesmo conjunto de dados (registros de entrevista e exame físico) frente a cada diagnóstico pontuado, era esperada maior proximidade entre os escores das diferentes avaliadoras para um mesmo diagnóstico.

As propriedades de confiabilidade e de validade são estimadas por métodos independentes e podem produzir resultados em direções opostas como, por exemplo, validade alta e confiabilidade baixa. Mesmo que um instrumento tenha índices elevados de validade, a confiabilidade baixa contra-indica o seu uso e põe em dúvida a validade estimada⁽¹⁵⁾.

As análises empreendidas neste estudo mostraram que a EADE não tem confiabilidade adequada. Os índices de Kappa foram próximos ou muito próximos de zero e também o teste de Friedman mostrou que não há correlação entre os escores aplicados pelas avaliadoras. Em estudo em que o Lunney Scoring Method for Rating Accuracy of Nursing Diagnoses foi aplicado por peritos em diagnósticos de enfermagem os índices de confiabilidade foram de 0,95, estimados pelo teste de correlação de Spearman, indicando alta confiabilidade⁽¹⁶⁾. Os testes de correlação não são os mais adequados para se estimar concordância entre avaliadores⁽¹⁷⁻¹⁸⁾. No entanto, o estudo citado é a única referência de que dispomos sobre estimativas de confiabilidade do instrumento em sua versão original. Realizamos também o teste de correlação (neste caso o de Friedman por tratar-se de mais que 2 avaliadores) para que pudéssemos ter base de comparação com o resultado do estudo referido⁽¹⁶⁾. Também nessa análise a correlação foi muito fraca. Apesar de se ter aplicado teste de correlação, o teste de Kappa, na análise de concordância entre avaliadores, é mais adequado porque ele ajusta o resultado para as concordâncias que ocorrem ao acaso⁽¹²⁾.

A confiabilidade da EADE ficou muito abaixo dos limites aceitáveis e isso pode lançar dúvidas acerca da idéia de diagnóstico de enfermagem e também acerca do conceito de raciocínio diagnóstico na enfermagem. Nesse sentido é importante observar que em outras áreas clínicas os estudos

têm identificado concordâncias entre diagnosticadores geralmente baixas, como os citados por GARB⁽¹⁷⁾.

Uma característica boa do instrumento é que todos os escores (de -1 a +7) foram aplicados pelas 4 avaliadoras. No entanto, considerando que o que foi julgado foram hipóteses diagnósticas, dentre as quais estariam diagnósticos com diferentes graus de acurácia, seria esperado que os escores intermediários tivessem maiores proporções de indicações e isso não ocorreu. Seria esperado que os escores +2 e +3 apresentassem proporções mais elevadas de aplicação, ou então que houvesse maior uniformidade nas linhas da Figura 1.

Esses resultados sugerem que há algum problema com os escores intermediários da escala. Ao analisá-los (Tabela 2) frente aos resultados da Figura 1, observamos que, especialmente os escores +2, +3, e +4, podem não estar suficientemente claros, o que teria contribuído para a baixa confiabilidade observada pelos testes de concordância.

O critério do escore +2 (O diagnóstico é indicado por algumas das pistas, mas as pistas relevantes são insuficientes para o diagnóstico e/ou o diagnóstico tem prioridade menor que outros diagnósticos) contém uma afirmação, uma restrição, além de uma adição ou alternativa. Esse conteúdo contribui para dificuldades para apreender o enunciado do critério e, conseqüentemente, em inconsistências na sua aplicação.

O critério do escore +3 (O diagnóstico é consistente com muitas das pistas, mas falha em refletir a especificidade de pistas disponíveis) contém uma restrição à afirmação principal. Além disso, o critério do escore +4 (O diagnóstico é consistente com a maioria ou todas as pistas e suportado por pistas relevantes, mas falha em refletir uma ou poucas pistas relevantes) compartilha boa parte do seu conteúdo com o critério do +3. Os dois (+3 e +4) consideram a falha em refletir pistas disponíveis, com a diferença de que o +4 se refere a especificidade das pistas e o +3 não.

Essas observações fundamentam a necessidade de revisão da EADE, especialmente nos seus escores intermediários, que parecem não ter a simplicidade desejável, visto que um dos atributos necessários para indicadores ou critérios é que eles sejam simples. Isto é, *quanto mais simples de buscar, calcular e analisar, maiores são as chances e oportunidades de utilização*⁽¹⁹⁾.

Além de possíveis problemas com o instrumento em si, é preciso considerar aspectos metodológicos que podem ter contribuído para os resultados obtidos. Os estudos com o instrumento original, em língua inglesa, tiveram delineamentos diferentes deste estudo. Destaca-se o fato de que a confiabilidade, no estudo com o instrumento original, foi estimada a partir da aplicação da escala por duplas de enfermeiros que haviam, os dois, interagido com o paciente

para a avaliação. Ao apreciar os presentes resultados, a autora do instrumento pondera que talvez ele seja aplicável apenas a situações em que os avaliadores têm oportunidade de entrevistar e examinar o paciente para o qual o diagnóstico está sendo pontuado. É uma possibilidade a ser considerada que, se confirmada, limitaria muito a utilidade do instrumento.

Outro aspecto a ser considerado é a possibilidade de as avaliadoras não terem sido adequadamente preparadas para a aplicação do método, ou ainda, que necessitem de maior experiência com os processos de raciocínio diagnóstico.

Como já foi antecipado, optamos por realizar os testes de confiabilidade em primeiro lugar e, como ela não foi demonstrada, decidimos não realizar estimativas de validade.

CONCLUSÃO

A acurácia em diagnósticos de enfermagem é um importante assunto para ser pesquisado. A Escala de Acurácia de Diagnóstico de Enfermagem de Lunney - versão brasileira

(EADE) é o único método de avaliação para conduzir este desafio. Os estudos disponíveis sobre a aplicação da versão original em língua inglesa mostram que o instrumento tem estimativas adequadas de confiabilidade em situações reais, na qual o diagnosticador tem oportunidade de assistir o paciente, identificar os fenômenos, diagnosticando-os e validando-os.

Esse estudo mostrou que a versão brasileira do LSM não tem estimativas aceitáveis de confiabilidade, o que contra-indica a sua utilização quando esse rigor é necessário, especialmente em situações em que o avaliador não avalia, ele mesmo, o paciente.

Considerando a importância de estudos epidemiológicos sobre os diagnósticos de enfermagem, e que a validade de seus resultados depende da acurácia dos diagnósticos, os esforços para desenvolver formas de avaliá-la devem ser encorajados.

Os aspectos considerados vulneráveis na EADE servirão de base para o seu aprimoramento ou para o desenvolvimento de outro instrumento que seja válido e confiável.

REFERÊNCIAS

1. Lunney M. Measurement of accuracy of nursing diagnoses. In: Paquette M, editor. *Classification of Nursing Diagnosis: Proceedings of the Tenth Conference*; 1994 Apr. 25-29; San Diego, CA. Philadelphia: Lippincott; 1994. p. 281-4.
2. Lunney M. Accuracy of nursing diagnosis: concept development. *Nurs Diag*. 1990;1(1):12-7.
3. Lunney M. Self-monitoring of accuracy using an integrated model of the diagnostic process. *J Adv Med Surg Nurs*. 1989;1(3): 43-52.
4. Mynatt CR, Doherty ME, Twenney RD. Consequences of confirmation and disconfirmation in a simulated research environment. *Q J Exp Psych*. 1978;(30):395-406.
5. Aspinall MJ. Nursing diagnosis: the weak link. *Nurs Outlook*. 1976;24(7):433-7.
6. Tanner CA, Padrick KP, Westfall UE, Putzier DJ. Diagnostic reasoning strategies of nurses and nursing students. *Nurs Res*. 1987;36(6):358-63.
7. Lunney M, Karlik BA, Kiss M, Murphy P. Accuracy of nurses' diagnoses of psychosocial responses. *Nurs Diagn*. 1997;8(4): 157-66.
8. Cruz DAL, Perez RCFG. Accuracy of nursing diagnosis: interrater agreement. *Int J Nurs Terminol Classif*. 2003;14 (4 Suppl):S47.
9. Volpato MP. Diagnóstico de enfermagem em clientes internadas em unidade médico-cirúrgica [dissertação]. São Paulo: Escola de Enfermagem, Universidade de São Paulo; 2001.
10. Azevedo SL. Diagnósticos de enfermagem: orientadores do cuidado ao cliente diabético [dissertação]. Rio de Janeiro: Escola de Enfermagem Anna Nery, Universidade Federal do Rio de Janeiro; 2001.
11. Fontes CMB. Fissuras lábio-palatais: diagnósticos de enfermagem no pré e pós-operatório [dissertação]. São Paulo: Escola de Enfermagem, Universidade de São Paulo; 2001.
12. Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York: Wiley; 1981.
13. Waltz CF, Strickland OL, Lenz ER. Reliability and validity or norm-referenced. In: Lenz ER, editor. *Measurement in nursing research*. 2nd ed. Philadelphia: Davis; 1991. p. 161-94.
14. Hair JF, Anderson RE, Tatham RL, Black WC. *Multivariate data analysis*. 5th ed. New Jersey: Prentice Hall; 1998.
15. Nunnally JC, Bernstein IH. *Psychometric theory*. 3rd ed. New York: McGraw-Hill; 1994.
16. Lunney M. *Critical thinking and nursing diagnosis: case studies and analyses*. Philadelphia: NANDA International; 2001.
17. Garb HN. *Studying the clinician: judgment research and psychological assessment*. Washington, DC: American Psychological Association; 1998.
18. Shoukri MM. *Measures of interobserver agreement*. Boca Raton: Chapman & Hall/CRC; 2004.
19. Bittar OJNV. Indicadores de qualidade e quantidade em saúde. *Rev Adm Saúde*. 2001;3(12):21-8.

Agradecimentos: Dra. Margaret Lunney pela autorização para adaptar a escala e suporte na interpretação dos resultados; à Brauns & Brauns pelo suporte nas traduções, à Valéria Troncoso Baltar pelas orientações e realização dos testes estatísticos