UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Métodos de verificação das suposições e da qualidade de ajuste dos modelos TRI cumulativos unidimensionais

Autor: Tiago Henrique Lenhard

Orientador: Professora Drª. Stela Maris de Jezus Castro

Porto Alegre, 13 de Dezembro de 2013.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Métodos de verificação das suposições e da qualidade de ajuste dos modelos TRI cumulativos unidimensionais

Autor: Tiago Henrique Lenhard

Monografia apresentada para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Drª. Stela Maris de Jezus Castro
Nome do convidado: Drª. Suzi Alves Camey

Porto Alegre, 13 de Dezembro de 2013.

Dedico este trabalho a minha família, minha base e inspiração.

Agradecimentos

Agradeço minha professora orientadora Stela Maris de Jezus Castro pelo apoio e escolha do tema deste trabalho e por me fazer conhecer um pouco do mundo da teoria da resposta ao item.

Agradeço também a professora Jandyra Fachel, pelos conselhos e ensinamentos, tanto profissionais quanto pessoais que me fora passado, muito do profissional que sou, devo a ela. A todos os professores e profissionais do Núcleo de Assessoria Estatística da universidade pelos apoio dado durante a graduação.

Agradeço aos meus colegas e amigos pelo apoio e amizade dados durante esses anos de graduação. Especialmente agradeço a minha colega Natalia Giordani, pela bela bonita amizade criada desde os tempos de NAE. A Bárbara Pederiva, minha melhor amiga, por ter trilhado esse caminho junto comigo, apoiando nos momentos difíceis e comemorando os felizes. Jamais será esquecida. Ao Douglas Mesquita, meu colega, meu amigo e meu irmão, teu apoio foi essencial para o meu crescimento e amadurecimento como pessoa. Que a nossa sociedade dure sempre, não importando a distância que nos encontramos.

Finalmente, quero agradecer aos meus pais, pelo esforço de ambos para dar condições de realizar a minha graduação, pela compreensão nas minhas escolhas, pelo amor incondicional dado e pelos ensinamentos que sempre me passaram. Agradeço também a minha irmã emprestada Geila, pelos conselhos e apoios nos mais diversos momentos. Obrigado por tudo.

Resumo

Na Teoria de Resposta ao Item (TRI), como em outros modelos, existem suposições que devem ser atendidas para sua adequada implementação e testes pra verificar a qualidade do ajuste dos mesmos. Nesse trabalho, foram revisados alguns dos métodos e técnicas que verificam essas suposições e ajustes. O trabalho está estruturado em 3 grandes tópicos: inicialmente são apresentados métodos para verificar a suposição de unidimensionalidade; em seguida, são apresentados métodos que verificam a qualidade do ajuste do modelo; por fim, os métodos implementados em softwares livres são ilustrados através da utilização de exemplos.

Palavras-chave: Teoria da Resposta ao Item, Unidimensionalidade, Qualidade de ajuste do modelo, Aplicação em softwares livres.

Abstract

In Item Response Theory (IRT), as in other models, there are assumptions that must be met for its adequate implementation and adequate testing to ensure their adjusting quality. In this paper, some of the methods and techniques that investigate these assumptions and adjustments are revised. The work is structured in 3 major topics: initially, methods are presented to verify the assumption of unidimensionality; then methods that verify the goodness of fit of data model are presented; and, finally, implemented methods in free softwares are illustrated through examples.

Keywords: Item Response Theory, Unidimensionality, Goodness of fit of data model, Application in free softwares.

Sumário

1. Introdução

A Teoria da Resposta ao Item é um conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros do item e da habilidade do respondente, a qual denominamos traço latente. (Andrade, Tavares e Valle (2000)).

Na Teoria de Resposta ao Item (TRI) há, como em outros modelos, suposições que devem ser atendidas para que seu uso seja adequado e testes pra verificar a qualidade do ajuste do modelo selecionado. Como esta é uma área em desenvolvimento, diversas metodologias têm sido propostas para estes fins, entretanto, os pesquisadores têm encontrado dificuldades para lidar com a subjetividade na escolha e interpretação dos resultados destes métodos.

Em vista disso, este trabalho tem como objetivo reunir algumas técnicas que são úteis para a verificação das suposições de uma subclasse dos modelos TRI: os modelos cumulativos unidimensionais, e, também, apresentar métodos estatísticos que visam avaliar a qualidade do ajuste dos mesmos.

Inicialmente foram consideradas as suposições de unidimensionalidade e de independência local. Segundo Hambleton & Swaminathan (1991) (em Andrade, Tavares e Valle (2000)), a unidimensionalidade implica em independência local, sendo assim, basta verificar apenas uma destas suposições. Existem diversos métodos na literatura que se prestam para verificar a suposição de unidimensionalidade (McDonald (1981), Pasquali, (2003), Stout (1987), Bock e Aitkin (1981) e Horn (1965)). Neste trabalho, foram mostradas a análise fatorial, análise paralela, e um teste não paramétrico chamado teste de Stout.

O segundo tópico, e mais explorado por este trabalho, refere-se a métodos que avaliam a qualidade de ajuste dos modelos TRI, que podem ser vistos nos trabalhos de Bock (1972), Yen (1981), Mckinley & Mils (1985), Orlando & Thissen (2000), Mair, Reise & Bentler (2008), Stone, Mislevy & Mazzeo (1994), Liang (2010) e Yin (2007).

Uma vez que as suposições do modelo estão atendidas e que o modelo TRI é ajustado (etapa conhecida como calibração dos itens), é necessário verificar a qualidade do ajuste do mesmo. No contexto TRI não há um único método ou teste específico que é utilizado pela grande maioria dos pesquisadores desta área. Neste artigo foram reunidos os métodos mais

utilizados e, também, novas abordagens, segmentando-os em dois grandes grupos: métodos clássicos, compostos pelo teste Qui-Quadrado de Bock, estatística Q_1 de Yen e a estatística G^2 proposta por Mckinley & Mils; e métodos alternativos, dos quais foram explorados o teste baseado no escore total de Orlando & Thissen, usando os testes da regressão logística (*Collapsed Deviance*, *Hosmer-Lemeshow Test*, *Casewise Deviance* e *Rost's Deviance*) vistos no trabalho de Mair, Reise & Bentler (2009), usando a estatística do esperado à *Posteriori* proposta por Stone, Mislevy, & Mazzeo (1994), o teste não paramétrico *RISE* e uma abordagem usando os resíduos no software *ResidPlots-2* vistos no trabalho de Liang (2009).

Por último, foram citados os *softwares* que contém as técnicas apresentadas, e, no caso de *software* do tipo *freeware*, foram citados os procedimentos e comandos para execução do devido teste a fim de auxiliar pesquisadores dessa área.

2. Unidimensionalidade

Citando Cuesta (1996):

> *Não existe uma definição de unidimensionalidade comum a todos os autores. A dificuldade de obtenção de uma definição universal prende-se com o fato de a maioria definir unidimensionalidade em função da forma como a vão avaliar*.

2.1. Análise fatorial

Por se tratar de uma técnica exploratória, o resultado deste método é subjetivo, mas se um instrumento é unidimensional espera-se que um único fator sobressaia, ou seja, concentre grande parte da variação total do conjunto de itens que o compõem.

McDonald (1981) define que um conjunto de medidas é unidimensional se, e apenas se, a sua variação é explicada por um *common fator model*, ou seja, se um único fator explica a maior quantidade de variação possível. Esse conceito é estendido para a definição de unidimensionalidade suficiente, que considera que um conjunto de itens apresentará unidimensionalidade suficiente, de modo que seja pertinente a utilização dos modelos TRI, se o primeiro fator for preponderante ante os demais. Os critérios para considerar a unidimensionalidade de um conjunto de itens são bastante subjetivos. A fim de criar um critério

objetivo, Reckase (1979) (em Vitoria, Almeida & Primi (2006)) propôs que um conjunto de itens pode ser considerado com unidimensionalidade suficiente se o percentual da variância explicada pelo primeiro fator é no mínimo de 20%.

2.2. Análise paralela

Na análise paralela, Reise, Comrey & Waller (1999) descrevem que amostras aleatórias são simuladas reproduzindo o mesmo número de itens e sujeitos que existem na matriz de dados reais. Então pode ser aplicada a análise fatorial, ou a análise de componentes principais, nos dois conjuntos de dados (reais e simulados). Compara-se os autovalores dos fatores gerados através do gráfico *screeplot*, contrastando os dois modelos. Dessa forma, por inspeção visual, o pesquisador tem uma ideia de quantos fatores são extraídos. O mais interessante desta técnica é o fato de ser possível visualizar o nº de fatores nos quais os autovalores da matriz de dados reais são maiores do que os autovalores do grupo de dados simulados, indicando assim o nº de fatores dominantes.

2.3. Análise não paramétrica de dimensionalidade de Stout

Stout (1987) definiu a unidimensionalidade essencial, que se refere à existência de uma dimensão dominante. Esse método mostrou ter um baixo erro tipo I e um bom poder estatístico (menos para amostras pequenas ou pequeno número de itens) (Childs & Oppler(2000)). Essa técnica testa a hipótese de que $d = 1$, onde d representa o número de dimensões em um conjunto de itens do teste. Uma das restrições desse teste é apenas analisar itens com respostas dicotômicas.

3. Qualidade do ajuste

Uma vez que um modelo TRI está definido e calibrado, é importante verificar a qualidade de ajuste do mesmo. Para isso, foram criados vários testes que têm essa finalidade, porém ainda não existe unanimidade sobre este assunto entre os pesquisadores dessa área. Na revisão de

literatura foi observado que os testes clássicos conseguem medir esse ajuste, porém só sendo válidos se forem supridas uma série de restrições. A seguir são apresentados os testes clássicos.

3.1. Testes clássicos de ajustamento

O método mais tradicional utilizado para verificar a qualidade do ajuste dos modelos TRI é o teste Qui-Quadrado nas suas variações (Bock (1972), Yen (1981) e Mckinley & Mils (1985)). A estatística é obtida comparando as respostas do modelo real com as do esperado, dado que as suposições do modelo TRI foram atendidas. Basicamente, a ideia do teste é montar uma tabela de contingência, onde cada linha representa a habilidade (ou traço latente (θ)) que é dividida em subgrupos e nas colunas são postas as categorias de resposta dos itens, conforme mostra a tabela 1.

	Item *i*		
Habilidade (θ)	**0**	**1**	**Total**
(-5) - (-3)	F_{11}	F_{12}	$F_{.1}$
(-3) - (-1)	F_{21}	F_{22}	$F_{.2}$
(-1) - 1	F_{31}	F_{32}	$F_{.3}$
1 - 3	F_{41}	F_{42}	$F_{.4}$
3 - 5	F_{51}	F_{52}	$F_{.5}$
Total	$F_{.1}$	$F_{.2}$	N

F_{ij} é a frequência observada na linha *i* e coluna j

Tabela 1: exemplo de construção de tabela de contingência para um item *i*.

3.1.1. Teste Qui-Quadrado de Bock:

Segue a estatística de teste do Qui-Quadrado de Pearson proposta por Bock (1972):

$$\chi_i^2 = \sum_{k=1}^{K}\sum_{j=1}^{J}\frac{n_k(O_{kj}-E_{kj})^2}{E_{kj}} \sim \chi^2_{K*(J-1)-m}$$

onde,

k é um subgrupo da habilidade,

K é o total de subgrupos da habilidade,

j é a categoria respondida pelo individuo no item *i*,

J é a maior categoria de resposta do item *i*,

n_k é a frequência observada no subgrupo da habilidade k,

O_{kj} é a resposta observada j com subgrupo de habilidade k,

E_{kj} é a resposta esperada j com subgrupo de habilidade k,

m é quantidade de parâmetros estimados.

O valor observado é obtido através da frequência de respostas em cada intervalo de habilidade. O valor esperado é estimado utilizando a estimativa dos parâmetros dos itens e a mediana estimada em cada intervalo de habilidade. Uma das restrições desse método é apenas avaliar modelos dicotômicos.

3.1.2. Teste Qui-Quadrado de Yen

Uma modificação do teste Qui-Quadrado foi proposta por Yen (1981). A estatística Q_1 proposta neste método é parecida com a de Bock, avalia também só modelos dicotômicos, mas tem duas restrições. Uma delas é que a divisão das habilidades será em 10 subgrupos com amplitudes iguais; a segunda restrição define que a E_{kj} é a média da probabilidade da resposta j de pessoas com a habilidade no subgrupo k. A estatística Q_1 tem distribuição aproximada χ^2 com $10 - m$ graus de liberdade.

3.1.3. Estatística G²

Baseado na estatística Q_1, Mckinley & Mills (1985) formalizaram a estatística G^2, que segue aproximadamente uma distribuição χ^2. Segue a estatística de teste do G²:

$$G_i{}^2 = 2\sum_{k=1}^{K}\sum_{j=1}^{J} O_{kj}\log\left[\frac{O_{kj}}{E_{kj}}\right] \sim \chi^2{}_{10*(J-1)-m}$$

Onde,

k é um subgrupo da habilidade,

K é o total de subgrupos da habilidade,

j é a categoria respondida pelo individuo no item i,

J é a maior categoria de resposta do item i,

O_{kj} é a proporção observada de respostas j com subgrupo de habilidade k,

E_{kj} é a proporção esperada de respostas j com subgrupo de habilidade k,

m é quantidade de parâmetros estimados.

É o primeiro teste que pode ser utilizado tanto para modelos dicotômicos como para modelos politômicos.

3.1.4. Considerações sobre os testes clássicos

Os testes clássicos tem muitas limitações. Uma delas é o tamanho de amostra, pois para ser assintoticamente distribuído com distribuição Qui-Quadrado, um tamanho de amostra grande é necessário. No entanto, se for grande demais, corre-se o risco de rejeitar-se a hipótese nula, que postula que o item está bem ajustado, com pequenas diferenças que, na prática, não interfeririam no ajuste do mesmo.

Outro problema é a falta de uma distribuição uniforme das categorias de resposta nas classes de habilidades. Ela também é causada pelo tamanho de amostra. Esse problema ocorre quando o tamanho de amostra é pequeno ou quando há uma amostra com itens com muitas categorias de resposta.

Da mesma forma, a subjetividade nas divisões dos subgrupos da habilidade também pode interferir no resultado do teste.

Para mais detalhes e comparações desses métodos, levando-se em conta a distribuição, o erro tipo I, o poder de teste, entre outros desempenhos, como também métodos para corrigir a dispersão das respostas, ver Yin (2007) e Stone & Zhang (2003)

3.2. Teste alternativos para obter o a qualidade do ajuste

Os testes a seguir apresentados foram criados na tentativa de resolver ou contornar os problemas gerados pelas limitações dos testes clássicos.

3.2.1. Estatística de ajuste usando o *esperado a posteriori*

Foi desenvolvida por Stone, Mislevy & Mazzeo (1994), e foi proposta para corrigir ou amenizar o problema causado pela incerteza na estimação do parâmetro de habilidade (θ). Isso ocorre quando há instrumentos com poucos itens. Bastante similar aos métodos clássicos, esse método calcula também o χ^2e o G^2. A diferença está na forma de construir a tabela de contingência. Aqui ela é construída por uma abordagem que representa a incerteza na estimativa de habilidade e permite que o ajuste do item seja feito quando as estimativas não são precisas. Uma das desvantagens desse método, como nos clássicos, é a falta de uma distribuição uniforme das respostas. Uma correção do método foi feita por Stone (2000). Mais comparações de desempenho desse método com outros testes podem ser vistas em Stone & Zhang (2003) e Yin (2007).

3.2.2. Estatística condicionada no escore total do teste

Orlando & Thissen (2000) dividiram os indivíduos não mais em subgrupos de habilidade, mas sim no escore total de cada um. Foram desenvolvidas duas novas estatísticas, a *S*-χ^2e a *S*-G^2, baseadas nos métodos clássicos.

O valor observado para essas estatísticas é a proporção observada de respostas corretas no item *i* no subgrupo de escore *k*, e o valor esperado é a proporção esperada de respostas corretas no item *i* no subgrupo de escore *k*. As notações e interpretações são as mesmas dos métodos clássicos. Essas duas estatísticas também seguem aproximadamente uma distribuição Qui-Quadrado.

Uma comparação de desempenho das estatísticas *S*-χ^2e *S*-G^2 foi feita por Orlando & Thissen demonstrando que a estatística *S*-χ^2tem um desempenho superior em relação aos testes clássicos na detecção da falta de ajuste. A generalização para os casos de respostas politômicas pode ser vista em Kang e Chen (2007).

A grande vantagem desse método, em relação aos demais, é o fato dele não utilizar as estimativas da habilidade para obter os valores observados.

3.2.3. Abordagem da regressão logística

Mair, Reise & Bentler (2008) testaram uma abordagem da regressão logística para respostas dicotômicas nos modelos TRI.

Para testar as diferentes estatísticas de ajuste, os autores utilizaram o modelo de Rasch de resposta dicotômica. Foram testados e comparados, quanto ao poder do teste e Erro Tipo I, os testes, *Collapsed Deviance*, *Hosmer-Lemeshow Test*, *Casewise Deviance* e *Rost's Deviance*.

Os Resultados das simulações demonstraram que o teste de *Collapsed Deviance* teve um resultado muito bom (exceto para tamanhos de amostra pequenos) não é limitado como os demais podendo ser aplicado em outros modelos da TRI, além do de Rasch. Já o *Hosmer-Lemeshow Test* somente obteve resultados satisfatórios para amostras grandes conjuntamente com muitos itens. Os demais testes não obtiveram bons resultados, pois inflacionavam o erro tipo I e apresentaram baixo poder na maioria das simulações.

3.2.4. Análise gráfica dos resíduos

Para Liang, Han & Hambleton (2009), os testes de significância estatística para a verificação da qualidade do ajuste não estão isentos de deficiências. Tais métodos tendem a ser concentrados num aspecto particular da relação entre o modelo e os dados, frequentemente resumindo a avaliação em um único número descritivo ou em um resultado de teste.

A abordagem tradicional da análise residual consiste em estimar a habilidade, predizendo a performance dos vários subgrupos de habilidade com base no modelo TRI estimado. Esta abordagem compara a distribuição esperada com a distribuição observada usando os resíduos brutos ou os resíduos padronizados (Yin (2007)), mostrando, por exemplo, o comportamento dos resíduos com as suas respectivas barras de erro (o que dá uma ideia do quanto eles são heterogêneos numa classe de habilidade) fornecendo uma inspeção visual útil do ajuste do modelo.

3.2.5. Teste não paramétrico da Raíz da Integral do Erro Quadrático

O teste da Raiz da Integral do Erro Quadrático (do termo em inglês *root integrated squared error* (RISE)) foi proposta por Douglas & Cohen (2001) (em Liang (2010). O conceito desta abordagem baseia-se em construir a Curva Característica do Item (CCI), que é uma representação gráfica da probabilidade de acerto do item i para uma determinado grau de habilidade. A CCI é estimada por um Kernel suavizado que utiliza a técnica de Média Local (*Local Averaging*). Esse teste compara essa CCI não paramétrica com a CCI do modelo paramétrico ajustado. Se as duas CCI's forem muito diferentes, o modelo paramétrico é considerado mal ajustado.

A significância do teste RISE é determinada por um procedimento de reamostragem criando uma distribuição empírica do mesmo. Se os valores observados do RISE forem maiores do que o percentil 95 da distribuição empírica, o ajuste do modelo é considerado inadequado.

Uma das vantagens deste método é poder representar graficamente essas duas curvas, podendo assim verificar visualmente o ajuste do modelo. Liang (2010) descreve mais detalhadamente sobre a técnica e faz uma comparação de desempenho com outros dois métodos, o G^2 e o S-χ^2, sendo o RISE superior em todas as simulações testadas, mantendo sempre um Erro tipo I baixo e um bom poder, considerando pequenas e médias amostras.

4. Uso dos Softwares

4.1. Unidimensionalidade

Para exemplificar as técnicas análise fatorial e análise pararela, foi utilizado o banco *Asthma34.csv* composto por 34 itens politômicos que são relacionados a asma, o qual foi aplicado em 622 crianças com idades entre 8 e 12 anos. Este banco de dados foi retirado das bases de exemplo do software IRTPRO Student.

4.1.1. Análise fatorial

Para executar a análise fatorial e a análise paralela pode ser utilizado o software R, versão 3.0.1 64 bits (disponível em http://www.r-project.org/). Para tanto, faz-se necessário instalar e carregar o pacote *psych.* Usando a função:

```
fa(x, nfactors=, ...)
```

onde *x* é a matriz dos dados e *nfactors* é o números de fatores que se deseja extrair. Por padrão da função é considerado número de fatores igual a 1 e a rotação do modelo é a *Varimax*. A hipótese nula testada postula que o conjunto de itens é suficientemente unidimensional, ou seja, existe um fator que consegue explicar pelo menos 20 % da variância total. Nos resultados é mostrada a matriz de correlação utilizada, e também algumas características do modelo. Os resultados mais importante estão destacados em vermelho.

```
                MR1
SS loadings    6.64
Proportion Var 0.20

Test of the hypothesis that 1 factor is sufficient.

The degrees of freedom for the null model are  561  and the objective function was  7.9 with Chi
The degrees of freedom for the model are 527  and the objective function was  2.64

The root mean square of the residuals (RMSR) is  0.06
The df corrected root mean square of the residuals is  0.09

The harmonic number of observations is  622 with the empirical chi square  2421.22  with prob <
The total number of observations was  622  with MLE Chi Square =  1605.22  with prob <  2.6e-109

Tucker Lewis Index of factoring reliability =  0.73
RMSEA index =  0.058  and the 90 % confidence intervals are  0.054 0.061
BIC =  -1784.94
Fit based upon off diagonal values = 0.92
Measures of factor score adequacy
                                                MR1
Correlation of scores with factors             0.95
Multiple R square of scores with factors       0.90
Minimum correlation of possible factor scores  0.79
```

Figura 1 - Saída da função *fa* suprimindo a matriz de correlação

Interpretação de resultados: Os dados evidenciam que o conjunto de 34 itens é suficientemente unidimensional com p-valor <0,0001. O primeiro fator resultante da análise explica 20% da variação total dos dados.

.

4.1.2. Análise paralela

Na análise paralela, a função do R é:

```
fa.parallel(x, n.obs=,...)
```

onde *x* é a matriz dos dados e *n.obs* é o tamanho de amostra que será utilizado para construir a distribuição simulada.. O resultado é apresentado através do *screeplot* (Figura 2).

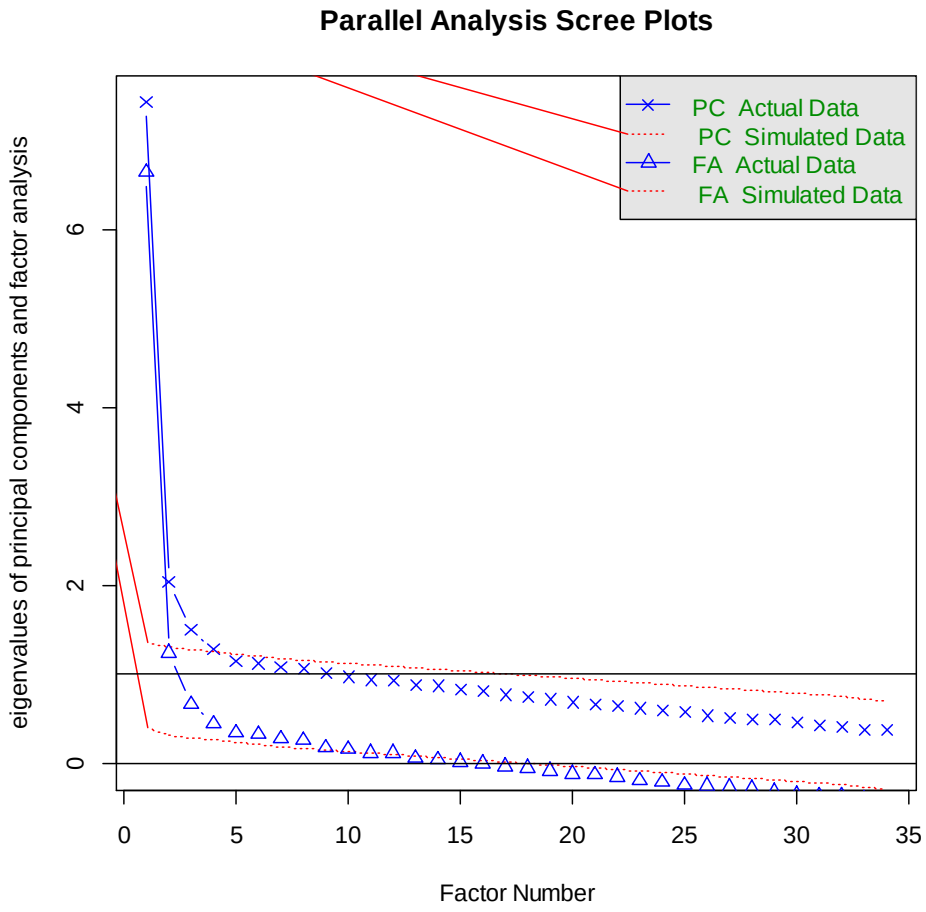

Figura 2 – *Screeplot* resultante da análise paralela

Interpretação de resultados: na Figura 2 pode-se verificar, tanto pela análise fatorial como pelos componentes principais, que há um fator preponderante (linhas azuis contínuas) em relação aos fatores gerados pela simulação (linhas tracejadas vermelhas), confirmando o resultado que já havia sido obtido no método anterior.

4.1.3. Teste de Stout

Para realizar o teste de Stout, foi idealizado o DIMTEST (Stout (1987)), disponível no software DIMPACK V1.0 (Stout (2006)). É testada a hipótese nula de que o modelo é unidimensional ($d = 1$). No exemplo será usado o banco lsat6.csv, que foi gerado por Thissen em 1982, no teste de Admissão da Escola de Direito, é composta por 5 itens dicotômicos e por 1000 observações. Para carregar os dados, o formato do arquivo deve estar em formato *.dat*, separado por espaços simples e sem cabeçalhos. Uma vez feito isso, carrega-se o banco. A tela inicial do software aparecerá novamente. O próximo passo, é selecionar alguns itens para a *AT List*: Para isso deve selecionar em *Assessment Subtest* a opção *User Specified* e selecionar alguns itens (em azul - Figura 3). Isso é necessário para o software conseguir cruzar os itens e calcular a estatística de teste. Em seguida, deve-se definir um endereço para salvar o arquivo do output (em vermelho - Figura 3). Por fim, clicar em *Run*.

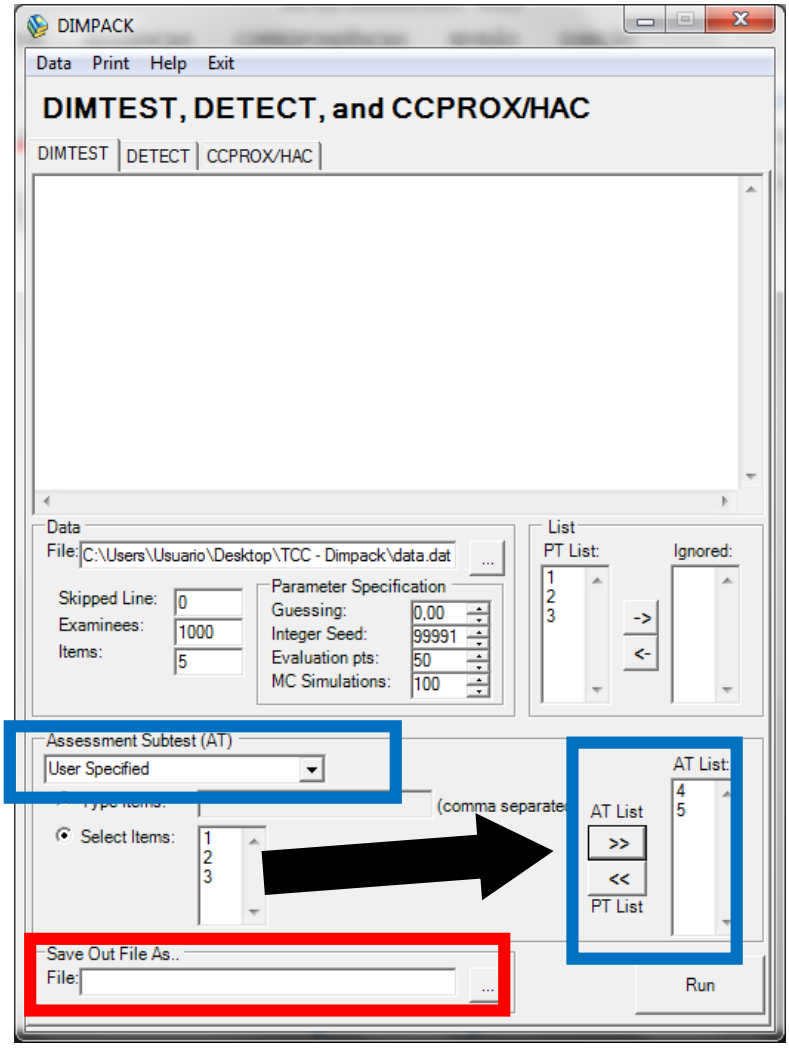

Figura 3 – Tela Principal do software Dimpack.

O resultado do teste é apresentado na tela principal do software:

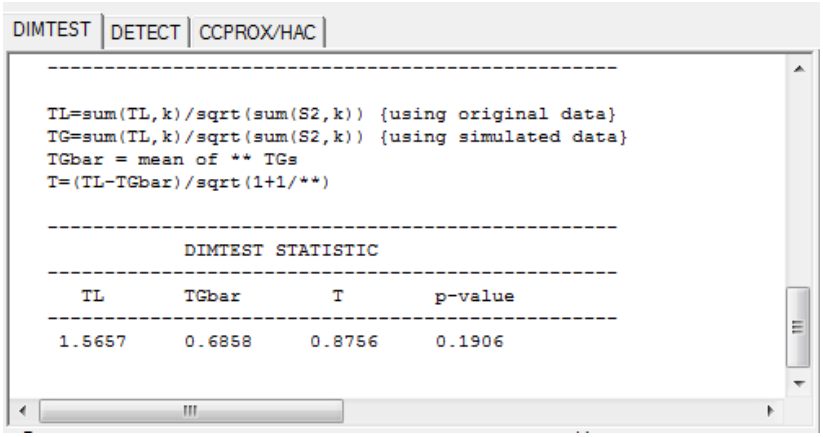

Figura 4 – Resultado da estatística DIMTEST.

Interpretação de resultados: há evidencia estatística de que esse conjunto de itens é unidimensional (p-valor = 0,1906).

4.2. Modelos clássicos de ajustamento

4.2.1. Teste Qui-Quadrado Bock e Yen

Para realizar esses dois testes de ajuste pode-se utilizar o software R (na mesma versão apresentada pelo outros testes). Para esses dois testes, os dados para o exemplo são os mesmos utilizados anteriormente no Teste de Stout (teste com cinco itens dicotômicos). A Hipótese nula para esses dois testes postula que os itens estão bem ajustados ao modelo TRI escolhido. O primeiro passo é instalar e carregar o pacote *ltm.* Em seguida, é necessário calibrar o modelo, isto é, ajustar o modelo da TRI desejado. Supondo que o modelo escolhido seja o modelo de Rasch, a função utilizado é:

```
modelo = rasch(dados, ...)
```

Após o ajuste do modelo, executa-se a função:

```
item.fit(modelo, G = 10, FUN = mean,...)
```

onde *G* é o número de subgrupos formados a partir da habilidade e *FUN* é o procedimento que será utilizado para encontrar valores esperados (pode ser mediana ou a média). O primeiro resultado mostrado, o teste de Yen, é apresentado na Figura 5.

```
> item.fit(modelo, G = 10, FUN = mean)

Item-Fit Statistics and P-values

Call:
rasch(data = dados)

Alternative: Items do not fit the model
Ability Categories: 10

        X^2 Pr(>X^2)
C1  60.2428  <0.0001
C2 158.1368  <0.0001
C3 233.3817  <0.0001
C4 131.3946  <0.0001
C5  83.2155  <0.0001
```

Figura 5 – Resultado do teste Qui-Quadrado de Yen.

Interpretação de resultados: percebe-se que nenhum item está bem ajustado ao modelo de Rasch (valor p < 0,0001 para os 5 itens).

Para executar o teste de Bock, pode-se mudar o número de subgrupos de habilidade, conforme conhecimento do pesquisador, e deve-se mudar o *FUN* para a mediana conforme a Figura 6:

```
> item.fit(modelo, G = 5, FUN = median)

Item-Fit Statistics and P-values

Call:
rasch(data = dados)

Alternative: Items do not fit the model
Ability Categories: 5

        X^2 Pr(>X^2)
C1  63.0849  <0.0001
C2 152.4665  <0.0001
C3 154.3890  <0.0001
C4 129.7745  <0.0001
C5  86.0359  <0.0001
```

Figura 6 – Resultado do teste Qui-Quadrado de Bock.

Interpretação de resultados: novamente nenhum item está bem ajustado ao modelo de Rasch (valor p < 0,0001 para os 5 itens).

4.2.2. Estatística G^2

A estatística G^2 não é encontrada em softwares gratuitos. Um dos softwares pagos que apresenta essa estatística é o IRTPRO (disponível em http://www.ssicentral.com). Não apresentamos uma aplicação dessa técnica, pois a única versão gratuita desse software, limitada na quantidade de testes, no número de casos e na quantidade de itens (aceita 1000 observações e 25 itens no máximo), é a versão STUDENT, a qual disponibiliza apenas a estatística condicionada no escore total do teste.

4.3. Modelos alternativos de ajustamento

4.3.1. Estatística de ajuste usando o *esperado a posteriori*

Não foi formalizada como um teste em nenhum software. O próprio autor do método comenta que foi desenvolvido no Software SAS® em forma de Macro (disponível em Hansen (2004)).

4.3.2. Estatística condicionada no escore total do teste

A estatística condicionada no escore total do teste é encontrada em diversos softwares pagos, tais como PARSCALE, BILOG-MG e IRTPRO (disponíveis em http://www.ssicentral.com). O software IRTPRO, disponível gratuitamente na versão STUDENT, será utilizado para demonstrar o método. O software trabalha com diversos formatos de arquivos de dados, como por exemplo, *.dat*, *.raw*, *.txt*, *.csv*, etc. Esse método é o padrão do software, mas pode ser acessado clicando em *Analysis*, em seguida em *Advanced Options*; na aba *Miscellaneous* marca-se a opção *Compute Chen-Thissen LD and item fit statistics* (Figura 7).

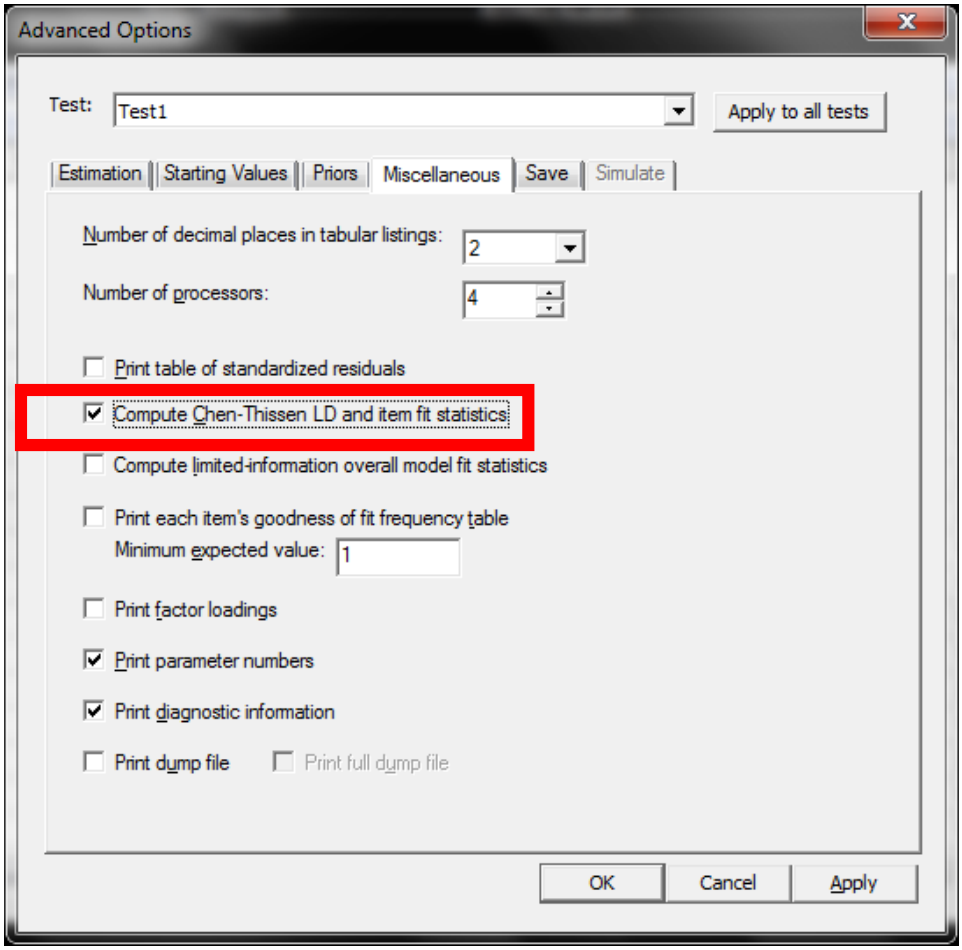

Figura 7 – Aba *Miscellaneous* do sub menu *Adavanced Options*.

Para exemplificar, foi utilizado o banco de dados *Anxiety14.csv*, que tem 14 itens, mas serão utilizados apenas 6 desses que tem relação com a ansiedade, foram observadas 518 pessoas, essa base foi retirada dos exemplos do próprio software. A hipótese nula para esse teste postula que os itens estão bens ajustados ao modelo TRI escolhido. O modelo TRI calibrado foi o modelo de resposta gradual. O software determina um método de estimação automaticamente pelo tipo de resposta, mas o usuário pode alterar para qualquer outro método todos os itens, ou alterar também somente alguns conforme a necessidade. O resultado obtido após calibrar o modelo é mostrado na Figura 8.

Summed-Score Based Item Diagnostic Tables and X^2s for Group 1 (Back to TOC)

S-X^2 Item Level Diagnostic Statistics

Item	Label	X^2	*d.f.*	Probability
1	Calm	42.47	38	0.2836
2	Tense	56.15	41	0.0576
3	Regretful	56.33	51	0.2818
4	AtEase	32.30	35	0.5999
5	Anxious	50.09	38	0.0905
6	Nervous	55.81	48	0.2044

Figura 8 — Saída de resultados do software IRTPRO .

Interpretação de resultados: adotando 5% de significância, conclui-se que os itens estão bem ajustados ao Modelo de Resposta Gradual.

4.3.3. Abordagem da regressão logística

Para realizar os testes de ajuste da regressão logística pode-se utilizar o R, através do pacote *eRm*. Primeiro, ajusta-se o modelo de Rasch usando o comando:

```
modelo = RM(dados)
```

Em seguida, executa-se a função:

```
pres= person.parameter(modelo)
```

que ajusta os parâmetros do indivíduo. Por fim, digitar os comandos:

```
GOF = gofIRT(pres,...)
```

```
summary(GOF)
```

No exemplo, foi utilizado o banco disponível no pacote *eRm*, o *raschdat1*, que é composto de 30 itens dicotômicos em 100 observações simuladas. A hipótese nula a ser testada postula que o modelo TRI de Rasch está bem ajustado. Os resultados estão apresentados na Figura 9.

```
Goodness-of-Fit Tests
                       value          df p-value
Collapsed Deviance   770.574         780   0.588
Hosmer-Lemeshow        6.937           8   0.543
Rost Deviance       2564.654 1073741794   1.000
Casewise Deviance   3221.328        2945   0.000

R-Squared Measures
Pearson R2: 0.275
Sum-of-Squares R2: 0.275
McFadden R2: 0.287
```

Figura 9 – Resultados dos métodos com a abordagem da regressão logística.

Interpretação de resultados: adotando um nível de significância de 5% os principais testes, o *Collapsed Deviance* e o *Hosmer-Lemeshow*, não rejeitaram a hipótese de que o modelo esteja bem ajustado (valores p 0,588 e 0,543, respectivamente).

4.3.4. Análise Gráfica dos resíduos

Esse método está disponível no software livre desenvolvido por Liang, Han & Hambleton (2009), o ResidPlots-2. O manual do software exemplifica todos os tipos de gráficos de resíduos disponíveis e faz comparações dos diferentes modelos de TRI. Porém ele trabalha somente com a sintaxe dos modelos já calibrados, sendo alienado aos softwares pagos PARSCALE e BILOG-MG. Para explicar, gráficos dos resíduos brutos com barras de erro são mostrado em uma comparação entre um modelo bem ajustado e outro com falta de ajuste nas figuras 10 e 11, respectivamente. Na primeira figura podemos ver que o praticamente nenhum resíduo está fora do intervalo, conferindo um bom ajuste do modelo as dados. Já na segunda imagem, poucos pontos estão dentro dos intervalos, mostrando que esse modelo não é o ideal para esse conjunto de dados.

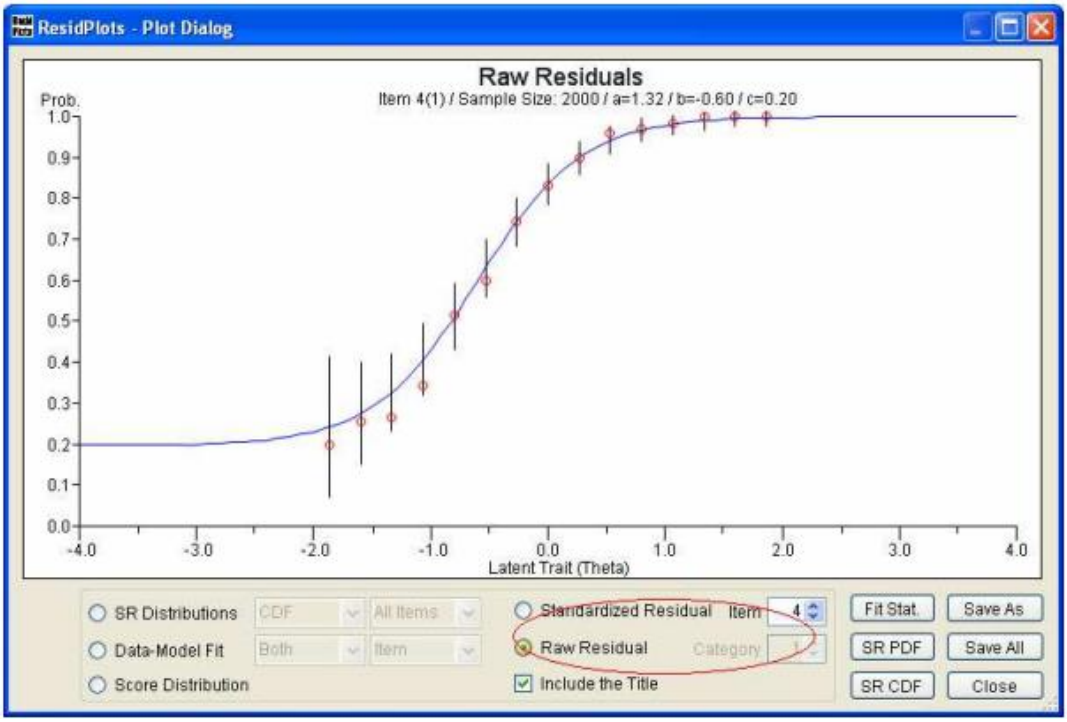

Figura 11 – Exemplo de modelo bem ajustados aos dados.

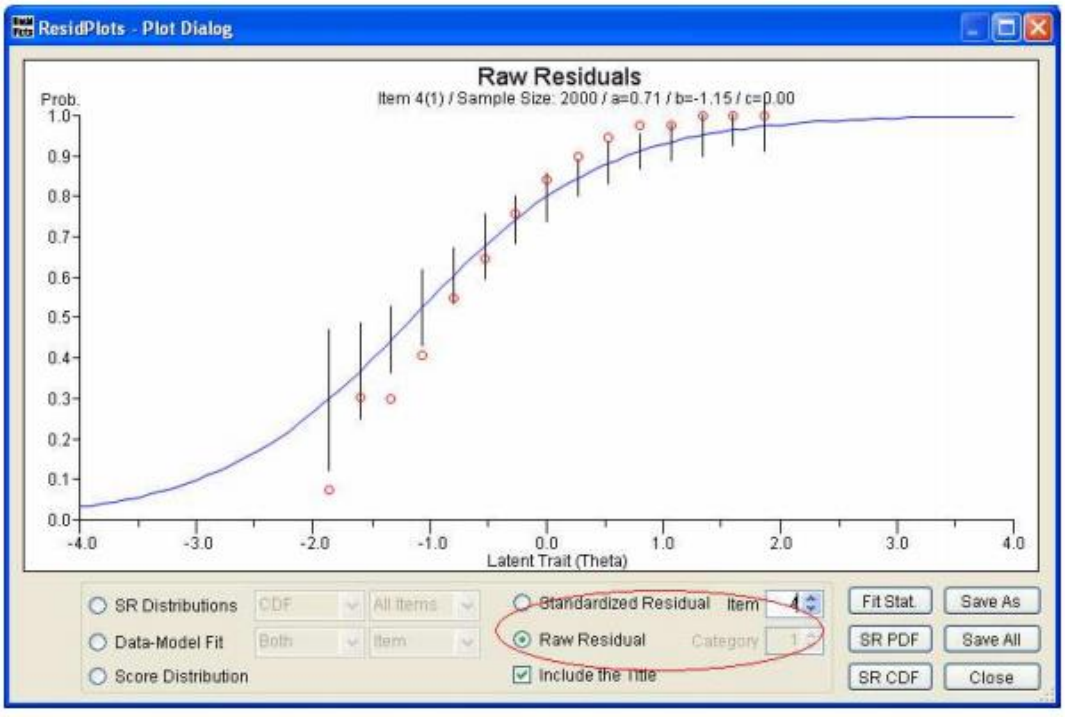

Figura 12 – Exemplo de modelo mal ajustados aos dados.

4.3.5. Teste não paramétrico da Raíz da Integral do Erro Quadrático

O teste não é encontrado em nenhum software comercial ou pacotes em softwares livres até o momento, sendo que a autora do artigo usado para compreender a técnica implementou-o no FORTRAN. Apenas a visualização gráfica que está disponível no ResidPlots-2.

5. Conclusão

Este trabalho permitiu conhecer algumas técnicas que são importantes na construção dos modelos TRI. Foram apresentadas as principais técnicas de verificação da unidimensionalidade, tanto a análise fatorial como a análise paralela, que podem ser usadas para itens dicotômicos e politômicos e estão disponíveis no software R. Já o Teste de Stout é mais limitado, estando disponível apenas para itens dicotômicos.

Em seguida foram abordadas as técnicas que verificam a qualidade do ajuste dos modelos TRI. Os primeiros testes que surgiram, os clássicos, tem muitas limitações, não sendo utilizados pela grande maioria dos pesquisadores. Para corrigir alguns desses problemas diversas abordagens alternativas foram criadas. Elas se mostraram superiores em alguns aspectos aos métodos clássicos. Deve-se ressaltar o teste RISE que obteve resultados satisfatórios em diversas situações simuladas e a análise gráfica dos resíduos, que, futuramente, poderão ser as principais técnicas para verificar a qualidade do ajuste. A desvantagem do teste RISE é não estar disponível em nenhum software livre, sendo necessário utilizar uma linguagem de programação para executá-lo. No caso dos gráficos dos resíduos, a implementação já foi feita, porém a execução está alienada a softwares pagos.

6. Referências Bibliográficas

ANDRADE, D. F., TAVARES, H. R., VALLE, R. C. (2000). Teoria da Resposta ao Item: Conceitos e Aplicações. SINAPE 2000.

BOCK, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37(1), 29-51.

BOCK, R. D. & AITKIN, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psycometrika, 46, 443-459.

CUESTA, M. (1996). Unidimensionalidade. InJ. Muñiz (Ed.), Psicometría. Madrid: Editorial Universitas.

HANSEN, M. A. (2004). Predicting the istribution of a goodness-of-fit statistic appropriate for use with performance-based assessments. Dissertation PhD., University of Pittsburgh.

HORN, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179-185.

KANG, T., CHEN, T. T. (2007). An Investigation of the Performance of the Generalized S-χ^2 Item-Fit Index for Polytomous IRT Models. ACT Research Report Series 2007-1.

LIANG, T. (2010). An assessment of the nonparametric approach for evaluating the fit of item response models. Dissertation PhD., University of Massachusetts - Amherst.

MAIR, P., REISE, S. P., BENTLER P. M. (2008). IRT Goodness-of-Fit Using Approaches from Logistic Regression. University of California.

MCDONALD, R. (1981). The dimensionality of testes and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

MCKINLEY, R., & MILLS, C. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 19, 49-57

ORLANDO M, THISSEN D (2000). Likelihood-based item-t indices for dichotomous item response theory models." Applied Psychological Measurement, 24, 50{64.

Liang, T., Han, K.T., & Hambleton, R. K. (2009). User's Guide for ResidPlots-2: Computer Software for IR T Graphical Residual Analyses (Version 2.1)

PRIMI, R.& ALMEIDA, L. S. (1998). Considerações sobre a análise factorial de itens com resposta dicotómica. Psicologia: Teoria, Investigação e Prática, 3, 225-23

REISE, S. P., COMREY, A. L., WALLER, N. G. (1999). Factor Analysis and Scale Revision. Psychological Assessment, 2000, Vol.12, No.3, 287-297.

SCIENTIFIC SOFTWARE INTERNATIONAL INC.: URL http://www.ssicentral.com

STONE, C.A., MISLEVY, R.J., MAZZEO, J. (1994). Classification error and goodness-of-fit in IRT models. Paper presented at the meeting of the American Educational Research Association, New Orleans. April.

STONE, C. A. (2000). Monte-Carlo based null distribution for an alternative fit statistic. Journal of Educational Measurement, 37, 58-75.

STONE, C.A. & ZHANG, B. (2003). Comparing three new approaches for assessing goodness-of-fit in IRT models. Journal of Educational Measurement, 4, 331-352.

STOUT, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

WIILIAN STOUT INSTITUTE OF MEASUREMENT (2006). Measured Progress ™. URL http://psychometrictools.measuredprogress.org

THE R PROJECT FOR STATISTICAL COMPUTING.: URL http://www.r-project.org/

VITORIA, F., ALMEIDA, L. S., PRIMI, R. (2006). Unidimensionalidade em testes psicológicos: conceito, estratégias e dificuldades na sua avaliação. PSIC- Revista de Psicologia da Vetor Editora, v. 7, nº 1, p. 1-7, Jan./Jun. 2006

YEN, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5(2), 245-262

YIN, Y. (2007). Using Beaton Fit Indices to Assess Goodness-of-fit of IRT Models. Dissertation PhD. University of Pittsburgh