Teoria de Resposta ao Item para otimização de escalas tipo likert– um exemplo de aplicação

For item response theory likert scale´s optimization- an application example

Carlos Henrique Sancineto da Silva Nunes[1], Ricardo Primi[2], Maiana Farias Oliveira Nunes[3], Monalisa Muniz[4], Tatiana Freitas da Cunha[5], Gleiber Couto[6]

RESUMO

A Teoria de Resposta ao Item (TRI) tem sido utilizada tradicionalmente para análise de testes com itens dicotômicos. Contudo, recentemente, algumas pesquisas têm indicado a utilidade dessa abordagem para a análise de testes baseados em escalas politômicas. O presente estudo teve como objetivo verificar as vantagens da utilização da TRI em relação a Teoria Clássica dos Testes no que diz respeito à quantidade necessária de categorias para a realização de medidas, aplicando a técnica dos créditos parciais. Para tanto, foram verificadas as propriedades psicométricas de cada categoria e seu número foi otimizado. Foram analisadas as respostas de 1.317 pessoas a uma escala Brasileira para avaliação de *Socialização no Modelo dos Cinco Grandes Fatores de Personalidade*, que

1. Doutor em Psicologia. Professor da Universidade São Francisco
Apoio: CNPq, CAPES e FAPESP. Os autores são pesquisadores vinculados ou colaboradores do Laboratório de Avaliação Psicológica e Educacional – LabAPE – USF. Correspondências devem ser enviadas para: Carlos Henrique Sancineto da Silva Nunes, Universidade São Francisco - Faculdade de Ciências Humanas, Rua Alexandre Rodrigues Barbosa, 45, CEP 13251-900, Itatiba – SP, Brasil. correio eletrônico: carlos.sancineto@pesquisador.cnpq.br
2. Doutor em Psicologia. Professor da Universidade São Francisco
3. Mestre em Psicologia. Doutoranda da Universidade São Francisco
4. Mestre em Psicologia. Doutoranda da Universidade São Francisco
5. Mestre em Psicologia. Pesquisadora colaboradora da Universidade São Francisco
6. Doutor em Psicologia. Professor da Faculdade de Estudos Administrativos de Belo Horizonte

originalmente utilizou uma escala *Likert* de 7 pontos. Três subescalas do instrumento e três itens do mesmo foram usados para exemplificar o processo de recodificação. A análise da adequação item/escala, da estrutura e da precisão foram comparadas em ambas as situações. Os resultados indicaram que a utilização de categorias otimizadas, em um menor número que na escala original, permitiu a mensuração do construto sem prejudicar os parâmetros psicométricos do instrumento, sendo que em alguns fatores as medidas de consistência interna e *misfit* foram superiores às originais.

Palavras-chave: Teoria de resposta ao item, escalas tipos *Likert*, otimização de escalas, psicometria, personalidade

ABSTRACT

Item Response Theory (IRT) has been traditionally used to analyze tests with dichotomous items. Nevertheless, in the past years, some research has shown the utility of adopting IRT for polythomous scales. This study aimed to verify the advantages of using IRT when compared to Classic Test Theory (CTT), regarding the necessary amount of answer categories to measure the construct, by using partial credits method of analysis. For that reason, the scale was recoded into a smaller one, and the number of categories required depended on its specific characteristics. The sample was composed by 1,317 people, who answered a Brazilian scale for Agreeableness assessment in the Five Factor Model of Personality, which used originally a 7-point Likert scale. Its three subscales and three items were used to exemplify the procedure of recoding. The analysis of item/scale fit, structure and reliability were compared in both situations. The results indicate that using optimized categories, within a smaller number, has allowed measuring the construct and has maintained the psychometric parameters of the scale. Also, in some subscales, internal consistency and misfit measures were better than the original ones.

Key words: Item Response Theory, Likert scales, optimizing scales, psychometrics, personality

INTRODUÇÃO

Muitos instrumentos construídos na psicologia, educação e em outras áreas utilizam itens com escalas tipo *Likert*. Nesses itens, geralmente se têm uma afirmação auto-descritiva e, em seguida, uma escala de pontos com descrições verbais, tais como, discordo totalmente (1), discordo (2), neutro (3), concordo (4) e concordo totalmente (5). Pretende-se, dessa forma, mensurar a intensidade do traço representado no item. A abordagem de medida que sustenta esses instrumentos decorre da teoria do escore verdadeiro, proveniente da teoria clássica dos testes, particularmente, das aplicações desse modelo no princípio da consistência interna.

Resumidamente, quando se tem duas afirmações que se supõe estarem medindo o mesmo construto e atrelado a elas se tem uma escala *Likert* de, por exemplo, cinco pontos, pode-se conceber a situação como dois mini-testes (de métrica 1 a 5) aplicados repetidamente em um mesmo sujeito. De acordo com a teoria clássica dos testes , quando se re-testa um sujeito com testes paralelos (medindo o mesmo construto, com a mesma intensidade) o escore verdadeiro, isto é, o valor que o sujeito possui no construto medido, é o mesmo nas duas situações . Entretanto, o resultado observado, isto é, a resposta observada (1 a 5) pode variar de um item para outro em razão do erro de medida, decorrente da imperfeição habitual que os instrumentos possuem ao tentar medir a posição do sujeito no construto (ou o escore verdadeiro do sujeito).

De acordo com essa lógica, quando se constrói um conjunto de afirmações pretendendo medir o mesmo construto utilizando uma escala *Likert* e se computa a soma das pontuações no item, esse escore é uma estimativa do escore verdadeiro do sujeito. A lógica desse procedimento está na replicabilidade da medida, isto é, na correlação entre os escores nos itens que é determinada pelo que existe de sistemático nessas pontuações, ou seja, o escore verdadeiro de cada sujeito que se repete a cada item. Baseado nisso, Cronbach criou o coeficiente de consistência interna (Alfa de Cronbach), que indica o quanto as pontuações de um teste (conjunto de itens somados para indicar um construto) é consistente como estimador do escore verdadeiro. Conversamente, esse mesmo indicador permite se estimar o montante de erro esperado. Esse coeficiente é fundamentalmente diretamente proporcional às correlações entre os itens. Assim, quanto maior forem as correlações inter-itens, pretendendo medir o mesmo construto, mais os escores observados refletem o escore verdadeiro (menor a influência do erro) e maior será a confiabilidade da medida.

Para a realização desse procedimento, verifica-se as correlações entre

os itens com a pontuação total da escala. Quanto maior for essa pontuação, mais o item está correlacionado com os itens restantes da escala e mais ele contribui para a consistência da medida. Ao mesmo tempo, observa-se a variabilidade dos itens já que escores mais variáveis são, a princípio, mais capazes de diferenciar os sujeitos e maior a possibilidade de se obterem correlações altas com os outros itens .

Segundo essa especificação, a construção de testes usando escalas *Likert* recomenda, dentre outras coisas, para garantir que todos sejam bons indicadores do mesmo construto, que tenham variância alta e que estejam correlacionados com os demais. Isso apregoa que escalas com mais pontuações, de 1 a 9, por exemplo, são melhores que escalas de 1 a 4, por exemplo. Isso ocorreria pois maior quantidade de opções de resposta implicaria maior variabilidade .

Por outro lado, existem dúvidas em relação a essas recomendações. Primeiro, será que as pessoas interpretam sempre no mesmo sentido "quantitativo" os números da escala *Likert*, isto é, a interpretação dada à primeira categoria para um item será sempre a mesma para duas pessoas? Segundo, será que a distância entre das categorias 2 para 3 indicaria a mesma intensidade do construto medido independentemente do item que se está respondendo? Ainda, será que as pessoas têm uma compreensão precisa o suficiente para distinguir, por exemplo, nove diferentes gradações de intensidade do construto como se espera quando se constroem escalas *Likert* de resposta de nove pontos?.

Além disso, é possível a ocorrência de inconsistências em função de uma desorganização semântica nos rótulos associados a cada categoria. Tal dificuldade decorre do fato que os rótulos dados aos diversos valores de uma escala likert como, por exemplo, "frequentemente" e "eventualmente", etc., podem ser interpretados de formas variadas. Dependendo da qualidade dos rótulos escolhidos, não apenas o pressuposto de intervalos constantes entre as categorias pode ser comprometido, como também a sua propriedade ordinal.

Tais questões têm intrigado pesquisadores e psicometristas a investigar os fundamentos e as possibilidades para otimização das escalas de avaliação. Dentre os avanços importantes na Psicometria, que trouxeram instrumentos de análise mais refinados para investigar essas questões, estão os modelos de Teoria de Resposta ao Item (TRI) para respostas politômicas . Essa teoria consiste em um conjunto de modelos probabilísticos destinados a representar parâmetros importantes para a mensuração incluindo as características dos itens e as medidas dos sujeitos. A principal diferença em relação ao modelo clássico é que a unidade básica de análise passa a ser o item e não o escore total composto pela soma de itens, como a Teoria do

Escore Verdadeiro trata. Assim, o modelo matemático na teoria clássica estabelece uma relação entre o escore observado e o escore verdadeiro (Escore Observado = Escore Verdadeiro + Erro de Medida). Já na TRI a relação é estabelecida entre o theta e a probabilidade de escolha de uma opção de resposta ao item. Embora a análise clássica leve em conta a correlação item-total, o que poderia sugerir uma unidade de análise também focada no item, os modelos matemáticos subjacentes são diferentes, o primeiro modelando o escore total e o segundo, a probabilidade de resposta ao item. A análise das correlações item-total na teoria clássica tem por objetivo indicar itens que contribuirão para o aumento da variância verdadeira no escore composto pela soma dos itens.

Embora a literatura em psicometria clássica demonstre que itens com mais categorias de resposta e com definições mais claras geralmente levem a resultados mais favoráveis em termos de consistência interna (Dawis, 1992; John & Benet-Martínez, 2000; Weems, 2004; Weng, 2004), outros estudos mais recentes, baseados na TRI, mostram que nem sempre isso ocorre. O que se tem demonstrado é que itens com geralmente 3 a 4 pontos fornecem a mesma informação do que itens com 7 a 9 pontos. Além disso, muitas vezes redundâncias no conteúdo podem inflacionar os coeficientes de precisão sem que isso corresponda a aumento na validade das escalas (Elliott & cols., 2006; Roberts, 1994; Stone & Wright, 1994) Ha ainda estudos com modelos de resposta ideal, mostrando que mesmo itens com baixa correlação item-total tem informação para as escalas (Chernyshenko, Stark, Drasgow, & Roberts, 2007). Como se verá nesse trabalho, a TRI é capaz de detalhar uma série de informações do teste como um todo, dos itens e, por fim, das categorias utilizadas nas escalas para cada item, trazendo informações empíricas mais detalhadas permitindo, com isso, responder à parte das questões levantadas acima.

Na aplicação da TRI para escalas tipo *Likert*, dois modelos freqüentemente usados foram desenvolvidos a partir dos modelos de Rasch : escalas graduadas (*Rating Scale Model*) e créditos parciais . Esses modelos concebem a relação entre as respostas dadas à escala *Likert* com o *theta*, dimensão subjacente inobservável que os itens pretendem estimar, assumindo que cada valor crescente da escala indique um passo cumulativo em direção a valores mais altos na variável latente. A diferença básica entre os dois modelos é que para escalas graduadas presume-se que os avanços nas pontuações *Likert* são constantes e iguais para todos os itens e no modelo de créditos parciais essa condição é relaxada podendo-se configurar diferentes distâncias entre as pontuações *Likert*, dependendo do item a ser con-

siderado. Essa característica do modelo de créditos parciais é interessante, pois, como já foi apontado acima, é difícil sustentar a assunção de constância dos intervalos nas escalas *Likert* como os intervalos numéricos supõem.

O modelo de créditos parciais é dado pela seguinte fórmula:

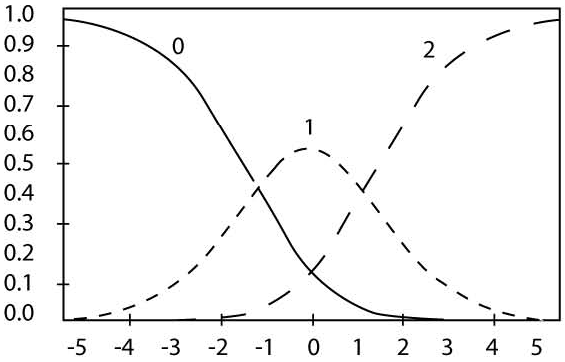
$$P_{nix}(\theta) = \frac{e^{\sum_{j=0}^{x}(\theta_n - \delta_{ij})}}{\sum_{r=0}^{m_i}\left[e^{\sum_{j=0}^{r}(\theta_n - \delta_{ij})}\right]}$$

Onde $\sum_{j=0}^{0}(\theta_n - \delta_{ij}) \equiv 0$,

Os escores no item têm a notação x = 0, ... *mi* para um item com *Ki* = *mi* + 1 categorias de resposta. Assim $P_{nix}(\theta)$ indica a probabilidade do sujeito *n* ter o escore *x* no item *i*. Os valores δ_{ij} (j = 1, ..., m_i) indicam os valores dos limiares de transição entre a categoria *j*-1 e a categoria *j*. Esses valores indicam o ponto de intersecção entre as curvas da categoria *j*-1 e *j*. Em situações ideais, esse ponto indicará a o momento a partir do qual a *j* passa a ser a mais provável, portanto o passo de transição entre a categoria menor *j*-1 e a categoria em consideração a qual esse parâmetro se refere.

Para facilitar a visualização do modelo, considere a Figura 1:

Figura 1. Exemplo das curvas do modelo de créditos parciais para um item com 3 categorias.



No eixo horizontal estão representados os valores de *theta* (dimensão latente medida pelo item) variando de -5 a +5 e, na dimensão vertical, a probabilidade de escolha das alternativas, variando de 0 a 1. Esse gráfico representa um item de 3 pontos (0, 1 e 2). Nota-se que há três curvas, uma para cada alternativa de resposta. Nota-se também que abaixo de *theta* -1 a cate-

goria "0" é a resposta mais provável, já entre -1 e +1, a categoria mais provável é a resposta "1" e acima de +1, a resposta "2". Existe, portanto, uma associação entre o incremento na dimensão subjacente e um correspondente acréscimo na probabilidade de escolha de respostas com pontuações mais altas na escala *Likert*, como se cada reposta indicasse um passo mais adiante na escala subjacente. Assim, à medida que se avança no *theta* medido é sucessivamente mais provável que pontuações mais altas sejam escolhidas.

Os limiares de transição δ_{ij} (j = 1, ..., m_i) entre duas categorias de um item indicam o ponto na dimensão subjacente em que a próxima categoria passa a ser mais provável que a anterior. A diferença básica entre o modelo de créditos parciais e o modelo de respostas graduadas, como já foi referido, é que os intervalos entre os limiares são constantes, condizentes com as representações numéricas das pontuações numéricas das escalas (2 – 1 = 1, 3-2=1 etc..). Nesse modelo, os limiares δ_{ij} são decompostos em dois componentes $\delta_{ij} = (\lambda_i + \delta_j)$. A variável λ_i é um parâmetro geral de localização e as variáveis δ_j são os parâmetros de intersecção entre as categorias que, como se nota, consiste em um número para cada categoria de resposta na escala (índice j) e igual para todos os itens (ausência nesse parâmetro do índice i). Esse modelo é dado pela equação:

$$P_{nix}(\theta) = \frac{e^{\sum_{j=0}^{x}[\theta_n - (\lambda_i + \delta_j)]}}{\sum_{r=0}^{m_i}\left[e^{\sum_{j=0}^{r}[\theta_n - (\lambda_i + \delta_j)]}\right]}$$

Onde, $\sum_{j=0}^{0}[\theta_n - (\lambda_i + \delta_j)] = 0$

No presente estudo, foi feita a aplicação do modelo de créditos parciais na análise de uma escala para avaliação de Socialização (EFS) no modelo dos Cinco Grandes Fatores de personalidade construídas dentro do modelo da Psicometria Clássica. Pretendeu-se depurar as propriedades psicométricas dos itens, especialmente a utilidade, em termos de informação, do uso de escalas *Likert* de sete pontos para a estrutura interna e precisão da escala.

Objetivou-se ainda com o presente estudo aprimorar o conhecimento da escala e, quando possível, otimizar suas propriedades psicométricas, demonstrando-se, portanto, as vantagens do uso da TRI nos procedimentos de análise dos dados na fase da construção e seleção de itens para instrumentos de medida.

MÉTODO

Participantes

A amostra utilizada para as análises foi composta por 1.317 pessoas, com idade média de 21,0 anos (com desvio

padrão de 6,3), sendo que 68% eram mulheres. A coleta de dados foi realizada em cinco estados brasileiros, Santa Catarina, São Paulo, Paraíba, Bahia e Rio Grande do Sul, sendo que esses dois últimos locais representaram, respectivamente, 49,9% e 23,3% da amostra. Os estudantes secundaristas representaram 37,2% da amostra e, entre os universitários, os cursos mais freqüentes foram psicologia e odontologia, com 42,7 e 7,4 % do grupo total, respectivamente.

Instrumentos

O presente estudo utilizou a Escala Fatorial de Socialização – EFS (Nunes & Hutz, 2007), que é um instrumento objetivo, que avalia um componente da personalidade humana a partir do modelo dos Cinco Grandes Fatores . A EFS é composta por 70 itens de auto-relato que descrevem sentimentos, atitudes e opiniões, a partir de assertivas. As pessoas devem indicar em uma escala tipo *Likert* de sete pontos quão bem as assertivas os descrevem. A instrução apresentada aos respondentes é para considerarem, para cada frase, o quão bem os descrevem. Se acharem que as frases os descrevem muito bem, devem marcar o valor "7" na grade de respostas. Se acharem que as sentenças absolutamente não os descrevem adequadamente, devem marcar o valor "1". É explicitamente salientado que todos os valores podem ser marcados e que quanto mais a frase for apropriada para descrevê-los, maior deve ser o valor indicado. Tal instrução é realizada para que o uso da escala seja bem compreendido, uma vez que os únicos rótulos apresentados são nas extremidades.

A Escala Fatorial de Socialização é composta por três subescalas, denominadas Amabilidade (S1), Pró-sociabilidade (S2) e Confiança nas pessoas (S3). O fator S1 agrupa itens que descrevem a disponibilidade para ajudar outras pessoas, uma tendência a ser empático e compreensivo, além de uma postura gentil e educada. Já S2 engloba itens que informam sobre o quão as pessoas aderem a normas sociais, tendência a comportamentos de risco, heteroagressividade e padrões de consumo de bebidas alcoólicas. O fator S3 reúne itens que dizem respeito ao nível de confiança depositado nas outras pessoas e as crenças sobre a existência de más intenções dos outros, assim como tendência a apresentar comportamentos de ciúmes. Escores muito baixos ou muito altos nessas facetas podem indicar padrões de interação com outras pessoas pouco adaptativos em variados contextos.

Procedimentos

A coleta de dados foi coletiva, sendo usualmente realizada nas salas

de aula das instituições de ensino procuradas (escolas de ensino médio, públicas e privadas; universidades públicas e privadas, cursos preparatórios para concursos, etc.). Nas instituições de ensino superior foram escolhidas preferencialmente turmas de disciplinas que reuniam estudantes de vários cursos com o objetivo de obter uma amostra mais diversificada. Os participantes, após serem informados dos objetivos do estudo, de que a sua participação era voluntária e da garantia de sigilo das respostas, receberam o caderno com os itens, a folha de respostas e instruções de preenchimento.

As instruções fornecidas seguiram um roteiro pré-estabelecido e foram lidas pelos aplicadores. Foi solicitado aos participantes que lessem os itens com atenção e que respondessem individualmente às questões. Também foi informado que não havia respostas certas ou erradas e que realmente era importante que dessem sua opinião sincera às situações, sentimentos e atitudes descritas nos itens. Àquelas turmas que apresentaram dificuldades para a compreensão dos itens, foi dada a orientação para que os deixassem em branco. Todos os participantes assinaram o termo de consentimento livre e esclarecido para a participação em pesquisa.

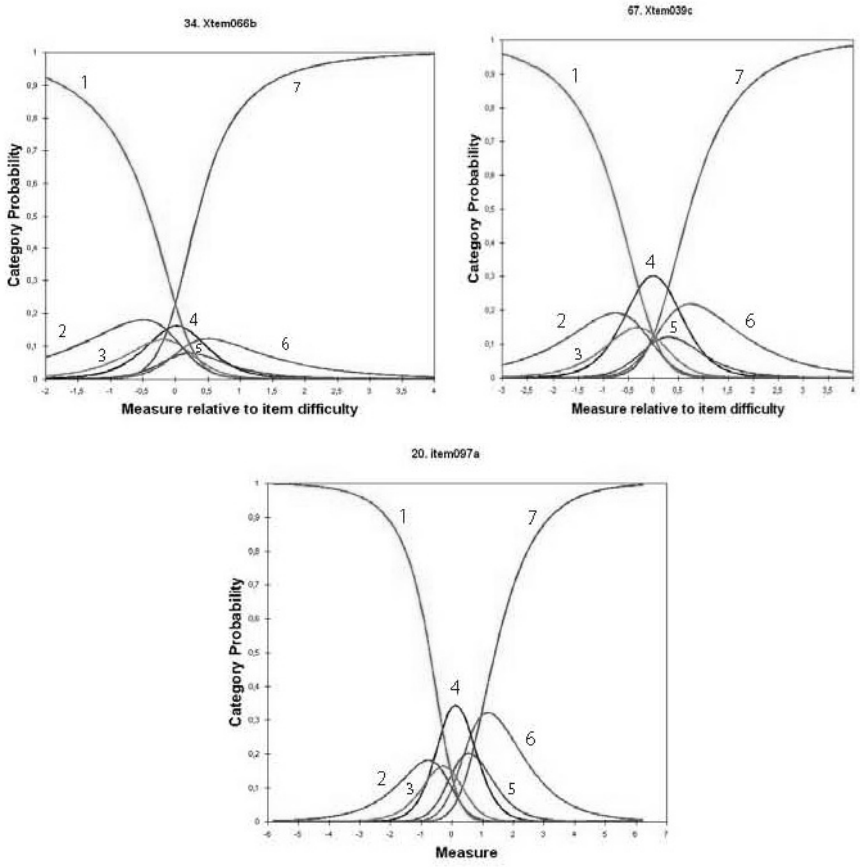
RESULTADOS E DISCUSSÃO

Os itens que compuseram as subescalas da EFS foram analisados por meio da técnica de Créditos Parciais utilizando o *software* Winsteps, que é capaz de gerar as curvas de probabilidade de cada alternativa, chamadas neste tipo de análise de categorias e informações sobre a precisão das escalas. Também são apresentadas informações sobre o resíduo (*misfit*) da escala como um todo, dos itens e das categorias utilizadas para cada item, além de dados sobre a adequação entre o nível de habilidade do sujeito e o seu padrão de resposta, isto é, qual categoria de resposta ele escolheu . A análise da Escala Fatorial de Socialização deu-se em algumas etapas, descritas abaixo:

Análise das categorias dos itens

A análise das curvas de probabilidade das categorias dos itens foi realizada para verificar quais delas efetivamente estavam contribuindo ao trazer informações sobre quantidades crescentes no construto latente medido. Com esse procedimento, foi possível verificar que haviam itens com padrões muito diversificados, apresentando um número variado de categorias realmente úteis. A ilustra tal resultado, apresentando os gráficos dos itens 34, 67 e 20. Nesses gráficos, o eixo x representa o *theta*, isto é, a variável latente medida, e o eixo y a probabilidade de aderência às categorias da escala.

Figura 2. Curvas de probabilidade originais das categorias dos itens selecionados



É possível verificar com na figura 2, referente às curvas do item 34, que a probabilidade de ocorrência da categoria "1" é maior que as demais para *theta´*s inferiores a aproximadamente 0 e que essa categoria cobre quase que totalmente a área avaliada pelas categorias "2" e "3". Isto significa que, para nenhum nível de *theta*, as categorias 2 e 3 foram as mais prováveis de serem indicadas pelos participantes e que, portanto, essas categorias se mostram pouco informativas sobre o valor do *theta* das pessoas que a escolhem. A análise da curva de probabilidade da categoria "7" indica que, para *theta´*s

superiores a zero, esta é a mais provável. É possível analisar as categorias independentemente das demais. Assim, por exemplo, apesar da categoria 2 em nenhum ponto da escala de *theta* ser a mais provável, verifica-se que esta teve maior probabilidade de aderência para um *theta* de aproximadamente -0,5. A mesma lógica pode ser utilizada para as demais categorias, indicando que há sobreposição das áreas de *theta* a que essas categorias estão associadas em termos de probabilidade relativa de escolha e os pontos que são mais elevados.

Por outro lado, na figura 2 (item 67), é possível notar que na curva da categoria 4, as regiões ao redor do *theta* 0 a 0,4, esta opção é a mais provável de ser escolhida, produzindo com isso mais informação sobre a localização dos os sujeitos que a escolhem. Esse caso demonstra uma categoria de resposta informativa do ponto de vista psicométrico. Portanto, resumidamente, esse procedimento analítico, aliado aos que serão descritos em seguida, permitiu uma análise visual da eficácia de cada categoria de cada item em termos de informação útil sobre a localização dos sujeitos que são atraídos a escolhê-las.

Também foram verificadas as tabelas apresentadas pelo Winsteps sobre as características das categorias utilizadas. A Tabela 1 apresenta um sumário da estrutura das categorias do item 20 e, na seqüência, são explicadas as informações presentes na mesma. Essas informações são apresentadas para o itens 34 na e para o item 62 na Tabela 3. Esses dados serão analisados posteriormente, com a sua comparação com a análise das escalas otimizadas.

Tabela 1. Sumário da estrutura original das categorias do item 20

Categoria	Observado		Média observada	Média esperada	Infit	Outfit	Calibração da estrutura	Medida da categoria
	Freq.	%						
1	44	3	.51	-.02	1.88	3.57	-	(-1.61)
2	36	3	*.12	.21	.82	.84	.08	-.75
3	72	5	.42	.39	1.12	1.34	-.61	-.27
4	258	20	.54	.55	1.06	1.30	-1.02	.12
5	203	15	.61	.72	.76	.61	.66	.56
6	317	24	.88	.94	.92	.83	.16	1.19
7	369	28	1.33	1.28	.88	.92	.73	(2.40)

Cont. Tabela 1. Sumário da estrutura original das categorias do item 20

Categoria	Estrutura		Escore para medida		50% Probabilidade acumulada	Coerência M → C	Coerência C → M	Estim. Discr.
	Medida	Erro padrão	Zona					
1	-	-	-INF	-1.17	-	100%	2%	-
2	.30	.17	-1.17	-.49	-.71	9%	2%	-1.87
3	-.39	.13	-.49	-.07	-.42	19%	12%	.38
4	-.81	.10	-.07	.33	-.17	37%	24%	.86
5	.87	.07	.33	.84	.38	21%	53%	.53
6	.38	.06	.84	1.78	.72	28%	42%	1.01
7	.94	.07	1.78	+INF	1.41	84%	20%	1.35

Tabela 2. Sumário da estrutura original das categorias do item 34

Categoria	Observado		Média observada	Média esperada	Infit	Outfit	Calibração da estrutura	Medida da categoria
	Freq.	%						
1	59	5	.32	.11	1.35	2.72	-	(-1.30)
2	39	3	*.22	.19	1.02	1.66	0.63	-.56
3	48	4	.3	.27	1.04	1.31	0.09	-.25
4	102	8	.41	.36	1.09	1.57	-0.37	-.04
5	70	5	*.36	.45	1.09	1.32	0.84	.16
6	134	10	.43	.56	1.14	0.61	-0.08	.44
7	825	63	.73	.72	1.02	1	-1.11	-1.04

Cont. Tabela 2. Sumário da estrutura original das categorias do item 34

Categoria	Estrutura		Escore para medida		50% Probabilidade acumulada	Coerência M → C	Coerência C → M	Estim. Discr.
	Medida	Erro padrão	Zona					
1	-		-INF	-.90		0%	0%	
2	.56	.14	-.90	-.38	-.46	25%	5%	-.05
3	.02	.12	-.38	-.14	-.23	25%	8%	-.75
4	-.44	.10	-.14	.06	-.11	11%	7%	.75
5	.78	.08	.06	.28	.06	9%	28%	.51
6	-.15	.07	.28	.72	.14	12%	54%	.81
7	-1.18	.06	.72	+INF	.27	86%	43%	1.00

Tabela 3. Sumário da estrutura original das categorias do item 67

Categoria	Observado		Média observada	Média esperada	Infit	Outfit	Calibração da estrutura	Medida da categoria
	Freq.	%						
1	51	4	-.13	-.18	1.07	1.11	-	(-1.85)
2	45	3	-.08	-.07	.98	1.02	.24	-.99
3	73	6	.04	.03	1.01	1.12	-.26	-.56
4	240	18	.08	.13	.84	.73	-.86	-.25
5	127	10	.2	.24	.89	.72	1.07	.06
6	240	18	.4	.37	.72	.8	-.09	.49
7	535	41	.56	.54	.96	.98	-.11	-1.37

Cont. Tabela 3. Sumário da estrutura original das categorias do item 67

Categoria	Estrutura		Escore para medida		50% Probabilidade acumulada	Coerência M → C	Coerência C → M	Estim. Discr.
	Medida	Erro padrão	Zona					
1	-		-INF	-1.41		0%	0%	
2	.00	.15	-1.41	-.75	-.96	16%	2%	.55
3	-.50	.11	-.75	-.40	-.65	5%	2%	.90
4	-1.11	.09	-.40	-.10	-.45	32%	17%	.98
5	.82	.07	-.10	.25	-.04	14%	48%	1.10
6	-.33	.06	.25	.91	.13	23%	55%	1.15
7	-.35	.06	.91	+INF	.47	79%	20%	1.06

Na metade superior da Tabela 1, nas primeiras três colunas são apresentados os valores das pontuações das categorias (*category label*), a freqüência de sujeitos que escolheram cada uma das categorias (*Observed count*) e a porcentagem correspondente (*%*). Em seguida, apresenta-se a média dos *thetas* dos sujeitos que escolheram cada uma das categorias (*Observed Average*) e é esperado que a média dos *thetas* aumente com o valor da categoria. Quando há alguma desordem neste parâmetro, isto é, quando a média de *theta* das pessoas que escolhem uma dada categoria não aumenta de maneira progressiva ao aumento da pontuação que supostamente se espera para cada categoria, esta é indicada com um asterisco ao lado do valor. A coluna Média Esperada (*Sample Expect*) apresenta o valor das médias de *theta* esperadas para cada categoria a partir do modelo.

Infit é uma medida que indica o nível de ajustamento dos padrões de respostas, sensível em categorias com valores de dificuldade próximos aos valores de *theta* da pessoa. Espera-se que valores neste parâmetro sejam próximos de um, sendo que valores substancialmente abaixo de 0,7 indicam que o dado empírico apresenta valores de discriminação superiores aos esperados pelo modelo de Rasch; valores substancialmente acima de 1,3 indicam ruído, ou seja, aponta para uma grande quantidade de respostas inesperadas. *Outfit* também é uma medida de ajuste, sensível a

padrões inesperados de respostas quando a diferença entre o *theta* das pessoas e a dificuldade das categorias é muito grande, ou seja, quando uma pessoa com um *theta* muito alto adere a uma categoria com dificuldade baixa ou vice-versa. Para este parâmetro, também são esperados valores próximos de um.

As colunas Calibração da estrutura (*Structure Calibration*) e Medida da estrutura (*Structure Measure*) são as medidas de transição entre categorias contíguas, isto é os limiares entre as categorias. Esses parâmetros representam o *Rasch-Andrich threshold,* ou simplesmente *threshold*, que são pontos em que as probabilidades de categorias adjacentes são as mesmas. A diferença entre esses dois parâmetros é que, o primeiro, é relativo ao índice de dificuldade dos itens definido como a média dos limiares brutos. Assim á média desses valores é igual a zero. No segundo caso se tem os valores originais das transições.

É esperado que esses valores sejam crescentes desde a transição da primeira até a da última categoria. A desordem dessas estimativas, ou seja, se elas não aumentam o valor à medida que mudam as categorias, pode ser causada pela baixa freqüência de observação da mesma ou por um problema inerente à interpretação ou organização das categorias apresentadas, o que pode sugerir problemas substanciais observados naquela categoria para uma adequada medida do construto avaliado. Vale destacar que o *threshold* da primeira categoria sempre será inexistente, uma vez que não há uma transição anterior.

A leitura da indica que existe desordem no item 20 em relação ao *threshold* às duplas de categorias dois e três, três e quatro, cinco e seis pois as categorias de menor pontuação apresentam um valor superior ao *threshold* subseqüente. Essa descontinuidade também pode ser verificada observando-se os valores para as três primeiras categorias no campo Média observada (*Observed Average*), que deveriam ser crescentes.

Já a Medida da categoria (*Category Measure*) representa o valor de *theta* associado à categoria, isto é, a magnitude de theta que cada categoria implica a partir do modelo de Rasch. Os parênteses indicam que a calibração correspondente tende ao infinito, ou seja, que seus valores são a menor representação possível da primeira categoria ou a maior representação possível da última categoria em uma dada escala de maneira aproximada.

Na parte inferior da Tabela 1, são apresentadas informações adicionais sobre as relações categorias-medida e sobre a adequação do modelo. As colunas Escore para medida (*Score-to-Measure*) são valores usados na conversão entre os valores brutos das categorias e os valores da medida na escala *theta*. A coluna 50% Probabilidade Acumulada (*Cumulative probability*) apresenta limiares a partir de outra

definição chamada *Rasch-Thurstone Thresholds*. Os pontos indicados representam os locais em que se tem chances iguais de se observar as categorias menores em comparação a categoria atual ou aquelas acima dela. Esses são os pontos onde em *theta* em que os intervalos das categoria iniciam.

As colunas sobre Coerência (*Coherence* M⟶C e C⟶M) indicam, respectivamente, qual o percentual de observações de uma dada categoria de fato previstas pelo *theta* e pelas relações *theta*-categoria mencionadas acima (M⟶C) ou, ao contrário, o percentual de medidas de *theta* de fato observadas partindo-se das categorias observadas e das relações *theta*-categorias (C⟶M). Espera-se que os valores percentuais dessas coerências sejam próximos de 100%. É interessante notar que, na Tabela 2, os valores apresentados nesses campos são extremamente baixos para a categoria 2. Tal resultado está associado à desordem observada no *threshold* para esta categoria. A última coluna traz a informação aproximada sobre a discriminação local considerando o modelo de dois parâmetros.

Otimização das escalas

Com as informações provenientes da primeira etapa da análise, foi possível verificar, para cada item, quais categorias estavam explicando mais adequadamente as respostas observadas na amostra. Todos os itens da EFS apresentaram características que justificavam o agrupamento de categorias, que é recomendado quando a probabilidade de ocorrência de uma dada categoria não é superior às das demais em toda a faixa de *theta* coberta pelo teste. Esses resultados são similares ao que outros estudos semelhantes têm encontrado (Elliott & cols. 2006; Roberts, 1994; Stone & Wright 1994).

A otimização dos itens foi realizada com o agrupamento de categorias adjacentes cujas curvas de informaçõ-es fossem muito próximas, o que também pode ser visualizado com as curvas de probabilidade das categorias, quando o valor máximo de algumas categorias ocorre em *thetas* muito próximos. Tal condição pode ser verificada na , na qual são apresentadas as curvas de probabilidade das categorias dos itens utilizados como exemplo. A análise dos gráficos sugere que as categorias podem ser mescladas de formas diferentes, utilizando dois, três e quatro categorias para os itens 34, 67 e 20, respectivamente.

É importante notar que, para cada valor de *theta* apresentado na Figura 2, a soma das probabilidades das categorias será igual a 1, supondo que para toda magnitude do traço avaliado, as únicas respostas possíveis são as existentes na escala. Por esse motivo, quando é feito o agrupamento de duas ou mais categorias, a curva da categoria resultante terá para cada *theta*, aproximadamente, a altura

somada daquelas que a compuseram. No entanto, as curvas das novas categorias podem não ser simplesmente a sobreposições das categorias que as geraram, pois os dados apresentados são resultados da modelagem das respostas observadas, de modo que a junção de várias categorias pode gerar novos modelos.

As escalas, que eram originalmente compostas por sete categorias de respostas, com valores entre 1 e 7 pontos, variaram entre duas a quatro categorias após a otimização. Essa otimização foi realizada por meio do procedimento de recodificação das categorias de resposta, no qual o valor original das categorias é transformado, a partir de um padrão definido. Os padrões de recodificação utilizados são apresentados na , assim como a quantidade de itens que foram recodificados em cada padrão.

Tabela 4. Padrões de respostas utilizados para otimização

Escala	Padrão		Quantidade de itens
S1 – Amabilidade	Orig	1234567	
	A	1112224	4
	B	1112244	15
	C	1112334	4
	D	1122334	10
S2 – Pró-Sociabilidade	Orig	1234567	
	E	1111333	12
	F	1112233	11
S3 - Confiança	Orig	1234567	
	G	1112244	8
	H	1122334	6

Um aspecto que foi amplamente discutido, inclusive com o autor do *software* Winsteps (Linacre, comunicação pessoal), diz respeito aos valores mínimo e máximo para as escalas otimizadas. Uma possibilidade seria, para cada padrão de resposta, utilizar como valor inicial a categoria "1" e atribuir às categorias subseqüentes valores numéricos aumentando em um ponto. Neste caso, por exemplo, itens com duas categorias de respostas teriam como valor mínimo de categoria o "1" e máximo igual a "2". Itens com três categorias iniciariam com "1" e teriam como valor máximo o "3" e assim sucessivamente. A principal dificuldade para a utilização desse

método é que, quando utilizadas simultaneamente na mesma escala, itens com padrões diferentes colaborariam de forma variada para o escore bruto total da mesma. Assim, itens com quatro categorias diferentes teriam o dobro do peso de itens dicotômicos no cálculo do escore geral. Tal característica não chega a ser um problema de medida, uma vez que todos os itens contribuem independentemente para avaliar alguma magnitude do construto e, de qualquer forma, não se pode garantir de antemão que todos os itens apresentam efetivamente a mesma contribuição na representação do construto mensurado e, por esse motivo, teriam que apresentar o mesmo peso.

Em contrapartida, como não haviam sido realizados estudos que apresentassem qualquer informação sobre a eficácia dos itens para detectar diferentes magnitudes dos traços avaliados, optou-se por adotar escalas cujas categorias apresentassem valores mínimo e máximo iguais para cada sub-escala. Como decorrência disso, o número de ocorrências de algumas categorias (COUNT) fica igual a zero, para as categorias intermediárias na versão otimizada sem associação com categorias da escala original, a exemplo dos itens 34 e 67.

As tabelas 5 a 7 apresentam as características psicométricas das classes após a sua otimização e a apresenta as suas curvas de probabilidade otimizadas. O item 20, por exemplo, que pertence a escala S1 e originalmente possui sete categorias de respostas, após a otimização ficou com 4 categorias no padrão D, conforme.

Tabela 5. Sumário da estrutura otimizada das categorias do item 20

Categoria	Observado		Média observada	Média esperada	Infit	Outfit	Calibração da estrutura	Medida da categoria
	Freq.	%						
1	80	6	.24	-.08	1.31	1.53	-	(-2.34)
2	330	25	.57	.61	.99	1.03	-1.38	-.48
3	520	40	.91	1.02	.88	.75	.08	1.08
4	369	28	1.61	1.48	.84	.87	1.30	(2.86)

Cont. Tabela 5. Sumário da estrutura otimizada das categorias do item 20

Categoria	Estrutura		Escore para medida		50% Probabilidade acumulada	Coerência M → C	Coerência C → M	Estim. Discr.
	Medida	Erro padrão	Zona					
1	-		-INF	-1.50		54%	7%	
2	-1.10	.13	-1.50	.31	-1.28	48%	28%	.82
3	.36	.07	.31	2.05	.33	45%	84%	.89
4	1.58	.07	2.05	+INF	1.80	79%	24%	1.41

Tabela 6. Sumário da estrutura otimizada das categorias do item 34

Categoria	Observado		Média observada	Média esperada	Infit	Outfit	Calibração da estrutura	Medida da categoria
	Freq.	%						
1	248	19	.41	.25	1.15	1.48	-	(-1.02)
2	0	0			.00	.00	NULL	-.05
3	1009	78	1.06	1.10	1.21	1.21	.00	(.92)

Categoria	Estrutura		Escore para medida		50% Probabilidade acumulada	Coerência M → C	Coerência C → M	Estim. Discr.
	Medida	Erro padrão	Zona					
1	-		-INF	-60		66%	5%	
2	NULL		-.60	.50	-.05	0%	0%	1.00
3	-.05	.08	.50	+INF	-.05	88%	78%	.55

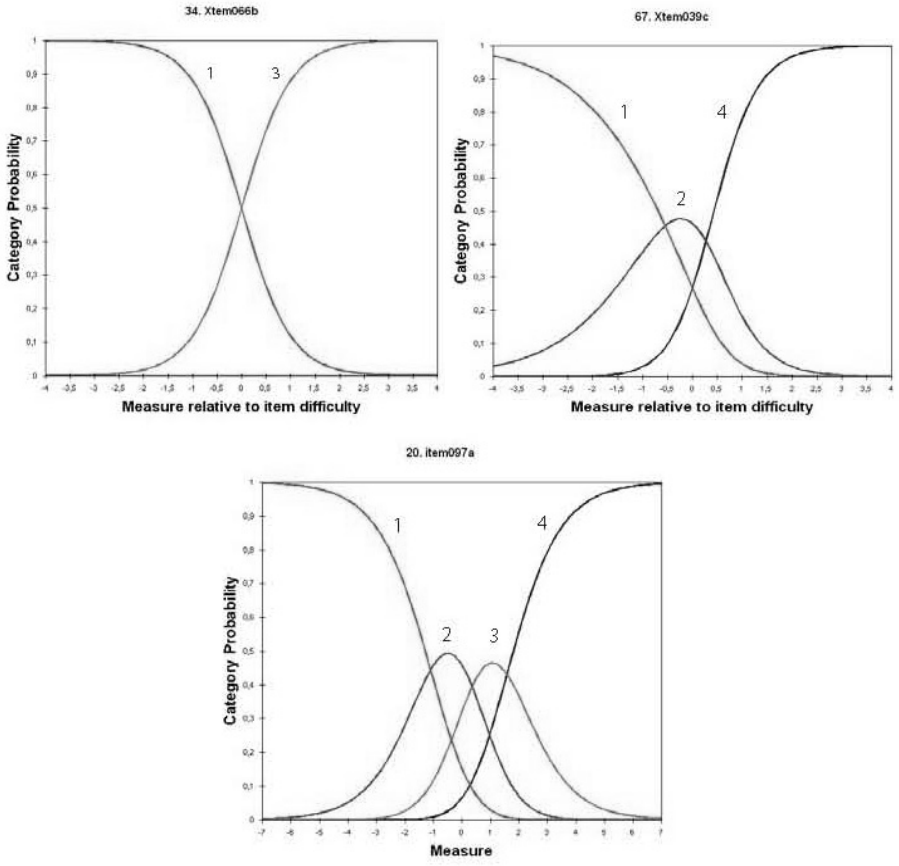
Tabela 7. Sumário da estrutura otimizada das categorias do item 67

Categoria	Observado		Média observada	Média esperada	Infit	Outfit	Calibração da estrutura	Medida da categoria
	Freq.	%						
1	169	13	-.31	-.33	1.02	1.09	-	(-2.10)
2	367	28	-.05	-.01	.91	.76	-.55	-.62
3	0	0			.00	.00	NULL	.14
4	775	59	.57	.56	.95	.98	.55	(.96)

Categoria	Estrutura		Escore para medida		50% Probabilidade acumulada	Coerência M → C	Coerência C → M	Estim. Discr.
	Medida	Erro padrão	Zona					
1	-		-INF	-1.32		66%	8%	
2	-.94	.09	-1.32	-.22	-1.08	49%	32%	.98
3	NULL		-.22	.58	.04	0%	0%	1.00
4	.15	.07	.58	+INF	.04	89%	40%	1.25

Ao se comparar os valores dos parâmetros da estrutura original das categorias com os valores para a estrutura otimizada dos itens 20, 34 e 67 (Tabelas 1, 2, 3 e Tabelas 5, 6, 7, respectivamente), pode-se notar que alguns parâmetros mostraram melhoras substanciais. Em primeiro lugar, os valores de dificuldade dos *steps* (valores dos *thresholds*) dos três itens na estrutura original apresentam desajustes, pois seus valores não apresentaram aumento em relação a progressão dos valores de algumas categorias. Já nas estruturas otimizadas, pode-se observar que apresentam um crescimento ordenado em relação às mudanças de categoria. Em segundo lugar, os erros padronizados de mensuração são menores na estrutura otimizada das categorias. Em terceiro lugar, observou-se que os valores de ajuste representados pelos parâmetros de *infit* e *outfit* apresentaram redução em quase todos os casos. Apesar desses valores terem indicado desajuste em apenas algumas categorias na estrutura original, na estrutura otimi-

Figura 3. Curvas de probabilidade otimizadas das categorias dos itens selecionados



zada os valores são mais adequados. Em último lugar, podemos observar um aumento nos valores percentuais dos parâmetros de coerência. No caso da capacidade de previsão da medida observada por meio da medida esperada (MC) os valores aumentaram em média 14%, 53% e 44% respectivamente, para os itens 20, 34 e 67. Já no caso de adequação da medida observada à medida esperada (CM), pode-se observar um aumento médio de 14%, 21% e 6%, respectivamente, para os itens 20, 34 e 67.

Com relação ao conteúdo abordado nos itens, é importante ressaltar

que este pode favorecer a verificação de categorias extremas ou de categorias intermediárias entre um pólo e outro da adesão aos itens. Para fins didáticos, será detalhado apenas o conteúdo dos itens já analisados anteriormente. O item 34 representa um exemplo de item que foi transformado em dicotômico ("Gosto muito de ter relações sexuais incomuns"), e seu conteúdo avalia um aspecto em que a população geral apresenta baixa aderência. Já os itens 67 ("Tenho um grande interesse pelas pessoas") e 20 ("Sou amaìvel com as pessoas") avaliam traços da personalidade em magnitudes menos extremas, o que suscitou uma maior flexibilidade de comportamento, o que é refletido em uma menor adesão a categorias extremas. De uma maneira geral, o único subfator em que os itens foram otimizados como dicotômicos, S2, é aquele que apresenta um maior número de itens que retratam características ou situações associadas a quadros clínicos.

Análise das características psicométricas dos subfatores da EFS

Após a otimização dos itens, foi feita a comparação das características psicométricas dos subfatores da EFS com a versão original. O programa Winsteps apresenta muitas informaçõ-es sobre as escalas, mas foram selecionadas apenas aquelas essenciais aos objetivos desta investigação.

Nas Tabelas 8, 9 e 10, o Escore bruto é calculado pela soma das respostas dadas aos itens que formam as escalas, após a inversão dos itens correspondentes. As colunas *Theta*, *Infit* e *Outfit* são interpretadas conforme as explicações já apresentadas e a precisão do instrumento é avaliada a partir de dois indicadores, dois coeficientes de consistência interna a partir das estimações da variância de erro pelo modelo (Precisão Modelo) e pelo erro observado (Precisão real) e a separação. Separação é a razão entre o desvio padrão ajustado (ADJ.S.D.) da pessoa ou do item, que é uma estimativa do desvio padrão verdadeiro, em relação ao RMSE, que é o erro de medida em unidades de desvio padrão. Fornece uma medida de razão de separação em unidades RMSE, que é mais fácil de interpretar que a precisão por correlação.

As medidas de consistência interna que são calculados a partir dos dados observados são apresentados com o rótulo "real", enquanto que aquelas que foram calculadas a partir do *theta* dos participantes são referidas como "modelo".

Análise de S1 - Amabilidade

Na são mostradas as estatísticas descritivas dos 1313 sujeitos participantes da pesquisa, com relação à escala S1 – Subfator Amabilidade, composto por 33 itens, bem como informações sobre a precisão da escala.

Tabela 8. Características estatísticas das respostas das pessoas à escala de Amabilidade

	Escala original				Escala otimizada			
	Escore Bruto	**Theta**	**Infit**	**Outfit**	**Escore Bruto**	**Theta**	**Infit**	**Outfit**
Média	187.6	0.84	1.16	1.13	102.7	0.98	1.07	1.08
Desvio Padrão	27.0	0.63	0.73	0.74	16.8	0.82	0.39	0.64
Máximo	229.0	3.44	5.35	6.53	131.0	3.95	3.06	9.42
Mínimo	42.0	-1.64	0.07	0.08	35.0	-3.96	0.30	0.26
Alpha	Real 0.86		Modelo 0.91		Real 0.89		Modelo 0.91	
Separação	Real 2.51		Modelo 3.19		Real 2.78		Modelo 3.15	

Comparando os dados entre a análise original e os resultados da análise com a escala otimizada, pode-se observar que a média de *theta* calculada por meio da análise de Rasch aumentou na escala otimizada, assim como o desvio padrão também indica maior variabilidade dos escores. Com relação aos índices de ajuste *Infit* e *Outfit* nota-se que na análise otimizada eles apresentam um melhor ajuste. Nessa segunda análise, o índice de precisão foi maior, passando de 0,86 para 0,89. A comparação dos resultados indica que a análise otimizada mostrou-se mais adequada, uma vez que houve ganhos em parâmetros importantes do itens. Desse modo, mesmo com a diminuição do número de categorias, a variabilidade das respostas foi contemplada e houve uma maior organização das mesmas.

Destaca-se ainda que a diminuição observada entre os escores brutos ocorreu devido a diminuição da amplitude das categorias de respostas, ou seja, a análise inicial considerava respostas que variavam entre um e sete e a segunda análise, respostas variando entre um e quatro. A mesma situação ocorreu com a otimização das duas outras sub-escalas.

Análise de S2 – Pró-sociabilidade

Na Tabela 9 são apresentadas as informações dos 1.307 participantes da pesquisa, referentes à escala S2 – Subfator Pró-sociabilidade formado por 23 itens e as informações sobre a consistência interna da escala originalmente e após a sua otimização.

Tabela 9. Características estatísticas das respostas das pessoas à escala de Pró-sociabilidade

	Escala original				Escala otimizada			
	Escore Bruto	**Theta**	**Infit**	**Outfit**	**Escore Bruto**	**Theta**	**Infit**	**Outfit**
Média	128.0	0.60	1.11	1.09	56.9	0.93	1.02	1.01
Desvio Padrão	20.1	0.46	0.56	0.70	8.1	0.78	0.29	0.61
Máximo	160.0	2.85	3.90	6.18	68	2.91	2	6.51
Mínimo	43.0	-0.69	0.17	0.18	26	-1.85	0.34	0.18
Alpha	Real 0.75		Modelo 0.82		Real 0.71		Modelo 0.75	
Separação	Real 1.73		Modelo 2.13		Real 1.57		Modelo 1.72	

Ao comparar os dados da análise original com os da análise da escala otimizada, verifica-se que a média de *theta* e seu desvio padrão aumentaram na escala otimizada, e os índices de ajuste *Infit* e *Outfit* tornaram-se mais favoráveis. A precisão da escala sofreu uma pequena queda, inferior a 0,1, o que não é substancial, uma vez que os ganhos gerais foram favoráveis à análise otimizada.

Confrontando os dois momentos da análise, encontraram-se dados que sustentam as vantagens do segundo modelo, especialmente no que tange aos índices de ajuste *Infit* e *Outfit*, e o aumento na variabilidade do *theta*.

Análise de S3 – Confiança nas pessoas

Na são apresentadas as informações dos 1.307 participantes da pesquisa, referentes à escala S3 – confiança nas pessoas, formado por 14 itens e as informações sobre a consistência interna da escala originalmente e após a sua otimização.

Tabela 10. Características estatísticas das respostas das pessoas à escala de Confiança

	Escala original				Escala otimizada			
	Escore Bruto	**Theta**	**Infit**	**Outfit**	**Escore Bruto**	**Theta**	**Infit**	**Outfit**
Média	68.0	0.33	1.04	1.05	38.0	0.28	1.04	1.05
Desvio Padrão	13.4	0.45	0.57	0.67	7.8	0.68	0.44	0.66
Máximo	97.0	3.05	4.27	6.62	55.0	3.21	3.41	8.51
Mínimo	23.0	-1.10	0.11	0.14	15.0	-3.39	0.11	0.11
Alpha	Real 0.76		Modelo 0.81		Real 0.74		Modelo 0.79	
Separação	Real 1.77		Modelo 2.07		Real 1.71		Modelo 1.94	

Nessa análise entre a escala original e a escala otimizada, novamente observa-se que para a escala otimizada foram obtidos resultados mais adequados com relação à variabilidade do *theta*, que pode ser observada no aumento do valor do DP e da amplitude e apesar de as médias do *infit* e *outfit* serem iguais nos dois casos, os DP's e a amplitude do *infit* são menores na escala otimizada. Apesar de a precisão ter diminuído na escala otimizada, este valor não é substancial.

Realizando-se uma análise dos sumários dos itens das três escalas nos dois momentos, inicial e otimizada, pode-se constatar que, apesar de não ter sido observado mudanças substanciais nos parâmetros destacados, os resultados demonstram vantagens nos indicadores de *misfit,* o que sugere que a escala otimizada conseguiu diminuir a quantidade de categorias de respostas, mantendo a qualidade psicométrica dos itens e garantindo que mesmo com menos opções de respostas é possível contemplar o universo de comportamento que a escala se propõe a medir.

CONSIDERAÇÕES FINAIS

O presente artigo teve como objetivo apresentar algumas das possibilidades da utilização da Teoria da Resposta ao Item em escalas politômicas. Foi possível verificar que alguns pressupostos utilizados na análise clássica podem ser questionados, como a relação direta entre valores de categorias e a magnitude do traço mensurado, bem como a idéia que aumento da variância dos itens corresponde a um aumento da precisão. Em relação a este aspecto, foi possível observar que, mesmo a utilização de escalas tipo *Likert* com âncoras nas extremidades não garantiu que todas as categorias apresentassem uma organização ordinal.

Também foi possível verificar que a otimização das categorias utilizadas para a avaliação de itens pode aumentar, em alguns casos, a consistência interna das respostas dadas a um instrumento, bem como diminuir medidas de resíduos, como o *infit* e *outfit*. Isso contradiz diretamente a assunção da psicometria clássica que o aumento do número de categorias da escala *Likert* aumentaria a variância e, conseqüentemente, a precisão da medida . De fato, isso só ocorre quando há um aumento de categorias que reflitam magnitudes diferentes e organizadas conforme o aumento do valor da escala, conforme se analisou nesse artigo. Como apontado no texto, a utilidade de uma maior quantidade de pontos em uma escala depende do conteúdo do item e da capacidade dos sujeitos de interpretarem nuances no construto. Por exemplo, algumas questões evocam respostas dicotômicas enquanto que outras geram respostas que se refletem em um número maior de categorias . No presente estudo, no entanto, os itens avaliados foram reorganizados com, no máximo, quatro categorias diferentes, indicando uma dificuldade dos participantes discriminarem nuances em sete pontos, o que parece ser um achado freqüente nos estudos que aplicam a TRI a análise de itens politômicos (Elliott e cols. 2006, Roberts, 1994, Stone & Wright 1994).

Uma questão ainda pode ser feita quanto ao formato dos itens que foram analisados nesse estudo, contendo descrições semânticas somente nas extremidades. Embora esse esquema tenha vantagens como se mencionou anteriormente, pode-se questionar se isso não favorece a pouca utilização dos pontos intermediários da escala, uma vez que não apresentam referências aos níveis que referem. O único estudo encontrado sobre essa questão, de Weng (2004), estudou a influência na precisão (consistência interna e teste-reteste) empregando a psicometria clássica. Portanto, em estudos futuros seria interessante verificar o efeito da ancoragem semântica dos pontos das escalas nos parâmetros estimados pela TRI no que diz respeito à utilidade das categorias.

Em alguns casos do presente estudo não foi constatada a melhora da precisão da escala ao otimizá-la reduzindo o número de categorias. No entanto, a variância das escalas aumentou após a otimização, o que sugere que não é necessário construir uma medida com muitas categorias de respostas para avaliar um determinado construto em toda a sua extensão. Também foi constatado que, dependendo do aspecto a ser mensurado, pede-se maior ou menor número de categorias. Essa é uma outra questão interessante para estudos futuros, ou seja, quais e como as características semânticas dos itens interferem nas propriedades psicométricas dos mesmos.

Em estudos futuros, pretende-se aplicar a EFS contando com um menor número de categorias de resposta, como foi sugerido nesse estudo, para investigar se os benefícios da diminuição de categorias, sugeridos pela análise por créditos parciais, pode ser replicado empiricamente. A questão que fica em aberto é se, diante de itens com um menor número de categorias, as pessoas responderão com padrões semelhantes aos os que foram identificados neste estudo, com as informaçõ-es das análises por créditos parciais.

Por fim, deve-se citar que a utilização da TRI para itens politômicos também apresenta outras vantagens em relação à teoria clássica dos itens, que fogem aos objetivos desse artigo mas que devem ser enumerados: a. possibilita a equalização de itens, ou seja, que resultados obtidos a partir de escalas diferentes, mas calculadas deliberadamente com a mesma métrica, sejam diretamente comparados; b. permite a criação e atualização de bancos de itens para a avaliação de construtos, o que permite a elaboração de múltiplas formas de testes. Com isso, são resolvidas questões referentes a aprendizagem de itens dos testes e divulgação de gabaritos em provas de alto impacto (como concursos públicos, avaliações educacionais de desempenho, certificação ocupacional, etc.); c. possibilita a aplicação adaptativa de provas, composta por itens que se aproximam mais da magnitude do traço latente apresentado pelas pessoas, aumentando assim a precisão dos resultados .

REFERÊNCIAS BIBLIOGRÁFICAS

Briggs, S. R. (1992). Assessing the Five-Factor Model of personality description. *Journal of Personality*, 60, 253-293.

Costa, P. T., Jr. & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and Five Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing Personality Scales Under the Assumptions of an Ideal Point Response Process: Toward Increasing the Flexibility of Personality Measures. Psychological Assessment, 19, 88-106.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

Dawis, R. V. (1992). Scale construction. In A. E. Kazdin (Ed.), *Methodological issues & strategies in clinical research*. (pp. 193-213). Washington, DC: American Psychological Association.

Digman, J. M. (2002). Historical Antecedents of the Five-Factor Model. In P. T. Costa & T. A. Widiger (Eds.), *Personality Disorders and the Five-Factor Model of Personality*. (2 ed., pp. 17-22). Washington, DC: American Psychological Association.

Elliott, R., Fox, C. M., Beltyukova, S. A., Stone, G. E., Gunderson, J., & Zhang, X. (2006). Deconstructing Therapy Outcome Measurement With Rasch Analysis of a Measure of General Clinical Distress: The Symptom Checklist-90-Revised . Psychological Assessment, 18, 359-372.

Embretson, S. & Reise, S. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Goodwin, L. D. & Leech, N. L. (2006). Understanding Correlation: Factors That Affect the Size of r. *The Journal of Experimental Education*, 74(3), 251-266.

Gulliksen, H. (1950). *Theory of menatl tests*. New York: John Wiley & Sons.

Hambleton, R. & Swaminatham, H. (1984). *Item Response Theory, Principles and Applications (Evaluation in Education and Human Services)*. New York: Springer.

John, O. P. & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation and scale construction. In C. M. Judd (Ed.), *Handbook of research methods in social and personality psychology*. (pp. 339-369).

Linacre, J. M. & Wright, B. D. (1991). *WINSTEPS - Rasch-Model computer programs*. Chicago: MESA Press.

Low, G. D. (1988). The semantics of questionnaire rating scales. *Evaluation and Research in Education*, 2 (2), 69-70.

McCrae, R. R. & John, O. P. (1992). An introduction to the Five-Factor Model and its applications. *Journal of Personality*, 60, 175-216.

Muñiz, J. (1990). *Theoría de respuesta a los items: un nuevo enfoque en la evolución psicológica y educativa*. Madrid: Ediciones Pirámide.

Muñiz, J. (1996). *Psicometría*. Madrid: Editorial Universitas.

Nunes, C. H. S. S. & Hutz, C. S. (2007). *Escala Fatorial de Socialização: Manual Técnico*. São Paulo: Casa do Psicólogo.

Nunes, C. H. S. S., Nunes, M. F. O., & Hutz, C. S. (2006). Uso Conjunto de Escalas de Personalidade e Entrevista Para Identificação de Indicadores de Transtorno Anti-social (no prelo). *Avaliação Psicológica*, 5 (2).

Pasquali, L. (1999). Testes Referentes a Construto: Teoria e Modelo de Construção. In L. Pasquali (Ed.), *Instrumentos Psicológicos: Manual Prático de Elaboração*. (pp. 37-71). Brasília, DF: Laboratório De Pesquisa em Avaliação e Medida – LabPAM.

Pasquali, L. (2003). *Psicometria: Teoria dos testes na Psicologia e na Educação*. Petrópolis, RJ: Vozes.

Primi, R. (1996). *Construção de um instrumento para a avaliação do raciocínio indutivo: aplicação da psicologia cognitiva e da teoria de resposta ao item*. Unpublished Prometo de Qualificação, Universidade de São Paulo, São Paulo.

Roberts, J. (1994). Rationg scale functioning. Retrieved 29/10/2006, http://www.rasch.org/rmt/rmt83r.htm

Stone, M. H. & Wright, B. D. (1994). Maximizing rating scale information. *Rasch Measurement Transactions*. Retrieved 29/10/2006, 2006, from http://rasch.org/rmt/rmt83r.htm

Weems, G. H. (2004). Impact of the Number of Response Categories on Frequency Scales . Research in the Schools, 11, 41-49.

Weng, L. (2004). Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability . Educational and Psychological Measurement, 64, 956-972.

Widiger, T. A., Trull, T. J., Clarkin, J. F., Sanderson, C., & Costa, P. T. (2002). A description of the DSM-IV personality disorders with the five-factor model of personality. In P. T. Costa & T. A. Widiger (Eds.), *Personality Disorders and the Five-Factor Model of Personality*. (2 ed., pp. 89-102). Washington, DC: American Psychological Association.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA.

Wright, B. D. & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA.