UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

DISSERTAÇÃO DE MESTRADO

Equalização das escalas NESSCA e SARA utilizando a Teoria da Resposta ao Item na avaliação do comprometimento pela doença de Machado-Joseph

Nicole Machado Utpott

Orientador: Prof. Dra. Vanessa Bielefeldt Leotti

Porto Alegre, 14 de setembro de 2018.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

DISSERTAÇÃO DE MESTRADO

Equalização das escalas NESSCA e SARA utilizando a Teoria da Resposta ao Item na avaliação do comprometimento pela doença de Machado-Joseph

Nicole Machado Utpott

Orientador: Prof. Dra. Vanessa Bielefeldt Leotti

A apresentação desta dissertação é exigência do Programa de Pós-graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul, para obtenção do título de Mestre.

Porto Alegre, Brasil.
2018

BANCA EXAMINADORA

Prof. Dra. Mônica Maria Celestina de Oliveira, Departamento de Saúde Coletiva, UFCSPA.

Prof. Dr. Jonas Alex Morales Saute, Programa de Pós-graduação em Ciências Médicas, UFRGS.

Prof. Dra. Stela Maris de Jezus Castro, Programa de Pós-graduação em Epidemiologia, UFRGS.

AGRADECIMENTOS

À Prof. Vanessa, pela paciência e por ter desempenhado tão bem este papel de me orientar do outro lado do oceano.

Às amigas do coração.

Ao Rodrigo, que faz com que a vida seja mais leve e doce.

Ao meu pai, Gilberto, à minha mãe, Stela, e ao meu irmão, Gustavo, pelo apoio em minhas decisões e por torcerem pelo meu sucesso.

SUMÁRIO

ABREVIATURAS E SIGLAS

1PL	*One-parameter Logistic Model*
2PL	*Two-parameter Logistic Model*
3PL	*Three-parameter Logistic Model*
CCI	Curva Característica do Item
CCT	Curva Característica do Teste
CCR	Curva de Categoria de Resposta
CII	Curva de Informação do Item
CINEG	*Common Item Nonequivalent Groups*
GPCM	*Generalized Partial Credit Model*
GRM	*Graded-Response Model*
HCPA	Hospital de Clínicas de Porto Alegre
ICARS	*International Cooperative Ataxia Rating Scale*
MJD	*Machado-Joseph disease*
NESSCA	*Neurological Examination Score for Spinocerebellar Ataxia*
NRM	*Nominal Response Model*
OSE	*Observed Score Equating*
RG	*Random Groups*
SARA	*Scale for Assessment and Rating of Ataxia*
SCA	*Spinocerebellar Ataxia*
SG	*Single Group*
SG-C	*Single Group with Counterbalancing*
TCT	Teoria Clássica dos Testes
TRI	Teoria da Resposta ao Item
TSE	*True Score Equating*

RESUMO

CONTEXTO: Escalas de medida são ferramentas integrantes da prática clínica e da avaliação do estado de saúde de pacientes. O estado de saúde é uma variável que pode ser classificada como traço latente e, que, apesar de não poder ser medido diretamente, pode ser inferido com base na observação de variáveis secundárias relacionadas à característica de interesse. Estatisticamente, traços latentes podem ser estimados através da Teoria da Resposta ao Item (TRI), que compreende um conjunto de modelos que conectam respostas a um traço latente. A doença de Machado-Joseph (SCA3/MJD) é uma patologia genética cujo comprometimento neurológico é avaliado através de escalas, como a NESSCA e a SARA. Com a utilização de diferentes escalas existe uma dificuldade em comparar resultados científicos. A equalização de escalas propõe estabelecer uma relação de equivalência entre instrumentos de medida distintos. **OBJETIVO:** Explorar o método de equalização de escalas e demonstrar sua aplicação através das escalas NESSCA e SARA, utilizando a abordagem da Teoria da Resposta ao Item (TRI) na avaliação do comprometimento pela SCA3/MJD. **MÉTODOS:** Os dados são de 227 pacientes do HCPA (Hospital de Clínicas de Porto Alegre) portadores da SCA3/MJD que possuem medidas completas para NESSCA e/ou SARA. O delineamento de equalização utilizado é o de grupos não equivalentes com itens comuns, com calibração separada. Os modelos TRI utilizados na estimação dos parâmetros foram o resposta gradual, para itens da NESSCA, e o crédito parcial generalizado, para SARA. Foi feita a transformação linear através dos métodos *Mean/Mean*, *Mean/Sigma*, *Haebara* e *Stoking-Lord* e foi aplicado o método da equalização do verdadeiro escore para obter uma relação estimada entre os escores das escalas. **RESULTADOS:** O escore NESSCA estimado pela equalização via escore SARA comparado com o escore NESSCA observado apresentou diferença mediana de 0,78 pontos pelo método *Mean/Sigma*. **CONCLUSÕES:** Com este estudo foi possível explorar a aplicabilidade da técnica de equalização via TRI no contexto da saúde e ilustrar sua utilização criando uma relação de equivalência entre os escores das escalas NESSCA e SARA.

Palavras-chave: Equalização de escalas, Teoria da Resposta ao Item, doença de Machado-Joseph.

ABSTRACT

CONTEXT: Measurement scales are guidelines for clinical practice and for the assessment of patient's health status. This status is a latent trait that, although it cannot be measure directly, it can be inferred based on the observation of variables related to the characteristic of interest. Statistically, latent traits can be estimated by Item Response Theory (IRT), which comprises a group of generalized linear models based on a sample of answers to scale measures. Machado-Joseph disease (SCA3/MJD) is a genetic pathology which neurological impairment is evaluated through scales, like NESSCA and SARA. Due to use of different scales, the difficulty arises in compare scientific results. Scale equating has the purpose to stablish an equivalence relationship between measuring instruments. **OBJECTIVE:** The aims of this work is to present and discuss the scale equating method under IRT approach, as well as build a relationship of equivalence between NESSCA and SARA scores. **METHODS:** Data came from 227 patients diagnosed with SCA3/MJD with valid measures for NESSCA and/or SARA. The equating design was CINEG (Common Item Nonequivalent Groups) with separate calibration. The IRT models used in the parameter estimation were GRM (Graded Response Model) for NESSCA and GPCM (Generalized Partial Credit Model) for SARA. Scale linking was calculated through Mean/Mean, Mean/Sigma, Haebara and Stocking-Lord methods, in order to obtain an estimated relationship between score scales. Data from patients evaluated for both scales, NESSCA and SARA, were used to evaluate results accuracy. **RESULTS:** Difference between NESSCA score estimated by SARA and observed NESSCA score has shown median of 0,78 points, by Mean/Sigma method - which presented best results between scale linking methods. **CONCLUSIONS:** This study extended the use of scale equating under IRT approach to health outcomes and established an equivalence relationship between NESSCA and SARA scores, making the comparison between patients and scientific results feasible.

Key Words: Scale equating; Item Response Theory; Machado-Joseph disease.

LISTA DE QUADROS

I – Revisão de Literatura

III – Anexos da dissertação

IV – Anexos do artigo

LISTA DE TABELAS

I – Revisão de Literatura

II – Artigo

IV – Anexos do artigo

LISTA DE FIGURAS

I – Revisão de Literatura

II – Artigo

III – Anexos da dissertação

IV – Anexos do artigo

1. APRESENTAÇÃO

Este trabalho consiste na dissertação de mestrado intitulada “Equalização das escalas NESSCA e SARA utilizando a Teoria da Resposta ao Item na avaliação do comprometimento pela doença de Machado-Joseph”, apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul, em 14 de setembro de 2018. O trabalho é apresentado em três partes, na ordem que segue:

1. Introdução, Revisão da Literatura e Objetivos;
2. Artigo;
3. Conclusões e Considerações Finais.

Documentos de apoio estão apresentados nos anexos.

2. INTRODUÇÃO

A equalização de escalas é uma técnica estatística utilizada para estabelecer relações de equivalência entre escalas e já bastante disseminada na avaliação educacional (de Andrade et al., 2000). Embora a área da saúde seja uma das que que mais desenvolvem e consomem testes e escalas de medida, a técnica de equalização ainda é pouco explorada (Chen et al., 2009; McHorney e Cohen, 2000). Com o crescente interesse no cuidado com a saúde, são desenvolvidos muitos instrumentos de medida, cada vez mais especializados e voltados para interesses específicos (Coluci et al., 2015). Com a utilização de diferentes escalas, existe a dificuldade em comparar os resultados publicados pela comunidade científica. Nesse sentido, a equalização torna possível a criação de uma relação entre diferentes escalas ou diferentes populações.

A ataxia espinocerebelar do tipo 3, conhecida como doença de Machado-Joseph (SCA3/MJD), é um exemplo de patologia cujo comprometimento neurológico dos pacientes é avaliado utilizando escalas de medida, como a ICARS (Trouillas et al., 1997), a SARA (Schmitz-Hübsch et al., 2006) e a NESSCA (Kieling et al., 2008). Embora a doença não tenha cura, a avaliação do comprometimento dos pacientes exerce grande influência nas decisões que se referem aos tratamentos para contornar os sintomas e melhorar a qualidade de vida dos pacientes. Nesse sentido, a utilização de escalas é essencial.

Tradicionalmente, as escalas de medidas são analisadas de acordo com os princípios da Teoria Clássica dos Testes (TCT), que se baseia na média ou na soma do escore obtido nos itens. Na ICARS, na SARA e na NESSCA utiliza-se a soma. A Teoria da Resposta ao Item (TRI) foi desenvolvida para suprir as principais limitações da TCT, como a ausência de discriminação entre itens, ou seja, respondentes com o mesmo escore total são considerados iguais mesmo que o conjunto de respostas tenha sido totalmente diferente. A TRI tem por objetivo descrever a associação entre a probabilidade de uma resposta a um item em particular e o nível de um respondente quanto a uma característica de interesse que não pode ser observada diretamente, conhecida por traço latente (Castro et al., 2010).

A NESSCA já foi avaliada pela perspectiva da TRI. Em 2013, Maciel (2013) avaliou a escala através do GRM (do inglês, *Graded-Response Model*) com dados de 106 pacientes, permitindo identificar os sintomas que ajudam a melhor explicar o comprometimento pela doença. Foram propostas alterações na escala, como a exclusão dos itens Câimbra e Vertigem e o agrupamento de categorias de resposta para alguns itens (Maciel, 2013). Não se tem conhecimento de estudos avaliando a ICARS ou a SARA sob a perspectiva da TRI.

Chen et al. (2009) utilizou a equalização de escalas com a abordagem da TRI em dados de dois estudos que avaliam a intensidade da dor em pacientes. Como resultado, foi estabelecida uma conexão entre as escalas que permitiu avaliar os pacientes em uma mesma métrica, independente de quais escalas os pacientes haviam sido submetidos.

Este trabalho apresenta uma revisão da TRI, trazendo os principais conceitos e modelos, bem como um aprofundamento nos diferentes métodos de equalização de escalas, explorando as possibilidades para execução da técnica e aspectos computacionais.

3. REVISÃO DE LITERATURA

3.1 DOENÇA DE MACHADO-JOSEPH

A doença de Machado-Joseph, também conhecida como ataxia espinocerebelar tipo 3 (SCA3/MJD), é um distúrbio neurodegenerativo autossômico dominante caracterizado por uma ataxia cerebelar que tem início, geralmente, em indivíduos que se encontram na fase adulta. A expressão ataxia espinocerebelar (em inglês, *spinocerebellar ataxia* - SCA) é utilizada para fazer referência às doenças genéticas do cerebelo, cujos sintomas são caracterizados pela falta de coordenação muscular durante movimentos voluntários. A primeira ataxia espinocerebelar, SCA1, foi identificada em 1974 – o número da ataxia refere-se à ordem em que foi catalogada (Jardim, 2000). Atualmente, existem mais de 40 diferentes tipos de mutações genéticas que deram origem a todas as SCAs (Bird, 1993).

A SCA3/MJD é causada por uma expansão no número de repetições do trinucleotídeo CAG na matriz de leitura do *gene* ATXN3, sendo que o diagnóstico se dá quando o alelo alterado apresenta 51 repetições ou mais (Saute e Jardim, 2015). Os primeiros casos foram descritos nos Estados Unidos na década de 1970 em famílias de origem luso-açorianas – hoje Portugal tem a maior concentração de indivíduos afetados pela doença no mundo, a proporção de portadores da SCA3/MJD chega a atingir 418/100.000 no arquipélago das Flores, território português (Bettencourt et al., 2008). A doença de Machado-Joseph é o tipo de ataxia espinocerebelar mais comum no mundo, representando aproximadamente 36% dos casos de SCAs. Em locais com descendência açoriana esse percentual é ainda mais elevado – no estado do Rio Grande do Sul, a SCA3 é responsável por 78% das manifestações de SCAs (Donis et al., 2016). Estima-se uma prevalência mínima de 6/100.000 na região sul do Brasil (Souza et al., 2016).

Dentre os principais sintomas está a crescente perda do controle muscular e da coordenação motora nos membros superiores e inferiores. Com o tempo, os indivíduos também apresentam dificuldade na fala e na deglutição, bem como alterações do movimento ocular. A doença causa dependência funcional, podendo levar o indivíduo a óbito. Apesar da severidade dos sintomas, as funções intelectuais, em geral, permanecem inalteradas (Jardim, 2000).

A doença de Machado-Joseph é uma ataxia crônica hereditária dominante, isso significa que, se um dos pais for portador do gene, o risco de transmitir a doença para cada filho é de 50%. De acordo com Donis et al. (2016), não foram identificadas diferenças na

prevalência entre os gêneros, homens e mulheres tendem a ser afetados nas mesmas proporções, com distribuições de idade semelhantes. O diagnóstico é feito baseando-se nos sintomas do paciente, em particular quando já existe histórico da doença na família, e através de um teste genético que avalia o número de repetições do trinucleotídeo CAG na região do cromossomo 14, que codifica o *gene* ATXN3 (Donis et al., 2016). A repetição CAG de um indivíduo normalmente varia de 12 a 43, enquanto que pessoas portadoras da doença apresentam mais de 51 repetições (Donis et al., 2016). No estudo de Kieling et al. (2007), a taxa de sobrevida após o início dos sintomas foi de 21 anos, em média. Neste mesmo estudo descobriu-se que o início precoce dos sintomas e o alto número de repetições CAG estão atrelados a uma taxa de sobrevida menor. Devido à variabilidade fenotípica da doença, pesquisadores costumam dividir os pacientes em subtipos, que variam de acordo com a idade do início dos sintomas e as principais manifestações. Coutinho e Andrade (1977) propuseram a seguinte classificação para os portadores da SCA3/MJD (as proporções podem variar de acordo com a região):

Tipo I (13% dos indivíduos): caracterizada por início precoce e espasticidade proeminente, rigidez, bradicinesia com poucos sintomas atáxicos. Está associada com maiores expansões CAG e é o tipo menos comum.

Tipo II (57% dos indivíduos): caracterizada por ataxia, sinais de neurônio motor superior com idade de início entre 20 e 45 anos.

Tipo III (30% dos indivíduos): manifestações tardias com ataxia e polineuropatia periférica com idade de início entre 40 e 60 anos.

Outros subtipos propostos na literatura foram omitidos pois a taxa de ocorrência é irrelevante e/ou desconhecida.

Apesar de não existir cura para a doença de Machado-Joseph, existem alguns tratamentos para contornar as manifestações clínicas, proporcionando melhor qualidade de vida aos pacientes. Dentre os principais cuidados estão o acompanhamento com fisioterapia, fonoterapia, terapia ocupacional e manejo específico dos sintomas mais graves (Donis, 2015). Nesse sentido, a avaliação da gravidade da doença é essencial.

3.1.1 Escalas de avaliação

Diversos instrumentos que avaliam o comprometimento neurológico da SCA3/MJD foram desenvolvidos com o passar dos anos. A primeira escala publicada, em 1997, foi a *International Cooperative Ataxia Rating Scale* (ICARS) (Trouillas et al., 1997). Em 2006 foi publicada a *Scale for Assessment and Rating of Ataxia* (SARA) (Schmitz-Hübsch et al., 2006)

e, na sequência, em 2008, a *Neurological Examination Score for Spinocerebellar Ataxia* (NESSCA) (Kieling et al., 2008). Neste trabalho serão utilizadas medidas de indivíduos que foram avaliados através das escalas NESSCA e SARA.

SARA

A SARA (Schmitz-Hübsch et al., 2006) surgiu como uma alternativa às escalas já consolidadas, devido à sua aplicação mais simples e rápida. Trata-se de um escore semiquantitativo para manifestações atáxicas. A escala possui 8 itens, cada um tem de cinco a nove categorias de resposta, cujo total varia de 0 até 40 pontos, onde escore 40 caracteriza a maior gravidade da doença. Os itens abordados pela SARA são: Marcha, Equilíbrio de Pé, Equilíbrio Sentado, Coordenação da Fala, Teste de Perseguição do Dedo, Teste Dedo-Nariz, Diadococinesia e Teste Calcanhar-Joelho-Canela.

No Anexo A deste trabalho encontra-se o instrumento de medida da SARA, completo, traduzido para o português.

NESSCA

Em 2001, um grupo de pesquisadores do Hospital de Clínicas de Porto Alegre (HCPA) iniciou o desenvolvimento de um novo instrumento de medida cuja publicação se deu em 2008, a NESSCA (Kieling et al., 2008), uma escala semiquantitativa para manifestações atáxicas e não atáxicas. A NESSCA é dividida em 18 itens, cada um possui de duas a cinco categorias de resposta, somando, no total, 40 pontos. Os sintomas abordados pela escala são: Ataxia de Marcha, Ataxia nos Membros, Nistagmo, Oftalmoplegia Externa Progressiva, Achados Piramidais, Disartria, Disfagia, Fasciculações, Perda Sensorial, Distonia, Rigidez, Bradicinesia, Retração Palpebral, Blefarospasmo, Amiotrofia Distal, Função do Esfíncter, Câimbra e Vertigem. A NESSCA leva de 30 a 40 minutos para ser administrado a um paciente (Kieling et al., 2008).

O diferencial em relação à SARA é que a NESSCA avalia também sintomas não-atáxicos, como por exemplo, Retração Palpebral, Rigidez e Função do Esfíncter. Dentre os resultados do artigo de Kieling et al. (2008), verificou-se que os escores da SARA tiveram alta correlação com os itens atáxicos e com os itens não atáxicos da NESSCA, sugerindo que a gravidade da doença também está correlacionada com os sintomas não atáxicos, contribuindo para a validade externa da escala proposta. Além disso, o escore da NESSCA mostrou-se altamente correlacionado com outras características utilizadas para mensurar a

severidade da doença (estágio, duração e repetições do trinucleotídeo CAG). O alfa de Cronbach's obtido foi de 0,77, corroborando com a consistência interna do questionário (Kieling et al., 2008).

No Anexo B deste trabalho encontra-se o instrumento de medida da NESSCA completo em português.

3.2 TEORIA DA RESPOSTA AO ITEM

Traço latente é o termo utilizado para descrever características de um indivíduo que não podem ser observadas diretamente, tais como: nível de satisfação do cliente, proficiência de um aluno em determinado assunto, qualidade de vida de uma comunidade e grau de depressão de um indivíduo (Moreira Jr., 2010). O traço latente pode ser inferido com base na observação de variáveis secundárias que estejam relacionadas a essa característica e que possam ser mensuradas através de instrumentos de medidas, como testes e questionários. De acordo com Castro (2008), tradicionalmente, a análise de dados provenientes dessas ferramentas é realizada através da Teoria Clássica dos Testes (TCT), que, geralmente, se baseia na soma ou na média de todos os itens. A necessidade de desenvolver uma técnica alternativa à TCT surgiu na psicometria, em meados dos anos 1960, em razão de limitações encontradas nesta metodologia. Nesse contexto, surgiu a Teoria da Resposta ao Item (TRI), que compreende de um conjunto de modelos que permite estimar uma variável não observável, isto é, um traço latente.

3.2.1 Teoria Clássica dos Testes

A Teoria Clássica dos Testes é um modelo mais antigo e ainda utilizado para a avaliação de traços latentes. Também conhecida como Teoria Clássica de Medida (TCM), o princípio fundamental é tomar como base o escore de todo o instrumento, geralmente a soma dos itens da escala. Essa teoria tem sido utilizada há mais de 100 anos na avaliação de testes educacionais. De acordo com Pasquali (2003), uma grande preocupação da comunidade psicométrica é com o erro contido na simples soma dos itens. O autor alerta para o fato de que esse escore representa, na verdade, uma magnitude, e que pessoas com mesmo escore podem ter comportamentos diferentes para o traço latente.

Uma das principais vantagens da TCT é que ela é um método simples, de fácil aplicação e não exige pressupostos rigorosos. Porém, é sabido que a TCT apresenta algumas limitações, dentre as quais se destacam (de Araujo et al., 2009):

a) Os parâmetros dos itens dependem da amostra de indivíduos na qual o instrumento foi aplicado: as avaliações do teste são válidas somente se a amostra for representativa ou se o instrumento for utilizado em outra amostra com características semelhantes;

b) Ausência de discriminação entre variáveis: o traço latente é medido com base no escore total do teste, pressupondo que todos os itens contribuam igualmente para a medida da variável latente. Ou seja, respondentes com o mesmo escore total são considerados iguais mesmo tendo um perfil diferente. Por exemplo, na NESSCA existem 18 itens, onde os indivíduos são classificados em diferentes níveis, que podem variar de zero a quatro, dependendo da severidade do sintoma. O escore máximo que pode ser obtido é de 40 pontos. Entretanto, existem milhares de combinações diferentes para obter um escore específico e indivíduos com um mesmo escore podem apresentar padrões muito diferentes no contexto clínico da doença.

c) Não permite a comparação entre populações submetidas a instrumentos de medida diferentes: a comparação entre indivíduos ou grupos de indivíduos somente é possível quando eles são submetidos aos mesmos testes.

Nesse sentido, as contribuições da Teoria da Resposta ao Item trouxeram alguns avanços na avaliação de traços latentes. Uma das grandes vantagens da TRI sobre a TCT é que ela permite a comparação de diferentes populações, pois sua principal característica é considerar como elementos os itens do instrumento e não apenas o escore final (Pasquali, 2003). Outro benefício dos modelos da TRI é o fornecimento de estimativas para os itens e para o traço latente que não variam com as características da população, tornando os resultados independentes do instrumento de medida e da amostra de respondentes (Castro, 2008). Hays et al. (2000) apontam para algumas vantagens da TRI sobre a TCT que se destacam no contexto clínico: a avaliação da contribuição de cada item, facilitando o processo de decisão quanto à exclusão ou inclusão de itens, a possibilidade de criar testes com diferentes tipos de itens, sem que aqueles com mais categorias impactem mais no escore final, como ocorre na TCT, e a CAT (*computerized adaptive testing*) – cujo item seguinte depende das respostas aos itens anteriores, adaptando o teste a medida que o respondente avança, permitindo estimar o traço latente utilizando apenas um subconjunto dos itens.

3.2.2 História da TRI

O termo análise de estrutura latente foi proposto por Paul F. Lazarsfeld, em 1959, para descrever uma classe de modelos matemáticos com o objetivo de estudar um comportamento através de atributos e indicadores relacionados (Kotz et al., 2006). Desde então, duas

abordagens foram traçadas para desenvolver os modelos da Teoria da Resposta ao Item, representadas pelo europeu G. Rasch e, nos Estados Unidos, por F. Lord.

Na Europa, o matemático dinamarquês Georg Rasch (1901-1980) conduziu as análises da avaliação de conhecimento de um grupo de recrutas para as Forças Armadas, logo após o término da Segunda Guerra Mundial (1947). O teste IGP (*Intelligence Group Test*), que consistia de 46 itens dicotômicos, foi aplicado a 1000 recrutas. Durante as análises, Rasch estimou o grau de dificuldade global para cada item e a inteligência de cada um dos recrutas foi estimada com base em uma avaliação da dificuldade dos itens que cada um poderia resolver e dos que não poderia (Boomsma et al., 2000). Em 1952 o modelo foi revisado e, posteriormente, deu origem ao modelo de Rasch, como é conhecido atualmente – cujo desenvolvimento foi publicado anos depois no livro *Probabilistic Models for Some Intelligence and Attainment Tests* (Rasch, 1993).

Paralelamente aos trabalhos conduzidos por G. Rasch, o início da Teoria da Resposta ao Item, nos Estados Unidos, é atribuído ao matemático Frederic Lord (1912-2000). Suas publicações nos anos 1950 fizeram com que a TRI ganhasse força, dando início ao desenvolvimento formal da teoria. Além disso, ele contribuiu para a posterior elaboração de programas computacionais, necessários para colocar a teoria em prática. Em 1968, junto com M. Novick (1932-1986) publicaram o clássico livro *Statistical Theories of Mental Test Scores* (Lord et al., 1968), o qual contém 4 capítulos sobre a TRI, contribuição de Allan Birnbaum (1923-1976).

Os conteúdos escritos por F. Lord, G. Rasch, juntamente com a contribuição de seus colegas, tornaram-se a base da moderna Teoria da Resposta ao Item. Nos anos seguintes, diversos pesquisadores aperfeiçoaram e desenvolveram outros modelos de TRI, estendendo os modelos originais para lidar com outras situações, com diferentes tipos de itens e com diferentes parâmetros para esses itens. Esses e outros modelos serão ilustrados no decorrer deste trabalho.

Um dos fatores que mais contribuíram para o crescimento e a utilização da TRI no contexto atual foi o avanço da tecnologia. As equações matemáticas que compõem o modelo podem se tornar complexas e a melhora no desempenho do processamento de computadores viabilizou os cálculos que o modelo TRI exige e permitiu refinamentos no método. A partir da década de 1970 ocorreu o desenvolvimento dos primeiros softwares e algoritmos apropriados para os cálculos do modelo (de Andrade et al., 2000).

3.2.3 Conceitos e modelos da TRI

A TRI é uma poderosa ferramenta estatística que surgiu para suprir as necessidades decorrentes das limitações da TCT. Oferece algumas das melhores alternativas para projetar e aperfeiçoar escalas, bem como realizar análises de itens (Castro 2008).

A TRI compreende um grupo de modelos lineares generalizados que conectam as respostas aos itens com o traço latente de interesse do pesquisador, respostas estas obtidas através de testes submetidos aos entrevistados (Castro et al., 2010). As equações matemáticas que compõem a TRI descrevem a associação entre a probabilidade de uma resposta a um item em particular e a magnitude de um indivíduo quanto a uma característica, usando uma função monótona não linear, exclusivamente crescente. No contexto da SCA3/MJD, a TRI modela a relação existente entre a probabilidade $P(\theta)$ de um indivíduo apresentar um sintoma e o nível de comprometimento pela doença, o traço latente θ. Essa relação entre a probabilidade e o traço latente é descrita pela curva característica do item (CCI) (de Andrade et al., 2000).

Para ilustrar essa relação, a Figura 1 mostra uma CCI fictícia de um sintoma dicotômico qualquer (que possui apenas duas opções de resposta – possui ou não possui o sintoma). O eixo horizontal θ representa o comprometimento com a doença de Machado-Joseph, considerando, neste caso, distribuição normal padrão com média igual a zero e desvio padrão igual a um. O eixo vertical $P(\theta)$ representa a probabilidade de o indivíduo apresentar o sintoma em questão.

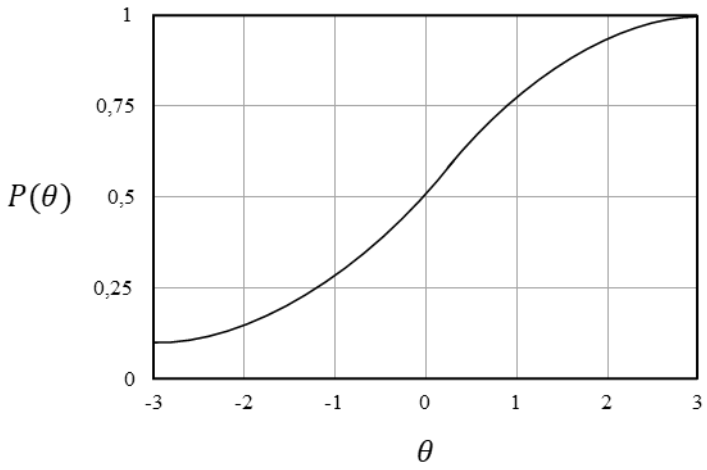

Figura 1. Exemplo de CCI.

Fonte: elaborado pelo autor.

São muitos os modelos de TRI propostos pelos autores, os quais podem ser facilmente diferenciados conforme três fatores fundamentais:

a) Natureza do item: dicotômicos (por exemplo, possui ou não possui o sintoma) ou politômicos (por exemplo, não apresenta o sintoma, sintoma fraco, sintoma moderado ou sintoma grave);

b) Número de populações envolvidas: uma ou mais;

c) Quantidade de traços latentes avaliados: um ou mais.

Vale lembrar que um mesmo teste pode ser formado por itens exclusivamente dicotômicos, exclusivamente politômicos ou também por uma combinação entre os dois, nesse caso, trata-se de um questionário de formato misto (em inglês, *mixed-format*). O instrumento de medida é criado conforme a necessidade do pesquisador.

Dentre os modelos TRI já estabelecidos na literatura, a maioria foi desenvolvida considerando unidimensionalidade – significa que somente um traço latente está sendo avaliado, ou, pelo menos, um se destaca em relação aos outros (de Andrade et al., 2000). Dessa forma, ao utilizar modelos unidimensionais, é importante verificar se essa suposição está satisfeita, de modo que todos os itens do instrumento de medida estejam mensurando o mesmo traço latente. Uma das formas de avaliar se a suposição de unidimensionalidade está atendida é através da análise fatorial (de Andrade et al., 2000). Autores sugerem que o primeiro fator deve explicar no mínimo 20% da variância total (Hattie, 1985). Há situações em que mais de um traço latente está envolvido, para esses casos, modelos multidimensionais também foram desenvolvidos (Reckase, 1997). Outra suposição que deve ser verificada é a propriedade de independência local, onde os itens não devem ser correlacionados uns com os outros, tomando como base o nível do traço latente. Se a suposição de unidimensionalidade estiver atendida, então a independência local também está satisfeita (Hambleton et al., 1991; Hays et al., 2000).

3.2.4 Modelos dicotômicos

Na prática, os modelos logísticos para itens coletados ou agrupados de forma dicotômica (certo/errado, concordo/discordo, sim/não, possui/não possui o sintoma) são os mais utilizados na educação (de Andrade et al., 2000). Dos 18 itens da NESSCA, cinco são avaliados de forma dicotômica, como por exemplo, o item Retração Palpebral, conforme Quadro 1:

Quadro 1. Item Retração Palpebral da NESSCA.

Item	Gravidade	Categoria de Resposta
13 – Retração Palpebral	Ausente.	0
	Presente.	1

Fonte: escala NESSCA.

Estes modelos podem ser diferenciados pela quantidade de parâmetros utilizados para descrever o item. São chamados de modelos logísticos de 1, 2 e 3 parâmetros, que consideram, respectivamente:

a) Dificuldade do item;

b) Dificuldade e discriminação do item;

c) Dificuldade, discriminação e a probabilidade de apresentar o sintoma em sujeitos cuja magnitude do traço latente é baixa ou nula.

As definições dos modelos apresentadas neste capítulo são baseadas em Andrade et. al. (2000), cujas interpretações foram adaptadas para o contexto deste trabalho, que trata do comprometimento pela SCA3/MJD.

Modelo Logístico de 3 Parâmetros

Por ser um modelo mais generalista, autores defendem que, dentre os modelos dicotômicos propostos pela TRI, o modelo logístico unidimensional de 3 parâmetros é o mais utilizado (de Andrade et al., 2000). Através do 3PL (em inglês, *Three-parameter Logistic Model*), proposto por F. Lord (Lord, 1980), podem ser obtidos, facilmente, os modelos com 1 e 2 parâmetros, 1PL e 2PL, respectivamente. A probabilidade de um indivíduo j, com comprometimento pela SCA3/MJD de θ_j, apresentar o sintoma descrito pelo item i é dada por:

$$P\left(U_{ij}=1\middle|\theta_j\right)=c_i+(1-c_i)\frac{1}{1+e^{-Da_i(\theta_j-b_i)}}, \tag{1}$$

com $i=1,2,\dots,I$ e $j=1,2,\dots,n$, onde:

U_{ij} é uma variável dicotômica que assume o valor 1 quando o indivíduo j possui o sintoma do item i ou 0 quando não possui o sintoma;

θ_j representa o comprometimento pela SCA3/MJD (traço latente) do j-ésimo indivíduo;

$P\left(U_{ij}=1\middle|\theta_j\right)$ é a probabilidade de um indivíduo j com comprometimento pela SCA3/MJD θ_j apresentar o sintoma descrito no item i;

a_i é o parâmetro de inclinação (ou discriminação) do item i;

b_i é o parâmetro de posição do item i, estimado na mesma unidade de θ, representando a gravidade do sintoma medido pelo item i;

c_i é o parâmetro do item que representa a probabilidade de indivíduos

apresentarem o sintoma descrito pelo item i mesmo quando tem baixo comprometimento da doença, isto é, baixo nível do traço latente;

D é um fator de escala, constante e igual a 1. Utiliza-se 1,7 quando se deseja que a função logística forneça resultados semelhantes ao da função ogiva normal;

I é o número de itens no instrumento de medida;

n é o número de indivíduos respondentes.

O modelo parte do princípio que indivíduos que se encontram mais comprometidos pela doença (SCA3/MJD), ou seja, com valor mais alto para o parâmetro θ_j, têm maior probabilidade de apresentarem os sintomas. Essa relação pode ser representada pela curva característica do item, cujo formato é definido pelos parâmetros dos itens a_i, b_i e c_i. Cada item é representado por um gráfico, onde cada curva representa uma categoria de resposta. A Figura 2 ilustra onde cada um dos parâmetros atua no comportamento da CCI.

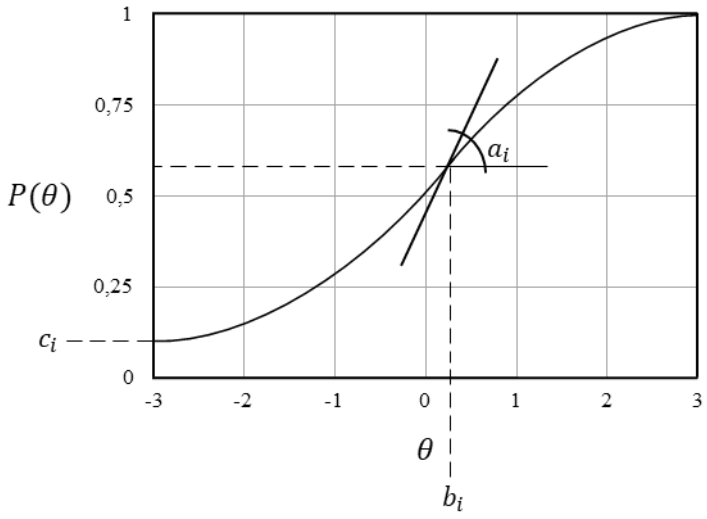

Figura 2. Parâmetros da CCI para um sintoma fictício ***i***.

Fonte: elaborado pelo autor.

O parâmetro a_i reflete o poder de discriminação (ou inclinação) do item quanto à intensidade do comprometimento pela doença e é proporcional ao ângulo de inflexão da curva. Itens com valores baixos de a_i indicam sintomas com pouco poder de discriminação, onde indivíduos com diferentes níveis de comprometimento têm aproximadamente a mesma probabilidade de apresentar o sintoma descrito pelo item. O parâmetro b_i é medido na unidade do traço latente e representa o nível necessário do traço latente para uma probabilidade de presença do sintoma igual a $(1 + c_i)/2$, assim, quanto maior o valor de b_i, mais grave o

sintoma. Também é no ponto $\theta = b_i$ onde ocorre a inflexão da curva. Por sua vez, o parâmetro c_i é uma probabilidade que varia de 0 a 1 e representa a probabilidade de um indivíduo com baixo comprometimento de SCA3/MJD apresentar um sintoma descrito por um determinado item. Conforme observado na Figura 2, o parâmetro c_i coincide com o local onde a curva corta o eixo vertical.

O parâmetro θ pode assumir qualquer valor entre $-\infty$ e $+\infty$. Porém, é comum a utilização de uma convenção escalar com média igual a zero e desvio padrão igual a um, como no gráfico da Figura 2. Essa escala facilita a interpretação. Por exemplo, um indivíduo com $\theta = 1$ está um desvio padrão acima da média. Nesse caso, o parâmetro b_i também sofre alterações e seus valores geralmente se concentram no intervalo (-3,3). A maioria dos softwares já adota esse comportamento padronizado, que também será utilizado ao longo deste trabalho.

A CCI permite visualizar o quanto mudanças na magnitude do traço latente são refletidas de acordo com uma resposta específica, ou seja, o quanto mudança na presença e/ou intensidade de cada sintoma se relacionam com o nível de comprometimento pela doença. Para ilustrar mudanças nos parâmetros, alguns exemplos de CCIs estão apresentadas nas figuras que seguem.

A Figura 3 ilustra variações no parâmetro de inclinação, a_i. Foram fixados os valores de $b_i = 0$ e $c_i = 0$, variando $a_i = 1$ (curva preta) e $a_i = 2$ (curva vermelha). Itens com maior valor para o parâmetro a_i têm a curva característica com inclinação mais acentuada, em outras palavras, o item representado pela curva vermelha é mais apropriado para discriminar os indivíduos, visto que as variações no traço latente impactam mais as probabilidades ao longo da curva. Por exemplo, um indivíduo com comprometimento $\theta = 1$ tem maior probabilidade de apresentar o sintoma representado pela curva vermelha do que o sintoma representado pela curva preta. Enquanto que um indivíduo com comprometimento $\theta = -1$ tem menor probabilidade de apresentar o sintoma representado pela curva vermelha do que o sintoma representado pela curva preta. Espera-se que os itens de um instrumento de medida sejam bastante discriminadores, com valores mais altos de a_i, geralmente a partir de um, variando no intervalo positivo de zero a dois (de Andrade et al., 2000).

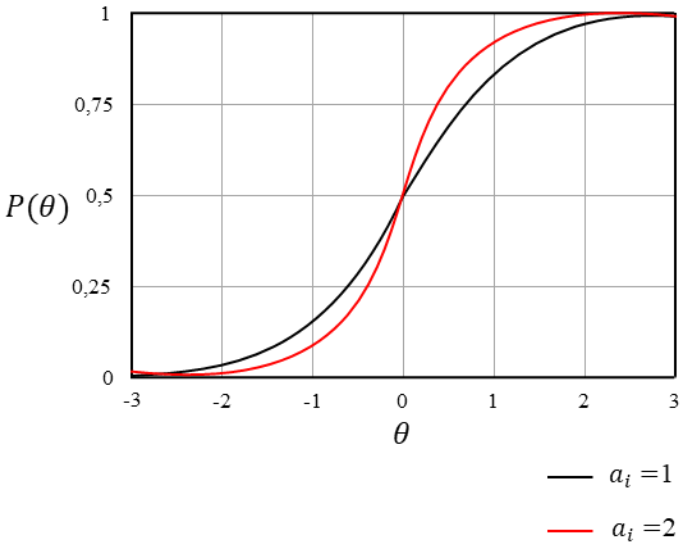

Figura 3. Exemplo de CCI variando o parâmetro $\boldsymbol{a_i}$.

Fonte: elaborado pelo autor.

Na Figura 4, fixando os parâmetros $a_i = 1$ e $c_i = 0$, foram traçadas duas curvas, representando dois sintomas diferentes, com valores de $b_i = -0{,}5$ (curva preta) e $b_i = 0{,}5$ (curva vermelha). A curva vermelha, com maior valor para o parâmetro b_i, significa que é preciso um comprometimento maior para que os pacientes comecem a apresentar este sintoma com probabilidade de 50% ou mais. Ou seja, para indivíduos com $\theta = 0{,}5$, se tem 50% ou mais de probabilidade de ocorrência do sintoma, enquanto esse mesmo percentual de 50% se observa na curva preta para pacientes a partir de $\theta = -0{,}5$, com meio desvio padrão abaixo da média. Dessa forma, o sintoma representado pela curva vermelha está associado a um sintoma mais grave, com maiores valores para o traço latente.

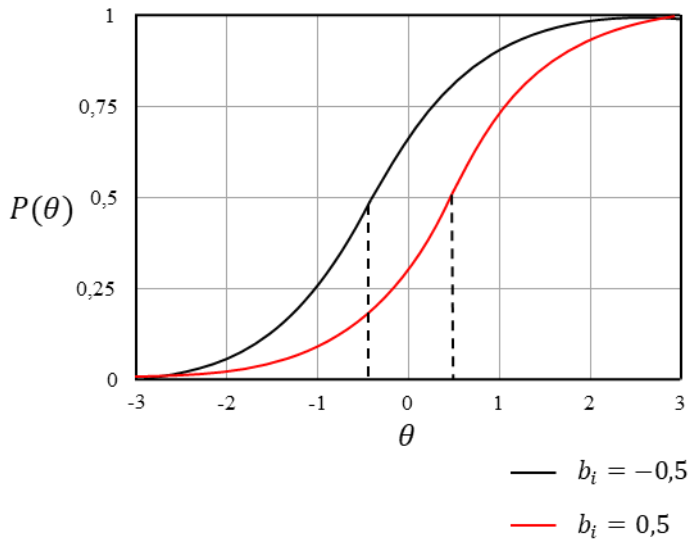

Figura 4. Exemplo de CCI variando o parâmetro $\mathbf{b_i}$.

Fonte: elaborado pelo autor.

Para ilustrar variações no comportamento do parâmetro c_i foram fixados os valores de $b_i = 0$ e $a_i = 1$, variando $c_i = 0$ (curva preta) e $c_i = 0{,}2$ (curva vermelha), conforme apresentado na Figura 5. Este parâmetro só pode assumir valores dentro do intervalo (0,1). A curva vermelha representa um sintoma que, mesmo indivíduos com baixíssimo comprometimento pela doença, $\theta = -3$, possuem 20% de chance de apresentarem o sintoma, enquanto que, para o item representado pela curva preta, estes mesmos indivíduos tem probabilidade praticamente nula de apresentarem o sintoma. Nas escalas clínicas, não espera-se que c_i seja muito diferente de zero.

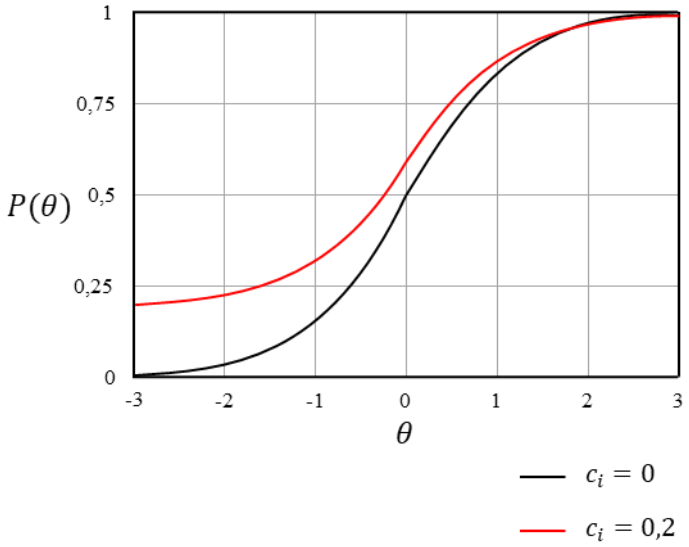

Figura 5. Exemplo de CCI variando o parâmetro $\mathbf{c_i}$.

Fonte: elaborado pelo autor.

A partir do 3PL podem ser obtidos os modelos 2PL e 1PL.

Modelo Logístico de 2 Parâmetros

Proposto inicialmente por F. Lord (Lord, 1952) e revisto por A. Birnbaum nos capítulos do livro *Statistical Theories of Mental Test Scores* (Lord et al., 1968), o modelo 2PL (em inglês, T*wo-parameter Logistic Model*) exclui o parâmetro c_i, mantendo apenas os parâmetros de gravidade, b_i, e inclinação, a_i, logo, nesse caso, $c_i = 0$. O modelo é descrito por:

$$P(U_{ij} = 1 | \theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}, \quad (2)$$

com $i = 1,2,\ldots,I$ e $j = 1,2,\ldots,n$.

Os parâmetros do 2PL são interpretados da mesma forma que os parâmetros do 3PL.

Modelo Logístico de 1 Parâmetro

Por sua vez, no 1PL (do inglês, *One-parameter Logistic Model*), além de ser retirado o parâmetro c_i, também é excluído o parâmetro a_i, considerando que todos os itens possuem o mesmo poder de discriminação e, consequentemente, a mesma inclinação. O 1PL, que também é conhecido como modelo de Rasch, em homenagem ao seu desenvolvedor (Rasch, 1960), é dado por:

$$P(U_{ij} = 1|\theta_j) = \frac{1}{1 + e^{-D(\theta_j - b_i)}}, \quad (3)$$

com $i = 1,2, \dots, I$ e $j = 1,2, \dots, n$.

Os parâmetros do 1PL são interpretados da mesma forma que os parâmetros do 3PL.

Estes modelos, que representam uma relação entre o traço e os itens de resposta dicotômica, são amplamente utilizados na avaliação educacional, onde geralmente há uma opção certa e outra errada. No entanto, tem baixa aplicação na área da saúde porque muitos dos instrumentos utilizados para avaliação de sintomas e diagnóstico são compostos por itens com múltiplas categorias, geralmente ordenadas. Os pesquisadores optam por este formato com mais opções porque o resultado do teste se torna mais informativo, possibilitando aos entrevistados graduarem suas respostas de acordo com o nível de concordância com o que é expresso pelo item, agregando mais informações do que uma resposta dicotômica (Castro, 2008).

Por exemplo, um dos sintomas frequentes na SCA3/MJD é a disartria, cuja avaliação é feita pelo item 6 do questionário NESSCA. Disartria é a dificuldade com que o indivíduo articula as palavras para expressar-se. Na NESSCA esse item foi dividido em cinco diferentes categorias, variando entre a ausência do sintoma até a total impossibilidade de articular palavras (anartria), conforme mostrado no Quadro 2.

Quadro 2. Item Disartria da NESSCA.

Item	Gravidade	Categoria de Resposta
6 – Disartria	Ausente.	0
	Leve: Dificuldade de fala, mas fácil de entender.	1
	Moderado: discurso compreensível, mas com dificuldade.	2
	Grave: discurso de difícil compreensão.	3
	Anartria.	4

Fonte: escala NESSCA.

Existem vários modelos TRI para dados politômicos, adequados para esse formato de resposta. Alguns são exclusivos para dados com respostas ordinais, como é o caso do item representado no Quadro 2.

3.2.5 Modelos politômicos

A seguir serão apresentados breves resumos dos principais modelos para dados politômicos (com mais de duas categorias de resposta, ordenadas ou não). Algumas equações serão omitidas por pertencerem a modelos que não serão adequados para modelar itens com níveis de resposta ordinal, com no caso da NESSCA e da SARA. Suas fórmulas podem ser consultadas em de Andrade et al. (2000).

Modelo de Resposta Nominal

Este modelo é baseado no 2PL para instrumentos de medida com itens de múltipla escolha. O modelo NRM (em inglês, *Nominal Response Model*), proposto por R. Bock (Bock, 1972), assume que não há nenhuma ordenação nas opções de reposta, dessa forma, não é adequado para modelar os dados obtidos através das escalas NESSCA e SARA.

Modelo de Resposta Gradual

O Modelo de Resposta Gradual (Samejima, 1969), diferentemente do NRM, assume que as categorias de um item tenham uma ordem. Conhecido por GRM (em inglês, *Graded-Response Model*), trata-se de uma generalização do 2PL e é considerado um modelo TRI "indireto" pois requer um procedimento adicional para calcular a probabilidade condicional de um indivíduo ter um determinado nível do sintoma. Uma vantagem do GRM é que os itens do instrumento não precisam ter a mesma quantidade de categorias de resposta, como ocorre na NESSCA e na SARA. Considerando que as possíveis categorias de um item sejam denotadas por $k = 0,1, \dots, m_i$ onde $m_i + 1$ é o número de categorias do item i, a probabilidade de um indivíduo j pertencer a uma particular categoria ou outra mais alta pode ser dada por:

$$P_{i,k}^{+}(\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_{i,k})}}, \tag{4}$$

com $i = 1,2, \dots, I, j = 1,2, \dots, n$ e $k = 0,1, \dots, m_i$, onde:

θ_j representa a intensidade do comprometimento pela SCA3/MJD (traço latente) do j-ésimo paciente;

a_i é o parâmetro de inclinação comum a todas as categorias de um mesmo item i;

$b_{i,k}$ é o parâmetro de posição da k-ésima categoria do item i, ou seja, representa o nível de comprometimento necessário para a escolha da categoria de resposta k, ou acima de k, com probabilidade igual a 0,50;

D é um fator de escala, constante e igual a 1. Utiliza-se 1,7 quando se deseja que a função logística forneça resultados semelhantes ao da função ogiva normal;

I é o número de itens no instrumento de medida;

n é o número de indivíduos respondentes.

Deverá existir uma ordenação entre as categorias de um dado item, ou seja:

$$b_{i,1} \leq b_{i,2} \leq \cdots \leq b_{i,m_i}$$

A probabilidade de um indivíduo j pertencer a categoria k no item i é dada pela expressão:

$$P_{i,k}(\theta_j) = P^{+}_{i,k}(\theta_j) - P^{+}_{i,k+1}(\theta_j) \tag{5}$$

onde $P^{+}_{i,0}(\theta_j) = 1$ e $P^{+}_{i,m_i+1}(\theta_j) = 0$, logo:

$$P_{i,k}(\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_{i,k})}} - \frac{1}{1 + e^{-Da_i(\theta_j - b_{i,k+1})}} \tag{6}$$

O número de parâmetros, por item, será dado pelo número de categorias k do item i.

Por se tratar de modelos para itens politômicos, com mais de duas categorias, a equação (6) gera as curvas de categoria de resposta (CCR), as quais são simbolizadas por $P_{ik}(\theta)$. Estas curvas ilustram a relação entre a probabilidade de um indivíduo com comprometimento θ pertencer a categoria k do sintoma do item i. O conjunto de todas as CCRs de um teste resulta na curva característica do teste (CCT) que representa a probabilidade de obtenção de um escore total em função de θ, pode ser interpretado como a média para um dado valor de θ. Para um teste com I itens, a CCT é dada por:

$$T(\theta) = \sum_{i=1}^{I} \sum_{k=1}^{m_i} U_{ik} P_{ik}(\theta) \tag{7}$$

onde U_{ik} é uma função do escore do item, ou seja, são os valores de escore possíveis de obter no item i.

Geralmente utiliza-se $U_{ik} = k - 1$ (quando uma resposta associada a primeira categoria recebe escore zero, que é o caso da NESSCA e da SARA) ou $U_{ik} = k$ (quando uma

resposta associada a primeira categoria representa escore igual a 1) (Kolen e Brennan, 2014). A coluna "Categoria de Resposta" presente nos Quadros 1 e 2 pode ser interpretada como U_{ik}.

Modelo de Escala Gradual

Este modelo se trata de um caso particular do GRM, também conhecido por RSM (em inglês, *Rating Scale Model*), foi proposto por Andrich (Andrich, 1978). Além de atender a suposição das categorias ordenadas, os escores das categorias deverão ser igualmente espaçados. Se os itens que compõem o instrumento de medida diferem na quantidade de categorias (como na NESSCA e na SARA), o RSM não será adequado para modelar os dados.

Modelo de Crédito Parcial

Proposto por Masters (Masters, 1982), este modelo é também conhecido como PCM (em inglês, *Partial Credit Model*). Assim como os modelos GRM e RSM, o PCM foi desenvolvido para análise de dados politômicos com duas ou mais categorias de respostas ordenadas. Ele difere, entretanto, por pertencer à família dos modelos de Rasch, pois é uma extensão do 1PL para itens dicotômicos. Logo, se pressupõe que todos os itens possuem o mesmo poder de discriminação ($a_i = 1$).

Supondo que o item i possui $m_i + 1$ categorias de resposta ordenadas, temos que o modelo é dado por:

$$P_{i,k}(\theta_j) = \frac{exp\left[\sum_{u=0}^{k}(\theta_j - b_{i,u})\right]}{\sum_{u=0}^{m_i} exp\left[\sum_{v=0}^{u}(\theta_j - b_{i,v})\right]}, \tag{8}$$

com $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, n$, $k = 0, 1, \ldots, m_i$ e $b_{i,0} \equiv 0$.

Os parâmetros do PCM são interpretados da mesma forma que os parâmetros do GRM.

Modelo de Crédito Parcial Generalizado

O modelo de crédito parcial generalizado foi desenvolvido por Muraki, em 1992, e consiste de uma generalização do PCM, relaxando a hipótese de poder de discriminação igual para todos os itens (Muraki, 1992). Ou seja, permite que os itens dentro de uma escala tenham diferentes parâmetros de inclinação, o que é interessante no contexto da NESSCA e da SARA. Também conhecido como *Generalized Partial Credit Model* (GPCM), supondo que o item i possui $m_i + 1$ categorias de resposta ordenadas ($k = 0, 1, \ldots, m_i$), temos que o modelo é dado por:

$$P_{i,k}(\theta_j) = \frac{exp\left[\sum_{u=0}^{k} Da_i(\theta_j - b_{i,u})\right]}{\sum_{u=0}^{m_i} exp\left[\sum_{v=0}^{u} Da_i(\theta_j - b_{i,v})\right]}, \quad (9)$$

com $i = 1,2,\dots,I, j = 1,2,\dots,n$ e $k = 0,1,\dots,m_i$, onde:

θ_j representa a intensidade do comprometimento pela SCA3/MJD (traço latente) do j-ésimo paciente;

$P_{i,k}(\theta_j)$ é a probabilidade de um indivíduo com nível de comprometimento θ_j ter um sintoma na categoria k dentre as $m_i + 1$ categorias do item i;

a_i é o parâmetro de inclinação do item i;

$b_{i,k}$ é o parâmetro do item que regula a probabilidade do sintoma ser k ao invés da categoria adjacente $(k-1)$ no item i. Cada parâmetro $b_{i,k}$ corresponde ao valor do traço latente no qual o indivíduo tem a mesma probabilidade de ser classificado nas categorias k e $(k-1)$, isto é, onde $P_{i,k}(\theta_j) = P_{i,k-1}(\theta_j)$. Pode ser interpretado como um parâmetro de interseção entre as categorias de resposta do item i;

D é um fator de escala, constante e igual a 1. Utiliza-se 1,7 quando se deseja que a função logística forneça resultados semelhantes ao da função ogiva normal;

I é o número de itens no instrumento de medida;

n é o número de indivíduos respondentes.

É importante observar que o parâmetro de inclinação (a_i) presente neste modelo não deve ser interpretado diretamente, da mesma forma como nos modelos dicotômicos, na interpretação do poder de discriminação dos itens. Nos modelos politômicos, a discriminação do item depende da combinação de a_i com a distribuição dos parâmetros $b_{i,k}$. Os parâmetros $b_{i,k}$ são os pontos na escala onde as curvas das categorias se cruzam, em qualquer ponto da escala θ_j. No geral, define-se $b_{i,0} = 0$. Além disso, frequentemente os parâmetros $b_{i,k}$ são decompostos em um parâmetro de posição b_i e nos parâmetros para as categorias, $d_{i,k}$, onde:

$$b_{i,k} = b_i - d_{i,k} \quad (10)$$

Para avaliar a contribuição de um item politômico pode-se observar a Curva de Informação do Item (CII). Essa curva indica a quantidade de informação que um determinado sintoma contribui para a medida do traço latente e em qual intervalo esse sintoma é mais informativo (Castro et al., 2010).

Com o objetivo de visualizar os modelos de forma conjunta, o Quadro 3 resume as características de aplicação para cada modelo apresentado neste trabalho. Tanto o modelo GRM quanto o GPCM são adequados para aplicação nos dados obtidos com os questionários da NESSCA e da SARA. Os modelos 2PL e 3PL podem servir para modelar os itens dicotômicos da NESSCA.

Quadro 3. Resumo dos modelos TRI.

Modelo	**Quantidade de Categorias**	**Parâmetros**			**Observação**	**NESSCA e SARA**
		(a_i)	(b_i)	(c_i)		
1PL	Dicotômico		✓			
2PL	Dicotômico	✓	✓			✓*
3PL	Dicotômico	✓	✓	✓		✓*
Resposta Nominal (NRM)	Politômico	✓	✓		Categorias não ordenadas.	
Resposta Gradual (GRM)	Politômico	✓	✓		Categorias ordenadas.	✓
Escala Gradual (RSM)	Politômico	✓	✓		Categorias ordenadas e na mesma quantidade.	
Crédito Parcial (PCM)	Politômico		✓		Categorias ordenadas mas não necessariamente na mesma quantidade.	
Crédito Parcial Generalizado (GPCM)	Politômico	✓	✓		Categorias ordenadas mas não necessariamente na mesma quantidade.	✓

*Somente para os itens dicotômicos da NESSCA.

Fonte: adaptado de Castro (2008).

3.2.6 Estimação dos parâmetros

Nos modelos da TRI, a probabilidade de uma resposta a um determinado item depende de duas coisas: do traço latente do respondente e dos parâmetros dos itens, assim é preciso estimá-los. Geralmente, tanto os traços latentes quanto os parâmetros dos itens são desconhecidos e a estimação é feita pelo método da máxima verossimilhança (de Andrade et al., 2000). Métodos bayesianos também costumam ser empregados (Mislevy, 1986).

A calibração pode ser conjunta ou separada, nesse último caso, estimando primeiro os parâmetros dos itens e depois os traços latentes. Na calibração conjunta, devido à quantidade de parâmetros a serem estimados simultaneamente, o processo é iniciado considerando

valores conhecidos para os traços latentes (escores padronizados, por exemplo) até que se obtenham estimativas para os parâmetros dos itens, seguido pela estimação dos traços latentes. Essa etapa ocorre através da aplicação de um método iterativo, até que seja atingido algum critério para cessar o processo. A calibração separada considera uma distribuição acumulada associada aos traços latentes dos indivíduos, possibilitando o uso do método da máxima verossimilhança marginal para estimação dos parâmetros dos itens. Estes métodos estão detalhados em no livro de Andrade et al. (2000) e ambos podem ser alcançados computacionalmente. Vale lembrar que o conceito de calibração separada, aqui, significa estimar primeiro os parâmetros dos itens e depois os traços latentes. Na equalização esse conceito será apresentado novamente mas com outro significado.

Os procedimentos descritos resultam em equações para estimação dos parâmetros dos itens, a_i, b_i e c_i (no caso do 3PL), as quais envolvem integrais que não possuem solução analítica. Um procedimento de aproximação bastante utilizado no contexto da TRI é método de quadratura gaussiana, onde obtém-se o valor da integral somando a área de uma quantidade finita de retângulos. Os nós ou pontos de quadratura são os pontos médios de cada um dos retângulos (de Andrade et al., 2000). Não há uma regra estabelecida no que se refere à quantidade de pontos de quadratura, alguns autores reportam que 10 pontos são insuficientes (Kolen e Brennan, 2014; Nering e Ostini, 2010) enquanto que 100 seria uma quantidade razoável (Kim e Lee, 2004). Chalmers (2012) propõe uma regra onde a quantidade de pontos varia conforme o número de traços latentes envolvidos no modelo, por exemplo, $1 =$ 61 pontos, $2 = 31$ pontos, $3 = 15$ pontos e $4 = 9$ pontos.

Softwares computacionais como BILOG-MG (Zimowski et al., 2003), MULTILOG (Thissen et al., 2003) e PARSCALE (Muraki e Bock, 2003) foram desenvolvidos especificamente para o problema da estimação de parâmetros da TRI e, por isso, são rápidos e eficientes. Cada um é especializado em um tipo de situação, por exemplo, enquanto que o PARSCALE permite utilizar simultaneamente modelos dicotômicos e modelos politômicos, para escalas de formato misto, o BILOG-MG só realiza análises de modelos do tipo 3PL, 2PL e 1PL (Kolen e Brennan, 2014). Versões *trial* de 15 dias dos softwares BILOG-MG, MULTILOG e PARSCALE podem ser encontradas na *Scientific Software International*: http://www.ssicentral.com/. Na linguagem SAS os principais procedimentos para calibração dos parâmetros são o PROC IRT e o PROC MCMC, esse último para métodos bayesianos. O software R, por ser um software livre, dispõe de uma série de pacotes que estimam os parâmetros da TRI, entre os mais utilizados estão o *mirt* (Chalmers, 2012) e o *ltm* (Rizopoulos, 2006).

3.3 EQUALIZAÇÃO DE ESCALAS

Na área da saúde, as escalas de medidas são ferramentas integrantes da prática clínica e da avaliação do estado de saúde do paciente, pois possuem grande influência sobre as decisões que se referem ao tratamento e que envolvem a gestão de políticas públicas direcionadas para o cuidado com a saúde da população. O avanço em pesquisas e as necessidades cada vez mais específicas contribuem para o surgimento de muitas ferramentas de medida (Coluci et al., 2015).

Um mesmo traço latente pode ser avaliado através de diferentes instrumentos, por exemplo:

a) Intensidade da dor: NRS (*Numerical Rating Scale*), VRS (*Verbal Rating Scale*) e VAS (*Visual Analogue Scale*);

b) Avaliação da qualidade de vida: WHOQOL (*World Health Organization Quality of Life*) e QOLS (*Quality of Life Scale*);

c) Intensidade dos sintomas depressivos: BDI (*Beck Depression Inventory*), CES-D (*Centre for Epidemiological Studies - Depression Scale*), PHQ (*Patient Health Questionnaire*) e HADS (*Hospital Anxiety Depression Scale*);

d) Comprometimento por ataxias espinocerebelares: ICARS, SARA e NESSCA.

Ao utilizarem-se diferentes escalas, existe uma dificuldade em comparar os resultados obtidos nas publicações científicas. Nesse sentido, a proposta da equalização é estabelecer uma equivalência entre escalas distintas ou diferentes populações, tornando os itens e/ou características de interesse (traços latentes) comparáveis. Essa oportunidade de poder confrontar os resultados é bastante interessante no contexto das ataxias espinocerebelares, pois permitirá comparar indivíduos que foram submetidos a diferentes escalas, além de possibilitar a criação de um novo instrumento de medida, ponderando os itens mais relevantes e com maior contribuição de cada questionário. Ainda, no caso de testes que causam incômodos e/ou cansaço ao paciente, a equalização de escalas permite que o indivíduo responda a apenas um instrumento, reduzindo o desconforto causado aos entrevistados.

Até a década de 1980, na comunidade acadêmica, apenas profissionais ligados à psicometria dedicavam-se a prática e a pesquisa em equalização. A partir desse momento, outros profissionais de áreas que fazem uso de testes, como na educação, enxergaram potencial na técnica e passaram a explorá-la (Kolen e Brennan, 2014). Nos anos 1990 se deram as primeiras publicações e, desde então, vem ganhando força e adeptos, inclusive no

Brasil. Na área da saúde, ainda são poucas as aplicações da equalização de escalas para avaliação de desfechos clínicos (Chen et al., 2009; McHorney e Cohen, 2000).

Na avaliação educacional, um exemplo da aplicação da equalização de escalas no Brasil é a prova anual do SAEB (Sistema de Avaliação da Educação Básica), cuja finalidade é realizar um diagnóstico da educação básica brasileira. No caso da 3ª série do Ensino Médio são criados, no total, 169 itens, divididos em 13 blocos de 13 itens cada. Cada aluno recebe um caderno com 3 blocos distintos, totalizado 39 itens. Diferentes blocos não possuem itens comuns, mas diferentes cadernos podem ter blocos comuns, ou seja, itens comuns. Isso permite avaliar os alunos em uma mesma escala de conhecimento, mesmo que sejam submetidos a diferentes cadernos de prova (de Andrade et al., 2000).

Na área da saúde, Chen et al. (2009) utilizou a equalização de escalas com dados de dois estudos que avaliam a intensidade da dor em pacientes. Como resultado, foi estabelecida uma conexão entre as escalas que permitiu avaliar os pacientes em uma mesma métrica, não importando a qual escala o paciente tenha sido submetido.

3.3.1 Conceitos da equalização

Existem conceitos similares que estão relacionados e surgem quando se está buscando por técnicas de comparação de escalas, os termos mais utilizados na literatura são *equating* e *linking*. O glossário Inglês-Português de Estatística (2011), editado em conjunto pela Associação Brasileira de Estatística (ABE) e pela Sociedade Portuguesa de Estatística (SPE), determina que a expressão *test equating methods* pode ser traduzida por "métodos de equiparação de testes", mas muitos autores utilizam o termo equalização (de Andrade et al., 2000). Apesar de não constar no glossário, *linking* pode ser traduzido de muitas formas, como ligação, encadeamento ou lincagem. Embora ambos façam uso de procedimentos estatísticos semelhantes, o termo *equating* pode ser visto como um caso particular de *linking*, este último sendo um procedimento genérico que busca a comparabilidade entre escores de diferentes instrumentos de medida. Enquanto que, no contexto de *equating*, estes diferentes instrumentos de medida devem ter sido construídos sob a mesma perspectiva, para mensurar um mesmo traço latente (Nering e Ostini, 2010).

3.3.2 Propriedades da equalização

Para obter uma equalização adequada, algumas propriedades devem ser observadas e asseguradas durante o processo. Sejam X e Y dois instrumentos de medida diferentes que

avaliam um mesmo traço latente. Os autores Kolen e Brennan (2014) propuseram um conjunto de propriedades que, quando atendidas, levam a um bom resultado:

a) Propriedade da simetria: a função utilizada para transformar um escore do instrumento de medida Y para a escala do instrumento de medida X deverá ser a inversa da função utilizada para transformar um escore do instrumento de medida X para a escala do instrumento de medida Y. Por exemplo, se o escore y na escala Y resulta no escore x na escala X, então o escore x na escala X deverá, também, ser convertido para o escore y na escala Y;

b) Propriedade das mesmas especificações: os instrumentos de medida devem mensurar o(s) mesmo(s) traços(s) latente em um mesmo nível de aprofundamento e devem ser construídos sob as mesmas condições e com mesmo grau de confiabilidade. Além disso, devem ser testados e validados conforme protocolo padrão;

c) Propriedade da equalização do escore observado: esta propriedade define que a equalização terá sido bem sucedida se a distribuição dos escores dos indivíduos que responderam ao instrumento de medida Y é semelhante à distribuição daqueles que responderam ao instrumento de medida X. Por exemplo, dentre os respondentes do instrumento de medida Y há uma dada proporção de indivíduos que obtiveram escore maior que y e, ainda, verificou-se que o escore y no instrumento Y equivale ao escore x no instrumento X. Essa proporção deverá ser semelhante entre os que responderam ao instrumento de medida X e obtiveram escore maior que x;

d) Propriedade da equidade: para um indivíduo com traço latente θ, a distribuição de probabilidade condicional $f(y|\theta)$ de um escore qualquer y deverá ser igual à distribuição de probabilidade condicional $f\left(e_y(x)\middle|\theta\right)$ do escore equalizado $e_y(x)$;

e) Propriedade da invariância populacional: a escolha da população para estimar a função de equalização não deve interferir, ou seja, a função de equalização deve ser independente do grupo de respondentes.

Na prática, a propriedade da simetria é um pré-requisito para que a relação entre as escalas seja de equalização e exclui a possibilidade de utilizar a regressão como um método de equalização, pois a regressão de Y em X, geralmente, difere da regressão de X em Y. A propriedade que tem por finalidade garantir as mesmas especificações entre os instrumentos de medida é essencial e, caso não seja cumprida, independentemente do procedimento estatístico utilizado, os escores não serão intercambiáveis.

A propriedade da equidade pode ser difícil de alcançar e sua aplicação é de caráter teórico, visto que só poderá ser atingida em sua totalidade se os instrumentos de medida forem idênticos, o que não faria sentido, embora instrumentos de medida muito diferentes em termos de abrangência e profundidade possam violar este item e também a propriedade da invariância populacional. Em função disso, Goldstein (1984) sugere que a propriedade de equidade pode ser parcialmente relaxada a ponto de tornar possível alcançá-la, a qual é chamada de propriedade da equidade de primeira ordem (em inglês, *first-order equity property*). Para respondentes com traço latente θ a esperança condicional de um escore y obtido no instrumento Y é igual a esperança condicional de $e_y(x)$, ou seja:

$$E\left[e_y(X)|\theta\right] = E(Y|\theta) \tag{11}$$

Espera-se que os respondentes obtenham o mesmo escore equalizado caso respondessem ao outro instrumento de medida, em média, mas não exige que as distribuições sejam exatamente iguais. Estes cinco requisitos devem ser utilizados como guias para a prática da equalização. Além disso, também é importante avaliar se as condições de medição foram semelhantes para todos os indivíduos, reduzindo possíveis vieses que possam ocorrer na etapa da coleta de dados. Por exemplo, se todos os entrevistadores receberam o mesmo treinamento e aplicaram os instrumentos da mesma forma, se tinham os recursos disponíveis para avaliar os sintomas dos indivíduos, tempo para aplicação do instrumento de medida, entre outras condições.

3.3.3 Delineamentos da equalização

Para conduzir a equalização, existem duas abordagens que podem ser escolhidas conforme o contexto da aplicação: mesma população ou grupos equivalentes (em inglês, *common population*) e grupos não equivalentes (em inglês, *nonequivalent groups*). O delineamento de mesma população parte do princípio que os indivíduos respondentes (de cada um dos instrumentos de medida) pertencem a uma mesma população, sendo considerados grupos equivalentes. Por outro lado, se essa suposição não for atendida, ainda assim é possível lidar com grupos não equivalentes, desde que existam itens comuns entre os instrumentos de medida, os quais irão servir para fazer uma conexão entre as diferentes populações (Kolen e Brennan, 2014).

É essencial que a amostra de respondentes na etapa de equalização seja representativa da população para a qual os instrumentos de medida serão administrados em situações

posteriores – conforme propriedade da invariância populacional, bem como, sob condições semelhantes de aplicação (Kolen e Brennan, 2014).

Grupos Equivalentes

São três os delineamentos para mesma população. As definições foram baseadas em Kolen e Brennan (2014) e foram mantidos os nomes em inglês, já consolidados na literatura. Nas explicações, consideraram-se dois instrumentos de medida (X e Y) a serem equalizados, no entanto, caso sejam mais de dois, basta extrapolar as mesmas instruções para os demais.

a) *Random Groups* (RG): os indivíduos são aleatoriamente distribuídos para os grupos X e Y, conforme ilustrado na Figura 6. Utilizando um método adequado para aleatorização, as diferenças resultantes entre os grupos são atribuídas às diferenças entre os instrumentos de medida. Por exemplo, supondo que dois grandes grupos uniformes, provenientes da mesma população, foram selecionados para responder aos instrumentos X e Y, se a média obtida pelo grupo que foi submetido ao instrumento X for 10 pontos acima da média do grupo que respondeu o instrumento Y, essa diferença pode ser atribuída às diferenças entre os instrumentos. As vantagens deste delineamento é que cada indivíduo precisará ser submetido a apenas um dos testes e é possível avaliar ao mesmo tempo quantos instrumentos forem de interesse, apenas acrescentando mais grupos ao estudo. No entanto, é importante garantir que o tamanho da amostra seja suficientemente grande para reduzir os erros presentes na equalização, os quais ainda serão abordados neste trabalho.

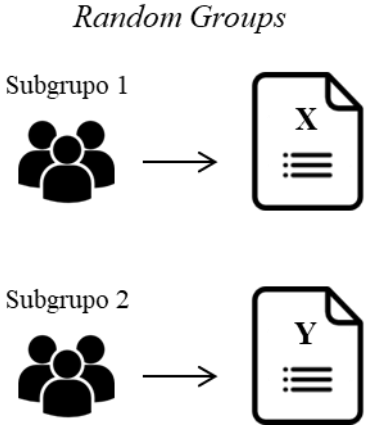

Figura 6. Delineamento *Random Groups*.

Fonte: elaborado pelo autor.

b) *Single Group* (SG): neste delineamento todos os indivíduos da amostra respondem aos dois instrumentos, conforme Figura 7. Este delineamento é interessante porque para cada respondente será obtido o escore de todos os instrumentos em avaliação, o que enriquece a

equalização, mas ao mesmo tempo é uma desvantagem, visto que nem sempre é possível submeter os indivíduos a mais de um teste. Outra consideração importante a respeito desse delineamento é que, se todos iniciarem pelo instrumento X, por exemplo, pode existir algum fator de confundimento associado à ordem de aplicação. Por isso, o balanceamento definido a seguir é uma alternativa para evitar a presença de fatores externos.

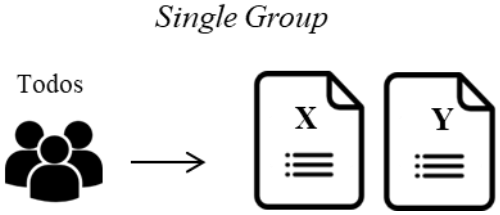

Figura 7. Delineamento *Single Group*.

Fonte: elaborado pelo autor.

c) *Single Group Design with Counterbalancing* (SG-C): todos respondem aos dois instrumentos, mas metade dos indivíduos inicia pelo instrumento X e a outra metade inicia pelo instrumento Y, conforme ilustrado na Figura 8. A definição de quais respondentes iniciarão por qual pode ser feita utilizando métodos de aleatorização. Esse delineamento foi utilizado na bateria de aptidão americana do Serviço das Forças Armadas durante os anos de 1976 a 1979, com o objetivo de introduzir um novo formulário de seleção de indivíduos para o serviço militar, em substituição a um antigo. Nos primeiros anos, os candidatos responderam aos dois testes, para poder ajustar a equalização, com o objetivo de utilizar apenas o novo formulário nos anos seguintes. No entanto, os respondentes souberam que seriam escolhidos apenas com base no escore do formulário antigo, pois estavam respondendo ao novo formulário apenas para fins de equalização. Como o questionário novo foi respondido com baixa motivação, o resultado da equalização foi equivocado e, nos anos seguintes, aproximadamente 350.000 pessoas que deveriam ter sido consideradas inelegíveis, acabaram ingressando no serviço militar americano. Esse é um exemplo de má administração do delineamento SG-C.

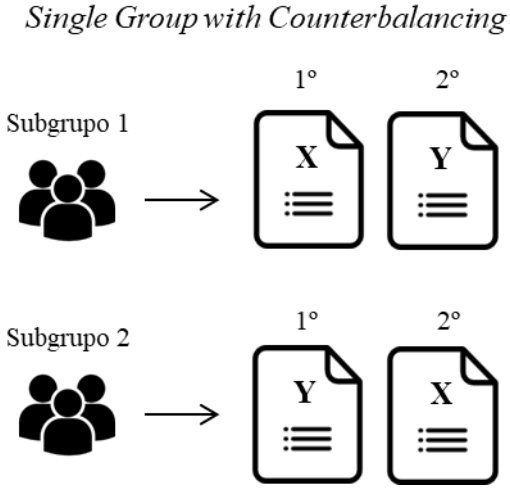

Figura 8. Delineamento *Single Group with Counterbalancing*.

Fonte: elaborado pelo autor.

Grupos Não Equivalentes

Quando os respondentes pertencem a populações diferentes, ou seja, oriundos de grupos não equivalentes, uma maneira de viabilizar a equalização é garantindo a presença de itens comuns nos instrumentos de medida que serão equalizados.

a) *Common Item Nonequivalent Groups* (CINEG): neste delineamento os instrumentos X e Y devem ser construídos com alguns itens em comum e administrados a diferentes grupos de pessoas, de acordo com a Figura 9. A quantidade de itens comuns deve ser representativa em relação ao total de itens do instrumento, não há uma regra estabelecida, porém alguns autores sugerem que 40% do total de itens sejam comuns entre os testes (Nering e Ostini, 2010). Recomenda-se que os itens comuns sejam idênticos em todos os aspectos: não deve existir diferença entre os enunciados e entre as categorias de resposta e os itens devem ocupar o mesmo lugar na ordem do teste. Os itens comuns devem ser representativos dos tipos de perguntas, por exemplo, se o instrumento for composto por itens dicotômicos e politômicos, é importante que tenham itens comuns também nessas condições. Neste delineamento, as diferenças obtidas nos resultados podem ser em razão das diferenças entre os grupos e das diferenças entre os instrumentos de medida. A questão central na equalização utilizando esse formato é separar essas diferenças, identificando quais são oriundas dos grupos e quais são as diferenças entre os instrumentos. Na Tabela 1 encontra-se um exemplo simplificado de aplicação deste delineamento. Supondo que existam dois instrumentos, X e Y, para avaliar a gravidade do paciente. Cada um desses instrumentos possui 100 itens dicotômicos, onde cada item se refere a um sintoma e as respostas possíveis são presente ou ausente. Dos 100 itens,

40 são comuns aos dois, restando 60 itens exclusivos de cada instrumento de medida. O grupo 1 respondeu ao instrumento X e o grupo 2 respondeu ao instrumento Y. As médias sugerem que a situação do grupo 2 está mais grave pois possui, em média, 24 (60%) dos sintomas descritos nos itens comuns, enquanto que o grupo 1 teve média de 20 (50%). Essa diferença de quatro sintomas a mais, em média, representa 10% no total de 40 itens comuns. A pergunta então é: qual seria o escore médio do grupo 2 caso os indivíduos tivessem sido sorteados para responder ao instrumento X? O grupo 2 obteve 10% a mais nos itens comuns, tomando como razoável esta linha de pensamento, é esperado que o grupo 2 tenha 10% a mais do que o grupo 1 no instrumento X, ou seja, 53 + 10 = 63. Logo, 63 é o provável escore médio que o grupo 1 teria obtido se fosse submetido ao instrumento Y. Esta é uma lógica simples para compreensão de como os itens comuns se relacionam, os métodos aplicados na prática serão descritos mais adiante neste capítulo. Este é um dos delineamentos mais utilizados na área da educação por se tratar de uma situação mais provável. Em 1986, os alunos que realizaram o teste NAEP (*National Assessment of Educational Progress*), que avalia o progresso dos estudantes das escolas americanas, tiveram resultados preliminares que mostraram uma redução significativa na etapa de *reading* (leitura), quando comparados aos escores dos alunos de 1984. Muitas investigações foram conduzidas até que se identificou uma potencial explicação para a queda: houve diferenças na administração dos testes entre os anos de 1984 e 1986. Os assuntos dos testes foram agrupados de forma diferente, itens comuns foram alocados em ordem diferente de um ano para o outro e o tempo disponível para completar o teste também não foi o mesmo. Este exemplo ilustra a importância de administrar ambos os instrumentos no mesmo contexto em todos os aspectos possíveis.

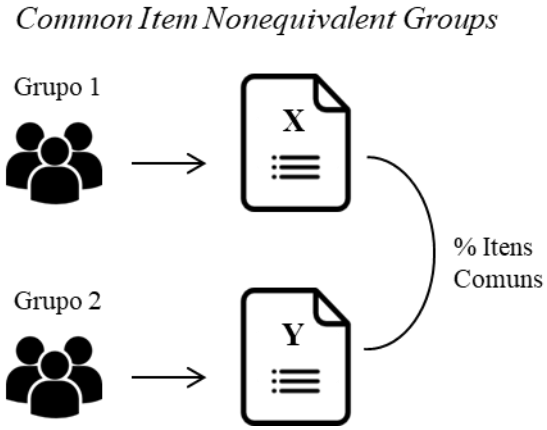

Figura 9. Delineamento *Common Item Nonequivalent Groups*.

Fonte: elaborado pelo autor.

Tabela 1. Exemplo de médias dos grupos 1 e 2 para dois instrumentos com itens comuns.

Grupo	Instrumento X (60 itens exclusivos)	Instrumento Y (60 itens exclusivos)	Itens comuns (40 itens)	Total (100 itens)
1	33 (55%)	-	20 (50%)	53 (53%)
2	-	36 (60%)	24 (60%)	60 (60%)

Fonte: adaptado de Kolen e Brennan (2014).

Dos quatro delineamentos apresentados, RG, SG, SG-C e CINEG, o mais utilizado na prática e, consequentemente, mais abordado na literatura, é o que possui itens comuns (CINEG). Com este delineamento não é preciso submeter um mesmo indivíduo a mais de um questionário e também não é necessário garantir a homogeneidade entre os grupos de respondentes, além de ser mais flexível e possivelmente mais barato (Kim e Kolen, 2005). Os autores também apontam que o delineamento RG produz estimativas mais estáveis do que se utilizado o delineamento CINEG com poucos itens comuns ou itens comuns de baixa qualidade.

3.3.4 Erros na equalização

Na estimação de relações de equalização estão presentes, geralmente, dois tipos de erro: o erro aleatório e o erro sistemático. Pode ser trabalhoso eliminá-los e por isso o objetivo é minimizá-los para que o resultado da equalização seja bem-sucedido (Kolen e Brennan, 2014).

Erro Aleatório

Quando se faz uso de amostras de respondentes para estimação dos parâmetros da equalização, não há como evitar a presença do erro aleatório, que, no contexto da equalização, é conhecido como erro padrão. Se fosse possível estimar com base em toda a população, esse erro seria zero. Logo, uma forma de controlar o erro padrão é acrescentar mais pessoas na amostra - à medida que o tamanho da amostra aumenta, o erro se torna desprezível. O erro padrão na equalização é único para cada escore e, tecnicamente, é o desvio padrão entre escores equivalentes estimados com base em repetições do processo de equalização. Os passos abaixo, propostos por Kolen e Brennan (2014), ilustram, conceitualmente, o significado do erro padrão na equalização, considerando dois instrumentos X e Y:

a) Extrair uma amostra de 1000 indivíduos de uma população de respondentes;

b) Através do método de equalização, encontrar o escore y do instrumento Y equivalente ao escore x do instrumento X usando dados desta amostra;

c) Repetir os passos (a) e (b) até que se tenha uma elevada quantidade de estimativas para o escore em Y equivalente ao escore x em X;

d) O desvio padrão dessas estimativas é uma estimativa para o erro padrão da equalização para o escore x do instrumento X.

Erro Sistemático

O erro sistemático está mais relacionado a violações nas premissas e condições para conduzir a equalização, por isso quantificá-lo pode ser uma tarefa difícil. Geralmente este erro ocorre quando o método de estimação introduz algum viés na relação de equalização. Por exemplo, no delineamento de grupos aleatórios (RG), a falta de aleatorização pode levar a um erro sistemático. No delineamento de grupo único (SG), o efeito da ordem de administração dos instrumentos pode ser considerado um erro sistemático, o que é resolvido no delineamento SG-C. No delineamento de grupos não equivalentes (CINEG), se as premissas estatísticas utilizadas para separar as diferenças entre grupos e entre instrumentos não forem atendidas, poderá configurar um erro sistemático, assim como se houver diferenças nos itens comuns entre os instrumentos (ordem e/ou enunciado diferente). Enquanto que o aumento da amostra reduz o erro padrão, o erro sistemático não é necessariamente sensibilizado pelo tamanho da amostra. Boas práticas na elaboração e no desenvolvimento da equalização, cuidados na etapa de coleta de dados, escolha adequada do delineamento e uso de técnicas estatísticas apropriadas ajudam a controlar o erro sistemático (Kolen e Brennan, 2014).

3.3.5 Métodos clássicos de equalização

Existem diferentes métodos para conduzir uma equalização, os quais foram desenhados para atingir um mesmo objetivo: a possibilidade de relacionar escores entre diferentes instrumentos de medida. As duas principais abordagens são os métodos clássicos de equalização, que integram a Teoria Clássica dos Testes, e os métodos por estimação via Teoria da Resposta ao Item, tecnicamente mais robustos. Os métodos clássicos estão detalhadas no Anexo C. Para aprofundamento do tema, consultar Kolen e Brennan (2014) e Davier (2011).

3.3.6 Métodos de equalização via TRI

Na prática, a obtenção de uma relação entre duas escalas, θ_X e θ_Y, utilizando os conceitos e recursos da TRI, pode ser feita por vários caminhos, que dependem de algumas decisões tomadas pelo pesquisador. O método de calibração é escolhido em função do delineamento definido na pesquisa e, dependendo do método de calibração utilizado, poderá ser necessário acrescentar uma etapa de transformação linear, para a qual existem diferentes técnicas. No final deste processo, ao escolher um indivíduo ao acaso, que respondeu ao instrumento X, cujo traço latente é θ_X, este valor pode ser relacionado a um θ_Y correspondente, ou seja, existirá uma relação direta entre θ_X e θ_Y mesmo que o respondente não tenha sido submetido ao instrumento Y (Nering e Ostini, 2010).

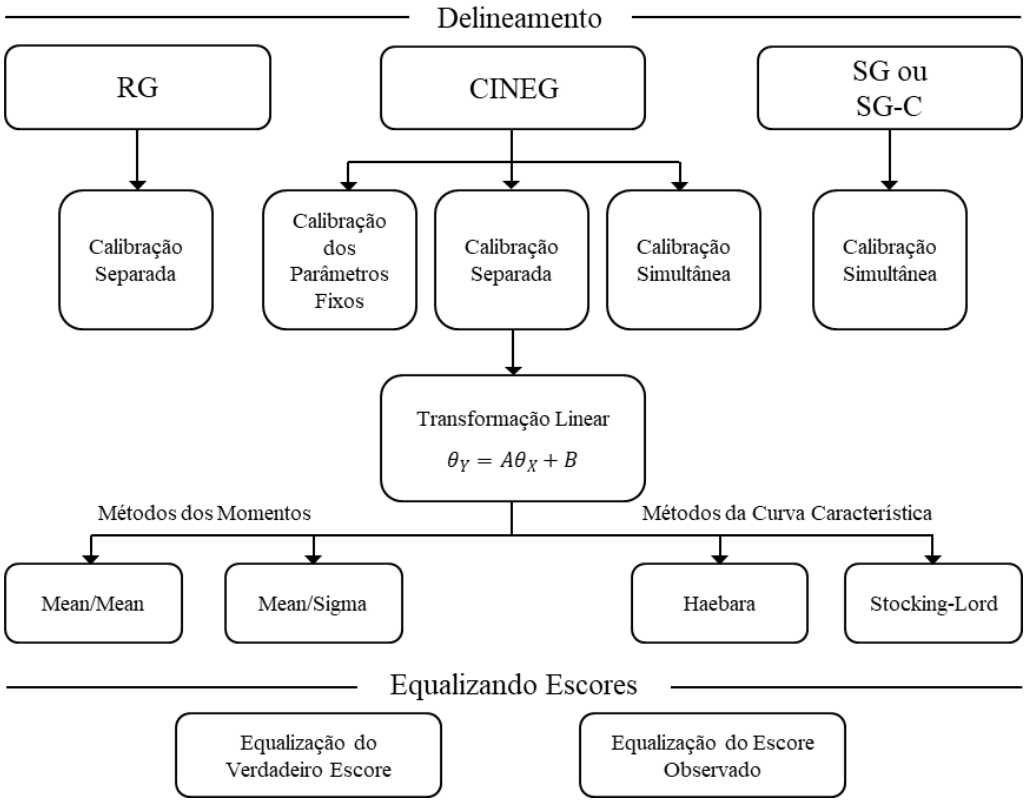

Figura 10. Diagrama das etapas de equalização utilizando a TRI.

Fonte: elaborado pelo autor.

No entanto, o resultado em termos de θ_X e θ_Y pode ser difícil de ser reportado aos pesquisadores e aos respondentes, que geralmente estão habituados a valores de escore absolutos, como na TCT (Kolen e Brennan, 2014; Nering e Ostini, 2010). Em função disto, foram desenvolvidas técnicas que relacionam, de fato, o escore absoluto no instrumento de medida X a um escore absoluto no instrumento de medida Y, mesmo que tenham sido

avaliados sob a perspectiva da TRI, facilitando a interpretação. Estes métodos são chamados de equalização do verdadeiro escore (*True Score Equating* – TSE) e equalização do escore observado (O*bserved Score Equating* - OSE) (Nering e Ostini, 2010) e serão abordados ao final desta seção.

Em virtude da quantidade de etapas da técnica, a Figura 10 foi elaborada para auxiliar na visualização do método. As etapas serão detalhadas no decorrer do texto.

3.3.6.1 Métodos de calibração

Os métodos de calibração descritos nessa seção foram baseados em Nering e Ostini (2010) e variam conforme o delineamento.

Random Groups

Os parâmetros dos itens do instrumento X podem ser estimados separadamente dos parâmetros dos itens do instrumento Y. Se utilizada a convenção escalar que pressupõe distribuição normal (média zero e desvio padrão um), não será necessária nenhuma etapa adicional pois os parâmetros já estarão na mesma escala. Isso está assegurado em virtude de que os grupos de respondentes representam uma amostra da mesma população.

Single Group/Single Group with Counterbalancing

Os parâmetros de todos os respondentes dos dois instrumentos devem ser estimados simultaneamente, logo, estarão na mesma escala.

Common Item Nonequivalent Groups

Para este delineamento, são três as formas de conduzir a calibração dos parâmetros:

a) Calibração separada: aqui, diferentemente do que foi apresentado na seção de estimação dos parâmetros, o conceito de calibração separada significa calibrar os instrumentos de medida separadamente. Como neste delineamento os grupos não são considerados equivalentes, mesmo utilizando a convenção que supõe normalidade dos traços latentes, não se tem a garantia de que os resultados estarão na mesma escala. A calibração é conduzida separadamente, a qual gera duas escalas dependentes dos grupos e, por isso, não estão ainda na mesma unidade. Após a calibração separada, será necessária uma etapa de transformação das estimativas dos parâmetros TRI da escala do instrumento X para a escala do instrumento Y, utilizando uma transformação linear para relacionar as duas escalas, a partir

dos itens comuns – alguns autores chamam essa etapa de *scale linking*. Essa relação linear entre as escalas se dá pela propriedade de invariância dos parâmetros, onde os parâmetros dos itens são invariantes entre grupos de respondentes e os parâmetros de traço latente são invariantes entre instrumentos de medida. Essa etapa de transformação linear será detalhada na sequência;

b) Calibração simultânea: uma única calibração é conduzida juntando os dados resultantes de ambos os instrumentos de medida. Nesse caso, adota-se um dos instrumentos de medida como referência de escala, por exemplo, o instrumento X. Ao final, todas as estimativas dos parâmetros estarão na escala do instrumento X;

c) Calibração de parâmetros fixos: trata-se de uma terceira abordagem que pode ser vista como um método misto entre as calibrações separada e simultânea. Nesse caso, primeiramente é conduzida uma calibração separada para o instrumento X, e em seguida, apenas os parâmetros dos itens não comuns do instrumento Y são estimados, fixando as estimativas dos itens comuns já obtidas do instrumento X. Ao utilizar deste artifício, as estimativas dos itens não comuns do instrumento Y já estarão na escala do instrumento X. Entretanto, quando os grupos forem muito diferentes entre si, essa abordagem pode levar a resultados viesados para os parâmetros dos itens, devido a diferença entre os grupos.

Em resumo, ao utilizar os delineamentos RG e SG (SG-C), a etapa seguinte de transformação não é necessária. Já para o delineamento CINEG, dependendo do método de calibração utilizado, poderá ser necessário acrescentar uma etapa de transformação linear, como no caso da calibração separada. A calibração simultânea e a calibração de parâmetros fixos resultam em um único conjunto de estimativas para os parâmetros dos itens comuns, enquanto que a calibração separada resulta em dois conjuntos diferentes de estimativas para os mesmos.

3.3.6.2 Transformação linear

A propriedade da invariância dos parâmetros esclarece que, dentro do contexto de uma transformação linear, os parâmetros dos itens não variam entre grupos de respondentes e os parâmetros dos traços latentes não variam entre instrumentos de medida. Ainda, a propriedade da invariância sugere que uma mudança de escala, através de uma transformação linear, pode ser feita desde que um conjunto de itens comuns forneça uma espécie de conexão entre as duas amostras de dados que foram obtidas a partir de diferentes instrumentos de medida (Kolen e Brennan, 2014; Nering e Ostini, 2010).

Considere, então, uma situação onde foi adotado o delineamento onde os grupos não são equivalentes (CINEG) com calibração separada. Para dar sequência ao procedimento de equalização, é necessário passar pela etapa de transformação linear. Sejam X e Y dois instrumentos de medida com i_X e i_Y itens cada um, sendo n_C os itens comuns. Sejam θ_X e θ_Y os traços latentes obtidos pela calibração separada dos dados dos instrumentos X e Y, respectivamente. Pela propriedade de invariância dos parâmetros, existe uma relação linear entre θ_X e θ_Y de forma que:

$$\theta_Y = A\theta_X + B, \tag{12}$$

onde A é o coeficiente angular e B é o intercepto.

As duas escalas, θ_X e θ_Y, são tidas como não equivalentes porque as respostas foram obtidas em grupos não equivalentes (se fossem equivalentes, então o delineamento seria RG). Teoricamente, os parâmetros de um item comum, i, entre as duas escalas, estimado em X e em Y, devem ser linearmente relacionados. Para os modelos 3PL, GRM e GPCM tem-se que:

$$a_{iY} = a_{iX}/A \tag{13}$$

e

$$b_{iY} = Ab_{iX} + B \tag{14}$$

e

$$c_{iY} = c_{iX} \tag{15}$$

onde a_{iY}, b_{iY} e c_{iY} são os parâmetros do item i na escala Y e a_{iX}, b_{iX} e c_{iX} são os parâmetros do item i na escala X. As constantes A e B devem ser estimadas com algum dos métodos que serão detalhados mais a diante.

O parâmetro c_i não é afetado porque se trata de uma medida de probabilidade, que varia entre zero e um, e independe da transformação linear. O modelo NRM deve ser tratado como um caso a parte devido às restrições impostas pelo modelo (Nering e Ostini, 2010).

Para estimar os valores de A e B existem duas classes de métodos estatísticos que vem sendo praticadas: métodos dos momentos e métodos da curva característica. Os métodos dos momentos correspondem ao método *Mean/Sigma* (Marco, 1977) e ao método *Mean/Mean* (Loyd e Hoover, 1980). Já os métodos conhecidos como da curva característica são *Stocking-Lord* (Stocking e Lord, 1983) e *Haebara* (Haebara, 1980). Os métodos dos momentos são conhecidos pela simplicidade na aplicação enquanto que os métodos da curva característica são empregados quando se busca por mais robustez e redução dos erros associados às estimativas (Kim e Lee, 2006).

Métodos dos Momentos

Os métodos dos momentos buscam encontrar constantes escalares combinando os dois conjuntos de estimativas dos parâmetros obtidas a partir das calibrações separadas para os itens comuns entre os instrumentos de medida X e Y. O resultado independe da direção da transformação, seja de Y para X ou de X para Y, os métodos levam a soluções simétricas.

Esta é a forma mais direta e simples para transformar escalas no contexto do delineamento CINEG. O raciocínio por trás dos métodos dos momentos é expressar as equações apresentadas em (13) e (14) para um item, em termos de um grupo de itens. A partir das equações (13) e (14), tem-se que:

$$A = \frac{\mu(a_X)}{\mu(a_Y)} \tag{16}$$

e

$$A = \frac{\sigma(b_Y)}{\sigma(b_X)} \tag{17}$$

e

$$B = \mu(b_Y) - A\mu(b_X) \tag{18}$$

onde $\mu(\cdot)$ é a média, $\sigma(\cdot)$ é o desvio padrão, a_X e a_Y representam os parâmetros de discriminação dos itens comuns e b_X e b_Y representam os parâmetros de posição dos itens comuns (Nering e Ostini, 2010).

Método *Mean/Sigma* (MS)

A partir das equações apresentadas em (17) e (18) e fazendo uso da média e do desvio padrão das estimativas dos parâmetros de discriminação e de posição do item, obtêm-se os coeficientes:

$$\hat{A}_{MS} = \frac{\sigma(\hat{b}_Y)}{\sigma(\hat{b}_X)} \tag{19}$$

e

$$\hat{B}_{MS} = \mu(\hat{b}_Y) - \hat{A}_{MS}\mu(\hat{b}_X) \tag{20}$$

onde $\hat{A}_{MS}$ e $\hat{B}_{MS}$ representam a estimativas pelo método *Mean/Sigma*.

Método *Mean/Mean* (MM)

Assim como o método *Mean/Sigma* utiliza a média e desvio padrão, intuitivamente, o método *Mean/Mean* utiliza apenas as médias para estimar os parâmetros A e B, conforme as seguintes equações:

$$\hat{A}_{MM} = \frac{\mu(\hat{a}_X)}{\mu(\hat{a}_Y)} \tag{21}$$

e

$$\hat{B}_{MM} = \mu(\hat{b}_Y) - \hat{A}_{MM}\mu(\hat{b}_X) \tag{22}$$

onde $\hat{A}_{MM}$ e $\hat{B}_{MM}$ representam a estimativas pelo método *Mean/Mean*.

Métodos da Curva Característica

Em alguns casos, os métodos dos momentos podem não trazer as melhores estimativas, principalmente quando os itens comuns incluírem itens calibrados através do NRM, devido às restrições impostas pelo modelo (Nering e Ostini, 2010). Outro problema se dá quando as estimativas dos parâmetros são muito diferentes mas as curvas estimadas são praticamente idênticas. Por exemplo, em itens com estimativas divergentes para o parâmetro b mas que geraram CCIs semelhantes, o método *Mean/Sigma* será mais influenciado pelas diferenças nas estimativas do que pelas semelhanças nas curvas. Isso ocorre porque os métodos dos momentos não consideram todas as estimativas para os parâmetros dos itens simultaneamente, conforme as equações apresentadas em (17) e (18) (Kolen e Brennan, 2014). Em virtude dessas limitações, alguns autores propuseram métodos que consideram as estimativas dos parâmetros conjuntamente, que são os métodos da curva característica.

Estes métodos são baseados na combinação das curvas de categoria de resposta (Método *Haebara*) ou na combinação das curvas características do teste (Método *Stocking-Lord*). Utilizando dados amostrais, a combinação nem sempre é exata, para isso, o grau de correspondência é avaliado por critérios que comparam as diferenças entre as curvas estimadas dos itens comuns (Kim e Kolen, 2005). As estimativas para A e B devem minimizar as diferenças entre as curvas. O método proposto por Haebara é simétrico (não importando se a transformação for feita de X para Y ou de Y para X), enquanto que o método de Stocking-Lord originalmente não é, mas já existe uma versão simétrica formulada por Kim e Lee (2006).

A lógica por trás dos métodos de curva característica parte da existência de uma correspondência exata entre as duas curvas dos itens comuns, cada uma expressa pelas escalas de θ_X e θ_Y. Dada a relação linear entre θ_X e θ_Y apresentada na equação (12), que também pode ser escrita em termos de θ_Y, a curva característica de uma categoria k de um item comum i de uma escala deve ser expressa na outra escala, utilizando as estimativas dos parâmetros transformados de uma escala para outra. Sejam $P_{ikX}(\theta_X)$ e $P_{ikY}(\theta_Y)$ as curvas de

categoria de resposta k do item comum i, expressas em θ_X e θ_Y, respectivamente, com seus parâmetros originais. Sejam $P^*_{ikX}(\theta_Y)$ e $P^{\#}_{ikY}(\theta_X)$ as curvas de categoria de resposta transformadas, expressas em θ_X e θ_Y com os parâmetros transformados para a outra escala (Nering e Ostini, 2010).

Logo, uma relação perfeita significa que, para cada item comum i:

$$P_{ikY}(\theta_Y) = P^*_{ikX}(\theta_Y) \tag{23}$$

e

$$P_{ikX}(\theta_X) = P^{\#}_{ikY}(\theta_X) \tag{24}$$

Tal relação pode ser escrita também em função da CCT:

$$T_Y(\theta_Y) = T^*_X(\theta_Y) \tag{25}$$

e

$$T_X(\theta_X) = T^{\#}_Y(\theta_X) \tag{26}$$

conforme apresentado na equação (7).

Dificilmente a igualdade entre as CCRs das equações (23) e (24) será atingida, o mesmo vale para as CCTs das equações (25) e (26). Dessa forma, com dados amostrais, os valores de A e B devem ser estimados de forma a otimizar as relações. Uma forma de lidar com esse problema é definir um critério para medir as diferenças entre os termos da igualdade e buscar minimizá-lo, que é a proposta dos métodos Haebara e Stocking-Lord (Kim e Kolen, 2005).

Método *Haebara*

A função de critério proposta por Haebara (1980) para expressar a diferença entre as CCRs é a soma dos quadrados das diferenças entre as CCRs para cada item e para cada valor de θ (Kolen e Brennan, 2014). Para obter as estimativas de A e B deve ser minimizada a função (Nering e Ostini, 2010):

$$Q(A,B) = Q_1(A,B) + Q_2(A,B) \tag{28}$$

onde

$$Q_1(A,B) = \sum_{g=1}^{G_Y} \left\{ \sum_{i=1}^{n_C} \sum_{k=1}^{m_i} \left[\hat{P}_{ikY}(\theta_{gY}) - \hat{P}^*_{ikX}(\theta_{gY}) \right]^2 \right\} W_1(\theta_{gY}) \tag{29}$$

e

$$Q_2(A,B) = \sum_{g=1}^{G_X} \left\{ \sum_{i=1}^{n_C} \sum_{k=1}^{m_i} \left[\hat{P}_{ikX}(\theta_{gX}) - \hat{P}^{\#}_{ikY}(\theta_{gX}) \right]^2 \right\} W_2(\theta_{gX}) \tag{30}$$

onde θ_{gY} e $W_1(\theta_{gY})$ representam os pontos de quadratura e seus respectivos pesos para a escala θ_Y, assim como θ_{gX} e $W_2(\theta_{gX})$ para a escala θ_X.

Nas equações 29 e 30, as curvas de categoria de resposta, $\hat{P}(.)$, $\hat{P}^*(.)$ e $\hat{P}^{\#}(.)$ são estimativas com base na amostra de respondentes.

O algoritmo de Newton-Raphson pode ser utilizado para minimizar a equação (28), já implementado computacionalmente na maioria nos softwares que fazem equalização de escalas.

Método *Stocking-Lord*

O método proposto por Stocking e Lord (1983) tem como objetivo combinar as CCTs de θ_X e θ_Y. A versão simétrica do método que minimiza os valores de A e B é dada pelo critério (Kim e Lee, 2006):

$$F(A,B) = F_1(A,B) + F_2(A,B) \tag{31}$$

onde

$$F_1(A,B) = \sum_{g=1}^{G_Y} \left[\hat{T}_Y(\theta_{gY}) - \hat{T}^*{}_X(\theta_{gY})\right]^2 W_1(\theta_{gY}) \tag{32}$$

e

$$F_2(A,B) = \sum_{g=1}^{G_X} \left[\hat{T}_X(\theta_{gX}) - \hat{T}^{\#}{}_Y(\theta_{gX})\right]^2 W_2(\theta_{gX}) \tag{33}$$

onde θ_{gY} e $W_1(\theta_{gY})$ representam os pontos de quadratura e seus respectivos pesos para a escala θ_Y, assim como θ_{gX} e $W_2(\theta_{gX})$ para a escala θ_X.

Nas equações 32 e 33, as curvas características do teste, $\hat{T}(.)$, $\hat{T}^*(.)$ e $\hat{T}^{\#}(.)$ são estimadas com base na amostra de respondentes.

Existem estudos comparando a eficiência dos quatro métodos para obtenção dos coeficientes de transformação linear A e B. Em um estudo com dados simulados, os autores Hanson e Béguin (2002) obtiveram melhores resultados com os métodos de curva característica. A mesma conclusão foi feita por Kim et al. (2006) em um estudo de simulação para comparar os quatro métodos, sob o delineamento RG e testes de formato misto. Os resultados obtidos através dos métodos da curva característica apresentaram erros menores, sendo o método Haebara o que obteve os resultados mais acurados, embora os autores ressaltem que os resultados podem não ser os mesmos quando utilizados dados reais (Hanson e Béguin, 2002; Kim e Lee, 2006).

Ao fim da etapa de transformação linear, o pesquisador irá obter os valores de A e B que resolvem a relação linear da equação (12). Ou seja, é possível traçar uma relação entre θ_X e θ_Y utilizando os dois coeficientes. Dessa forma, escolhendo, ao acaso, um indivíduo que respondeu ao instrumento X cuja medida de traço latente é θ_X, o valor de θ_Y pode ser obtido facilmente através da equação de transformação linear. Se utilizada a convenção escalar, tanto os valores de θ_X quanto os de θ_Y estarão contidos, na grande maioria, no intervalo (-3,3).

Conforme já exposto anteriormente, o resultado em termos de θ_X e θ_Y pode ser difícil de reportar, pois geralmente as pessoas esperam valores positivos e coerentes com o instrumento de medida. Pensando nisso, os métodos chamados de equalização do verdadeiro escore e equalização do escore observado foram desenvolvidos (Nering e Ostini, 2010).

3.3.6.3 Equalização do verdadeiro escore

Popularmente conhecido por TSE (*True Score Equating*), a ideia principal deste método é que um escore x do instrumento de medida X associado com um valor θ seja considerado equivalente a um escore y obtido através do instrumento de medida Y, associado ao mesmo valor de θ. Essa abordagem necessita que todos os parâmetros dos itens estejam na mesma escala θ, só assim será possível encontrar o escore do instrumento Y equivalente ao escore do instrumento X (Nering e Ostini, 2010).

Para obter a equalização do verdadeiro escore, primeiro é preciso compreender a relação existente entre o escore verdadeiro total e a curva característica do teste. Considerando dois instrumentos de medida de formato misto, X e Y, com I itens cada e sejam Z_{iX} e Z_{iY} os escores obtidos para uma resposta do item i para os instrumentos X e Y, respectivamente, dentre todas as opções possíveis, U_{ikX} e U_{ikY}. Conforme a Teoria Clássica dos Testes, o escore verdadeiro total V_X do questionário X é a esperança do escore total observado X, que é a soma do escore obtido em cada item i, definido pela equação $X = \sum_{i=1}^{I} Z_{iX}$. O mesmo vale para Y, logo, similarmente, temos o escore verdadeiro total V_Y e o escore total observado $Y = \sum_{i=1}^{I} Z_{iY}$. De acordo com Kolen e Brennan (2014), a esperança do escore observado X dado um nível de traço latente θ é equivalente a CCT, que remete a equação apresentada em (7), isto é:

$$E(X|\theta) = E\left(\sum_{i=1}^{I} Z_{iX}\,|\theta\right) = T_X(\theta) = \sum_{i=1}^{I}\sum_{k=1}^{m_i} U_{ikX}P_{ikX}(\theta) \tag{34}$$

onde U_{ikX} é uma função de escore ordenada para a categoria k do item i do instrumento X, ou seja, todos os valores de categoria de resposta disponíveis, conforme a coluna "Categoria de Resposta" dos Quadros 1 e 2.

Similarmente, para o instrumento Y, essa relação é:

$$E(Y|\theta) = E\left(\sum_{i=1}^{I} Z_{iY}\,|\theta\right) = T_Y(\theta) = \sum_{i=1}^{I}\sum_{k=1}^{m_i} U_{ikY}P_{ikY}(\theta) \tag{35}$$

Como a CCT tem por definição ser uma função monótona de θ, exclusivamente crescente e definida no intervalo $-\infty < \theta < +\infty$, os escores verdadeiros v_X e v_Y, dos instrumentos de medida X e Y, respectivamente, associam-se com um valor de θ dentro do seguinte intervalo:

$$\sum_{i=1}^{I}(U_{i1X} + \delta_i c_{iX}) < v_X < \sum_{i=1}^{I} U_{im_iX} \tag{36}$$

e

$$\sum_{i=1}^{I}(U_{i1Y} + \delta_i c_{iY}) < v_Y < \sum_{i=1}^{I} U_{im_iY} \tag{37}$$

onde $\delta_i = 1$ se o item i for um item do modelo 3PL, do contrário, $\delta_i = 0$.

Quando utilizado o modelo 3PL o valor mínimo de escore verdadeiro que pode ser obtido é a soma de todos os c_i, considerando que os U_{ik} iniciem pelo zero (Kolen e Brennan, 2014).

Considerando itens calibrados pelo modelo GPCM, por exemplo, o escore verdadeiro v_X deve ser maior do que a soma de todos os escores mínimos possíveis de cada item e deve ser menor do que a soma de todos os escores máximos possíveis de cada item. O mesmo vale para v_Y.

Estabelecida a relação apresentada nas equações (34) e (35), inicia-se o processo de equalização do verdadeiro escore. Para um dado valor de θ existem verdadeiros escores $V_X(\theta)$ e $V_Y(\theta)$ que são considerados equivalentes. Seja $e_Y(V_X(\theta))$ o escore verdadeiro da escala Y equivalente a um verdadeiro escore da escala X associado a um θ. Ou seja:

$$e_Y\big(V_X(\theta)\big) = V_Y(\theta) \tag{38}$$

Pela propriedade da equidade da equação (11) e seja V_X^{-1} a função inversa de V_X, a função de equalização para o escore verdadeiro é:

$$e_Y(v_X) = V_Y(v_X^{-1}) \tag{39}$$

onde $v_X^{-1} = \theta$.

A equação descrita em (39) implica que a equalização do escore verdadeiro, conforme definido por Kolen e Brennan (2014), é um processo de três etapas:

a) Especificar um escore verdadeiro v_X do instrumento X;

b) Encontrar o valor de $\theta = v_X^{-1}$ que corresponda a este escore;

c) Encontrar o verdadeiro escore equivalente no questionário Y, $V_Y(\theta)$, que corresponda ao valor de θ.

Geralmente os escores definidos no primeiro passo são os valores inteiros possíveis de obter na escala X, dentro da amplitude definida. O segundo passo requer o uso de técnicas iterativas, como o método Newton-Raphson, que se trata de um processo de sucessivas etapas iterativas cujo objetivo final é encontrar raízes de funções não lineares. Este e outros métodos numéricos já foram implementados nas ferramentas computacionais que conduzem a equalização do verdadeiro escore. Após encontrar o valor θ, o terceiro passo é direto, mas vale lembrar que é bastante provável que, para um valor inteiro na escala de X, se obtenha um valor não inteiro para Y.

A Figura 11 ilustra um resultado fictício de equalização de escore verdadeiro através de duas curvas características do teste hipotéticas dos instrumentos de medida X e Y, $T_X(\theta)$ e $T_Y(\theta)$. Por exemplo, o verdadeiro escore 26 no instrumento X corresponde a um valor de θ igual a 1,8, que equivale ao valor de 28 para a escala do instrumento Y.

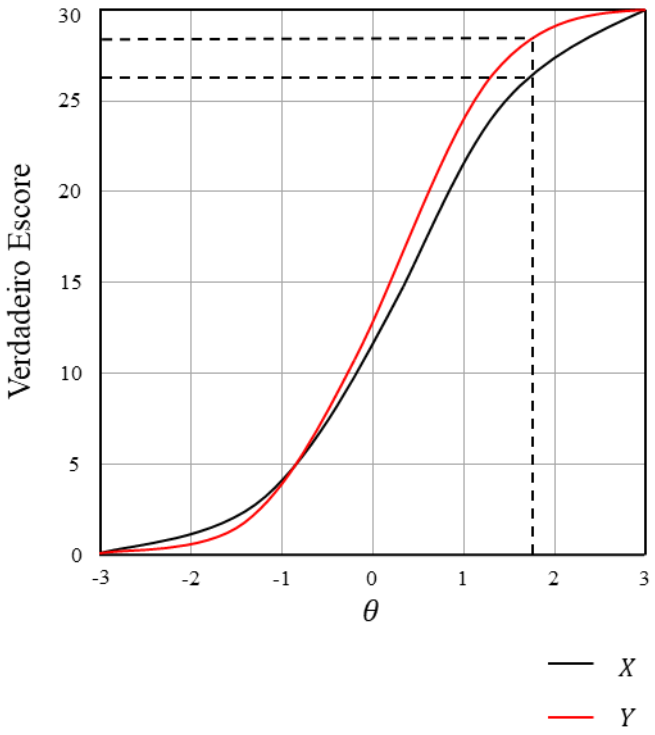

Figura 11. Curva característica do verdadeiro escore, comparando $\boldsymbol{T_X(\theta)}$ e $\boldsymbol{T_Y(\theta)}$.

Fonte: adaptado de Nering e Ostini (2010).

A aplicação do método da equalização do verdadeiro escore supõe que os parâmetros dos itens e os verdadeiros escores são conhecidos, logo, as curvas característica estão definidas. Porém, na prática, dificilmente os parâmetros dos itens e os verdadeiros escores estarão disponíveis. Embora sejam utilizadas as estimativas dos parâmetros e os escores observados, essa substituição pode levar a erros na equalização (Nering e Ostini, 2010). Além disso, não existem razões teóricas comprovadas de que o escore observado possa ser tratado como um verdadeiro escore (Kolen e Brennan, 2014). Assim, um procedimento adicional será necessário para converter os escores observados para além da amplitude de escores verdadeiros possíveis, conforme definido nos intervalos em (36) e (37). De acordo com Kolen (1981), os seguintes passos auxiliam na resolução deste problema:

a) Estabelecer o valor mínimo do escore observado do questionário X, $\min_X = \sum_{i=1}^{I} U_{i1X}$ igual ao mínimo escore observado do questionário Y, $\min_Y = \sum_{i=1}^{I} U_{i1Y}$;

b) Estabelecer o limite inferior de escores verdadeiros do questionário X, $\inf_X \sum_{i=1}^{I}(U_{i1X} + \delta_i c_{iX})$, igual ao limite inferior de escores verdadeiros do questionário Y, $\inf_Y \sum_{i=1}^{I}(U_{i1Y} + \delta_i c_{iY})$;

c) Através de interpolação linear, encontrar equivalência entre os escores mínimos e os limites inferiores;

d) Estabelecer o escore máximo observado do questionário X, $\max_X = \sum_{i=1}^{I} U_{im_iX}$, igual ao escore máximo observado do questionário Y, $\max_Y = \sum_{i=1}^{I} U_{im_iY}$.

Na ausência de itens calibrados pelo modelo 3PL, o segundo e o terceiro passo podem ser ignorados. Seja v_X^o um escore fora do intervalo de verdadeiros escores mas dentro dos valores possíveis de escore observado. A interpolação linear é dada por (Nering e Ostini, 2010):

$$e_Y(v_X^o) = \frac{\inf_Y - \min_Y}{\inf_X - \min_X}(v_X^o - \min_X) + \min_Y \tag{40}$$

Na prática, para a equalização do verdadeiro escore, as estimativas para os parâmetros dos itens são utilizadas para estabelecer uma relação estimada de verdadeiro escore. Então a conversão estimada dos verdadeiros escores é aplicada aos escores observados.

3.3.6.4 Equalização do escore observado

Este método é uma alternativa ao método da equalização do escore verdadeiro. A equalização do escore observado utiliza as estimativas dos parâmetros dos itens, obtidas no modelo TRI, para calcular a distribuição marginal de probabilidade dos escores observados

em cada instrumento, X e Y, os quais são equalizados utilizando o procedimento do equipercentil. O método do percentil é também um dos métodos clássicos de equalização e está descrito brevemente no Anexo C (Nering e Ostini, 2010). Nos softwares, a equalização do escore observado é conhecido por OSE (*Observed Score Equating*) e está detalhada no Anexo D.

Comparando OSE e TSE, a equalização do verdadeiro escore é computacionalmente mais simples e a conversão não depende da distribuição de θ, como na equalização do escore observado (Kolen e Brennan, 2014). No entanto, na TSE existe o problema teórico de que os verdadeiros escores, na prática, não são conhecidos. Existem evidências de que, para o delineamento RG, os métodos produzem resultados bastante diferentes (Kolen, 1981), sendo que a equalização utilizando o OSE obtém melhores resultados, enquanto que o método TSE apresenta resultados mais estáveis ao longo da escala. Utilizando o delineamento CINEG foram obtidos resultados muito semelhantes entre os métodos TSE e OSE (Lord e Wingersky, 1984). Além disso, o método do escore observado pode levar a resultados fora dos limites possíveis da escala, o que precisa ser avaliado do ponto de vista do pesquisador.

3.3.6.5 Aspectos computacionais

O conjunto de softwares disponibilizados pela *Scientific Software Internacional*, PARSCALE, STUIRT e POLYEQUATE, quando utilizados juntos, compreendem todos os passos da equalização via TRI, desde a escolha do delineamento e dos métodos de calibração, até a transformação linear e os métodos TSE e OSE. Recomenda-se utilizar os softwares em conjunto porque dentre os arquivos de *input* do POLYEQUATE estão as estimativas para os parâmetros dos itens e os pontos de quadratura, com os respectivos pesos, os quais podem ser encontrados no *output* do PARSCALE.

Uma alternativa em software livre são os pacotes disponíveis no R. No entanto, devido às subdivisões da técnica, cada pacote especializou-se em algumas situações. Por exemplo, o pacote *equate* (Albano, 2016) só lida com métodos clássicos de equalização, enquanto que o pacote *SNSequate* (González, 2017) possibilita utilizar os métodos da TCT como também a equalização via TRI, mas somente para testes inteiramente dicotômicos. O pacote *equateIRT* (Battauz, 2015) também só avalia itens utilizando os modelos 1PL, 2PL e 3PL. O pacote *kequate* (Andersson et al., 2013) é especializado nos métodos de suavização de Kernel. O pacote *lordif* (Choi, 2016) performa equalização via TRI para delineamento CINEG e

computa a transformação linear somente do método *Stocking-Lord*, além disso, só permite a utilização dos modelos GPCM e GRM. Para o delineamento de itens comuns, o pacote *plink* (Weeks, 2010) é o mais completo porque permite a utilização de testes de formato, faz análises unidimensionais e multidimensionais, computa todos os métodos de transformação linear e também TSE e OSE.

O Quadro 4 apresenta um resumo dos pacotes do R disponíveis para equalização adequados para cada situação.

Quadro 4. Pacotes do R para equalização.

Pacote	Equalização via TCT			Equalização via TRI			Dimensionalidade		Transformação Linear				Modelos		TSE	OSE	Outros Métodos	
	RG	SG SG-C	CINEG	RG	SG SG-C	CINEG	Uni	Multi	MM	MS	HB	SL	Dicotômicos	Politômicos			Kernel	Bayesianos
equate	X	X	X															
equateIRT				X	X	X	X		X	X	X	X	X		X	X		
kequate																	X	
lordif						X*	X					X		X**				
plink						X*	X	X	X	X	X	X	X	X	X	X		
SNSequate	X	X	X	X	X	X	X		X	X	X	X	X		X	X	X	X

*Somente o método de calibração separada.

**Somente os modelos GPCM e GRM.

Fonte: elaborado pelo autor.

4. OBJETIVOS

4.1 JUSTIFICATIVA

Ao utilizarem-se diferentes escalas para avaliação de pacientes, existe uma dificuldade em comparar as informações obtidas nas publicações científicas e criar hipóteses sobre as diferenças entre os achados. No contexto da SCA3/MJD, é interessante que se estabeleça uma equivalência entre a NESSCA e a SARA, a fim de torná-las comparáveis. Além disso, trata-se de um trabalho inovador dado o potencial da técnica, a qual é amplamente conhecida na avaliação educacional mas poucas são as aplicações na área da saúde.

Vale lembrar que a NESSCA já foi avaliada pela perspectiva da TRI (Maciel, 2013). Na ocasião, foram identificados os sintomas que ajudam a discriminar melhor o comprometimento pela doença e foram propostas algumas alterações na escala, como a exclusão dos itens Câimbra e Vertigem e o agrupamento de categorias de resposta para alguns itens. Além disso, não se tem conhecimento da avaliação da escala SARA sob a perspectiva da TRI, o que permitirá avaliar os itens quanto as suas contribuições para o traço latente.

4.2 OBJETIVOS

Objetivo Geral

Este trabalho tem por objetivo apresentar e discutir o método de equalização de escalas utilizando a abordagem da Teoria da Resposta ao Item e explorar sua utilização no contexto da área da saúde.

Objetivos Específicos

a) Demonstrar a utilização do método através da equalização das escalas NESSCA e SARA para avaliação do comprometimento pela doença de Machado-Joseph.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Albano AD. **equate**: An *R* Package for Observed-Score Linking and Equating. Journal of Statistical Software. 2016;74(8).

Andersson B, Branberg K, Wiberg M. Performing the Kernel Method of Test Equating with the package {kequate}. Journal of Statistical Software. 2013;55(6):1–25.

de Andrade DF, Tavares HR, Valle R da C. Teoria da Resposta ao Item: Conceitos e Aplicações. São Paulo, SP: ABE; 2000.

Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978;43(4):561–73.

de Araujo EAC, de Andrade DF, Bortolotti SLV. Teoria da resposta ao item. Revista da Escola de Enfermagem da USP. 2009;43(spe):1000–1008.

Battauz M. **equateIRT**: An *R* Package for IRT Test Equating. Journal of Statistical Software. 2015;68(7).

Bettencourt C, Santos C, Kay T, Vasconcelos J, Lima M. Analysis of segregation patterns in Machado–Joseph disease pedigrees. Journal of Human Genetics. 2008;53(10):920–3.

Bird TD. Hereditary Ataxia Overview. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJ, Stephens K, et al., organizadores. GeneReviews®. Seattle (WA): University of Washington, Seattle; 1993.

Bock R. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika. 1972;37(1):29–51.

Boomsma A, van Duijn MAJ, Snijders TAB. Essays on Item Response Theory. Groningen, Netherlands; 2000.

Castro SMJ. Teoria da Resposta ao Item: Aplicação na Avaliação da Intensidade de Sintomas Depressivos. [Porto Alegre]: Universidade Federal do Rio Grande do Sul; 2008.

Castro SMJ, Trentini C, Riboldi J. Item response theory applied to the Beck Depression Inventory. Revista Brasileira de Epidemiologia. 2010;13(3):487–501.

Chalmers RP. **mirt**: A Multidimensional Item Response Theory Package for the *R* Environment. Journal of Statistical Software. 2012;48(6).

Chen W-H, Revicki DA, Lai J-S, Cook KF, Amtmann D. Linking pain items from two studies onto a common scale using item response theory. J Pain Symptom Manage. 2009;38(4):615–28.

Choi SW. **lordif**: Logistic Ordinal Regression Differential Item Functioning using IRT. 2016;

Coluci MZO, Alexandre NMC, Milani D. Construção de instrumentos de medida na área da saúde. Ciência & Saúde Coletiva. 2015;20(3):925–36.

Coutinho P, Calheiros JM, de Andrade C. Sobre uma nova doença degenerativa do sistema nervoso central transmitida de modo autossómico dominante e afectando familiares originários dos Açores: nota prévia. 1977.

Datta D. **blandr**: A Bland-Altman Method Comparison package for *R*. 2017; Recuperado de: https://github.com/deepankardatta/blandr

Davier AA von. Statistical models for test equating, scaling, and linking. New York: Springer; 2011.

Donis KC. História natural da ataxia espinocerebelar tipo 3/Doença de Machado-Joseph com início na infância. [Porto Alegre]: Universidade Federal do Rio Grande do Sul; 2015.

Donis KC, Saute JAM, Krum-Santos AC, Furtado GV, Mattos EP, Saraiva-Pereira ML, et al. Spinocerebellar ataxia type 3/Machado-Joseph disease starting before adolescence. neurogenetics. 2016;17(2):107–13.

Goldstein H. Test Equating. British Journal of Mathematical and Statistical Psychology. 1984;37(1):131–4.

González J. **SNSequate**: Standard and Nonstandard Statistical Models and Methods for Test Equating. Journal of Statistical Software. 2017;

González J, Wiberg M. Applying Test Equating Methods Using R. Cham: Springer International Publishing; 2017.

Haebara T. Equating Logistic Ability Scales by a Weighted Least Squares Method. Japanese Psychological Research. 1980;22(3):144–9.

Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park, Calif: Sage Publications; 1991.

Hanson BA, Béguin AA. Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in the Common-Item Equating Design. Applied Psychological Measurement. 2002;26(1):3–24.

Hattie J. Methodology Review: Assessing Unidimensionality of Tests and ltems. Applied Psychological Measurement. 1985;9(2):139–64.

Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38(9 Suppl):II28-42.

Jardim LB. Aspectos clínicos e moleculares da doença de Machado-Joseph no Rio Grande do Sul: Sua relação com as outras ataxias espinocerebelares autossômicas dominantes e uma hipótese sobre seus fatores modificadores. 2000 [citado 14 de outubro de 2017]; Recuperado de: http://www.lume.ufrgs.br/handle/10183/164174

Jardim LB, Hauser L, Kieling C, Saute JAM, Xavier R, Rieder CRM, et al. Progression Rate of Neurological Deficits in a 10-Year Cohort of SCA3 Patients. The Cerebellum. 2010;9(3):419–28.

Kieling C, Rieder CRM, Silva ACF, Saute JAM, Cecchin CR, Monte TL, et al. A neurological examination score for the assessment of spinocerebellar ataxia 3 (SCA3). European journal of neurology. 2008;15(4):371–376.

Kim S, Kolen MJ. Methods for Obtaining a Common Scale Under Unidimensional IRT Models: A Technical Review and Further Extensions. Iowa: University of Iowa; 2005.

Kim S, Lee W-C. IRT Scale Linking Methods for Mixed-Format Tests. Iowa: ACT; 2004.

Kim S, Lee W-C. An Extension of Four IRT Linking Methods for Mixed-Format Tests. Journal of Educational Measurement. 2006;43(1):53–76.

Kolen MJ. Comparison of Traditional and Item Response Theory Methods for Equating Tests. Journal of Educational Measurement. 1981;18(1):1–11.

Kolen MJ, Brennan RL. Test equating, scaling, and linking: methods and practices. Third Edition. New York: Springer; 2014.

Kotz S, Balakrishnan N, Read CB, Vidakovic B, organizadores. Encyclopedia of statistical sciences. 2nd ed. Hoboken, N.J: Wiley-Interscience; 2006.

van der Linden WJ, Hambleton RK. Handbook of modern item response theory. New York: Springer; 1996.

Lord FM. A Theory of Test Scores. ETS Research Bulletin Series. 1952;1952(1):i–126.

Lord FM. Applications of item response theory to practical testing problems. 1980.

Lord FM, Novick MR, Birnbaum A. Statistical theories of mental test scores. Charlotte, NC: Information Age Publ; 1968.

Lord FM, Wingersky MS. Comparison of IRT True-Score and Equipercentile Observed-Score "Equatings". Applied Psychological Measurement. 1984;8(4):453–61.

Loyd BH, Hoover HD. Vertical Equating Using the Rasch Model. Journal of Educational Measurement. 1980;17(3):179–93.

Maciel TH. Aplicação da Teoria da Resposta ao Item ao escore NESSCA de avaliação da progressão da Doença de Machado Joseph. [Porto Alegre]: Universidade Federal do Rio Grande do Sul; 2013.

Marco GL. Item Characteristic Curve Solutions to Three Intractable Testing Problems. Journal of Educational Measurement. 1977;14(2):139–60.

Masters GN. A rasch model for partial credit scoring. Psychometrika. 1982;47(2):149–74.

McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. Med Care. 2000;38(9 Suppl):II43-59.

Mislevy RJ. Bayes modal estimation in item response models. Psychometrika. 1986;51(2):177–95.

Moreira Jr. FJ. Aplicações da teoria da resposta ao item (TRI) no Brasil. Revista Brasileira de Biometria. 2010;28(4):137–70.

Muraki E. A Generalized Partial Credit Model: Application of an EM Algorithm. Applied Psychological Measurement. 1992;16(2):159–76.

Muraki E, Bock R. PARSCALE. Chicago: Scientific Software International; 2003.

Nering ML, Ostini R. Handbook of polytomous item response theory models. New York, NY: Routledge; 2010.

Oliveira CM, Reckziegel ER, Augustin MC, Rocha AG, Bolzan G, Santos JA, et al. Causal factors behind early- and late-onset Machado-Joseph disease patients do not interfere with the rate of neurological deterioration. Pisa; 2017. Recuperado de: http://www.iarc2017.com/wp-content/uploads/2017/09/IARC-Abstract-Book.pdf

Pasquali L. Teoria da medida. L. Pasquali, Psicometria: Teoria dos testes na Psicologia e na Educação. 2003;1:23–51.

R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Recuperado de: https://www.R-project.org/

Raiche G. **nFactors**: An *R* package for parallel analysis and non graphical solutions to the Cattell scree test. 2010; Recuperado de: http://CRAN.R-project.org/package=nFactors

Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: Mesa; 1960.

Reckase MD. The Past and Future of Multidimensional Item Response Theory. Applied Psychological Measurement. 1997;21(1):25–36.

Rizopoulos D. **ltm** : An R package for Latent Variable Modelling and Item Response Theory Analyses. Journal of Statistical Software. 2006;17(5):1–25.

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Mongraph. 1969;17.

Saute JAM, de Castilhos RM, Monte TL, Schumacher-Schuh AF, Donis KC, D'Ávila R, et al. A randomized, phase 2 clinical trial of lithium carbonate in Machado-Joseph disease: Lithium Trial in Machado-Joseph Disease. Movement Disorders. 2014;29(4):568–73.

Saute JAM, Jardim LB. Machado Joseph disease: clinical and genetic aspects, and current treatment. Expert Opinion on Orphan Drugs. 2015;3(5):517–535.

Saute JAM, da Silva ACF, Souza GN, Russo AD, Donis KC, Vedolin L, et al. Body Mass Index is Inversely Correlated with the Expanded CAG Repeat Length in SCA3/MJD Patients. The Cerebellum. 2012;11(3):771–4.

Schmitz-Hübsch T, du Montcel ST, Baliko L, Berciano J, Boesch S, Depondt C, et al. Scale for the assessment and rating of ataxia: development of a new clinical scale. Neurology. 13 de junho de 2006a;66(11):1717–20.

Schmitz-Hübsch T, du Montcel ST, Baliko L, Berciano J, Boesch S, Depondt C, et al. Scale for the assessment and rating of ataxia: development of a new clinical scale. Neurology. 2006b;66(11):1717–20.

Souza GN, Kersting N, Krum-Santos AC, Santos ASP, Furtado GV, Pacheco D, et al. Spinocerebellar ataxia type 3/Machado-Joseph disease: segregation patterns and factors influencing instability of expanded CAG transmissions: Segregation patterns and CAGexp transmissions in SCA3/MJD. Clinical Genetics. agosto de 2016;90(2):134–40.

Stocking ML, Lord FM. Developing a Common Metric in Item Response Theory. Applied Psychological Measurement. 1983;7(2):201–10.

Thissen D, Chen W, Bock R. MULTILOG. Chicago: Scientific Software Internacional; 2003.

Trouillas P, Takayanagi T, Hallett M, Currier RD, Subramony SH, Wessel K, et al. International Cooperative Ataxia Rating Scale for pharmacological assessment of the cerebellar syndrome. The Ataxia Neuropharmacology Committee of the World Federation of Neurology. J. Neurol. Sci. 12 de fevereiro de 1997;145(2):205–11.

Weeks JP. **plink**: An R package for Linking Mixed-Format Tests Using IRT-Based Methods. Journal of Statistical Software. 2010;35(12):1–33.

Zimowski M, Muraki E, Mislevy R, Bock R. BILOG-MG. Chicago: Scientific Software Internacional; 2003.

6. ARTIGO

Equalização das escalas NESSCA e SARA utilizando a Teoria da Resposta ao Item na avaliação do comprometimento pela doença de Machado-Joseph

Nicole Machado Utpott[1], Vanessa Bielefeldt Leotti[1,2], Laura Bannach Jardim[3,4]

[1] Programa de Pós-Graduação em Epidemiologia, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil

[2] Departamento de Estatística, Instituto de Matemática e Estatística, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil

[3] Programa de Pós-Graduação em Ciências Médicas, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul, Brasil

[4] Serviço de Genética Médica, Hospital de Clínicas de Porto Alegre, Brasil

RESUMO

Contexto: A equalização de escalas é uma técnica estatística utilizada para estabelecer relações de equivalência entre diferentes escalas. Sua utilização é bastante popular na avaliação educacional, porém, incomum na área da saúde, onde escalas de medidas são ferramentas integrantes da prática clínica. Com a utilização de diferentes escalas, existe uma dificuldade em comparar resultados científicos, como é o caso das escalas NESSCA e SARA, instrumentos de avaliação do comprometimento pela doença de Machado-Joseph (SCA3/MJD). **Objetivo:** Explorar o método de equalização de escalas e demonstrar sua aplicação através das escalas NESSCA e SARA, utilizando a abordagem da Teoria da Resposta ao Item (TRI) na avaliação do comprometimento pela SCA3/MJD. **Métodos:** Os dados são de 227 pacientes do Hospital de Clínicas de Porto Alegre portadores da SCA3/MJD que possuem medidas completas para NESSCA e/ou SARA. O delineamento de equalização utilizado é o de grupos não equivalentes com itens comuns, com calibração separada. O modelo TRI utilizado na estimação dos parâmetros foi o de crédito parcial generalizado, para NESSCA e SARA. Foi feita a transformação linear através dos métodos *Mean/Mean*, *Mean/Sigma*, *Haebara* e *Stoking-Lord* e foi aplicado o método da equalização do verdadeiro escore para obter uma relação estimada entre os escores das escalas. **Resultados:** O escore NESSCA estimado pela equalização via escore SARA comparado com o escore NESSCA observado apresentou diferença mediana de 0,82 pontos pelo método *Mean/Sigma*. Este foi o melhor método de transformação linear dentre os testados. **Conclusões:** Com este estudo foi

possível explorar a aplicabilidade da técnica de equalização via TRI no contexto da saúde e ilustrar sua utilização criando uma relação de equivalência entre os escores das escalas NESSCA e SARA.

Palavras-Chave: Equalização de escalas; Teoria da Resposta ao Item; doença de Machado-Joseph; NESSCA; SARA.

ABSTRACT

Background: Scale equating is a statistical technique used to establish equivalence relations between different scales. Its use is quite popular in educational evaluation, however, unusual in the health area, where scales of measures are tools that integrate clinical practice. With the use of different scales, there is a difficulty in comparing scientific results, such as NESSCA and SARA scales, tools for assessing the commitment to Machado-Joseph disease (SCA3/MJD). **Objective:** Explore the method of scale equating and demonstrate its application through NESSCA and SARA scales, using the Item Response Theory (IRT) approach in assessing SCA3/MJD commitment. **Methods:** Data came from 227 patients from the Hospital de Clínicas de Porto Alegre with SCA3/MJD who have complete measures for NESSCA and/or SARA scales. The equating design used is that of non-equivalent groups with common items, with separate calibration. The IRT model used in the estimation of the parameters was the generalized partial credit, for NESSCA and SARA. The linear transformation was performed using the Mean/Mean, Mean/Sigma, Haebara and Stoking-Lord methods and the equation of the true score was applied to obtain an estimated relationship between the scores of the scales. **Results:** Difference between NESSCA score estimated by SARA and observed NESSCA score has shown median of 0.82 points, by Mean/Sigma method. This was the best method of linear transformation among the tested. **Conclusions:** This study extended the use of scale equating under IRT approach to health outcomes and established an equivalence relationship between NESSCA and SARA scores, making the comparison between patients and scientific results feasible.

Key Words: Scale equating; Item Response Theory; Machado-Joseph disease; NESSCA; SARA.

INTRODUÇÃO

A equalização de escalas é uma técnica estatística utilizada para estabelecer relações de equivalência entre diferentes escalas (1). Embora a área da saúde seja uma das áreas de

conhecimento que mais desenvolvem e consomem questionários e testes, a técnica de equalização ainda é pouco explorada neste contexto (2). Sua base teórica é fundamentada na psicometria e sua utilização é majoritariamente voltada para a avaliação educacional (2,3). Na área da saúde, as escalas de medidas são ferramentas integrantes da prática clínica e da avaliação do estado de saúde do paciente, pois possuem grande influência sobre as decisões que se referem ao tratamento e que envolvem a gestão de políticas públicas direcionadas para o cuidado com a saúde da população. O avanço em pesquisas e as necessidades cada vez mais específicas dos pacientes contribuem para o surgimento de muitas ferramentas de medida (4). Nesse sentido, a equalização torna possível a criação de uma relação entre diferentes escalas, possibilitando a comparação entre resultados de publicações científicas.

Tradicionalmente, as escalas de medidas são analisadas de acordo com os princípios da Teoria Clássica dos Testes (TCT), que se baseia na média ou na soma dos escores obtido nos itens. A Teoria da Resposta ao Item (TRI), tecnicamente mais robusta, foi desenvolvida para suprir as principais limitações da TCT, como a ausência de discriminação entre itens, ou seja, respondentes com o mesmo escore total são considerados iguais mesmo que o conjunto de respostas tenha sido totalmente diferente (1,5). A TRI tem por objetivo descrever a associação entre a probabilidade de uma resposta a um item em particular e o nível de um respondente quanto a uma característica de interesse que não pode ser observada diretamente, conhecida por traço latente (5). O estado de saúde de um paciente é uma variável que, na estatística, pode ser classificada como um traço latente. Apesar de não poder ser medido diretamente, o traço latente pode ser inferido com base na observação de variáveis secundárias que estejam relacionadas a essa característica de interesse e que possam ser mensuradas através de instrumentos de medidas (6). Os modelos TRI variam conforme a natureza do item (dicotômicos ou politômicos), o número de populações envolvidas e a quantidade de traços latentes avaliados (1). A escolha do modelo deve ser adequada às escalas.

Um exemplo de patologia que utiliza escalas para mensurar o estado de saúde do paciente é a doença de Machado-Joseph. Também conhecida como ataxia espinocerebelar tipo 3 (SCA3/MJD), é um distúrbio neurodegenerativo autossômico dominante caracterizado por uma ataxia cerebelar de início geralmente na idade adulta (7). Apesar de não existir cura para a doença, existem algumas opções para administrar os sintomas (8). Para isso, a avaliação do nível de comprometimento neurológico dos pacientes é essencial. Em 2006, foi publicada a *Scale for Assessment and Rating of Ataxia* (SARA) (9) que, por sua simplicidade, tem sido bastante utilizada nas publicações científicas. A escala

SARA é composta por 8 itens, cada um tem de cinco a nove categorias de resposta, cujo escore total varia de 0 a 40 pontos. Em 2008 foi publicada a *Neurological Examination escore for Spinocerebellar Ataxia* (NESSCA) (10), que diferencia-se da SARA por avaliar também sintomas não-atáxicos. A NESSCA é dividida em 18 itens, cada um possui de duas a cinco categoriais de resposta, somando, no total 40 pontos. Tradicionalmente, as escalas SARA e NESSCA estimam o traço latente considerando a soma do escore obtido em cada item, que é o pressuposto da TCT.

No contexto da SCA3/MJD, a TRI modela a relação existente entre a probabilidade de um paciente apresentar um sintoma e o nível do comprometimento da SCA3/MJD. Em 2013 a NESSCA foi avaliada através dos modelos da TRI com dados de 106 pacientes. Os autores identificaram os sintomas que ajudam a explicar melhor o comprometimento com a doença e propuseram alterações na escala, como a exclusão dos itens Câimbra e Vertigem, bem como o agrupamento de categorias de resposta para alguns itens (11). Não se tem conhecimento de estudos avaliando a SARA sob a perspectiva da TRI. Em geral, as escalas NESSCA e SARA são correlacionadas mas, apesar de ambas somarem os mesmos 40 pontos, não se sabe como comparar cada pontuação individualmente.

O objetivo principal deste artigo é apresentar o método de equalização de escalas e explorar sua utilização com a abordagem da TRI no contexto da saúde. Para ilustrar a utilização do método será construída uma relação entre os instrumentos de medida para avaliação do nível de comprometimento pela doença de Machado-Joseph, SARA e NESSCA, estabelecendo escores equivalentes entre as escalas.

MÉTODOS

Fontes de Dados[*]

Diversas fontes de dados foram utilizadas, referentes a estudos procedidos com pacientes diagnosticados com SCA3/MJD do Hospital de Clínicas de Porto Alegre (HCPA):

a) Avaliações da NESSCA de 156 pacientes de um estudo de história natural descrito por Jardim et al. (12);

[*] Os estudos que originaram os dados para este trabalho foram aprovados pelo comitê de ética do Hospital de Clínicas de Porto Alegre: (a) GPPG-HCPA-02194; (b) GPPG-HCPA-09418; (c) GPPG-HCPA-13-0303; (d) GPPG-HCPA-06-0613; (e) GPPG-HCPA-14-0625.

b) Avaliações da NESSCA e SARA basais de 60 pacientes de um ensaio clínico randomizado de Saute et al. (13);

c) Avaliações de NESSCA e SARA de 35 pacientes de um estudo descrito por Oliveira et al. (14);

d) Avaliações de NESSCA e SARA de 24 pacientes do estudo de Saute et al. (15);

e) Avaliações de NESSCA e SARA de 8 pacientes com início na infância descritos por Donis et al. (16).

Os dados foram divididos em dois grupos, o Grupo 1 compreende os pacientes que possuem apenas avaliações da NESSCA, estudo (a), e o Grupo 2 é composto pelos pacientes restantes, estudos (b), (c), (d), e (e), que possuem avaliações para NESSCA e SARA. No caso de estudos prospectivos, com mais de uma medida das escalas (estudos (a), (b), (c) e (e)), foi utilizada a mais antiga dentre as que estavam completas, sendo descartadas as restantes. Pacientes com ausência de informação para algum item foram excluídos da análise, sendo 53 da fonte de dados (a) e um da fonte (e). Dada a importância do sintoma de marcha para a doença, comum às escalas NESSCA e SARA, os únicos dois pacientes que apresentaram ausência desse sintoma em pelo menos uma das duas escalas foram excluídos, sendo um da fonte (c) e um da fonte (e). Por fim, a amostra foi composta por 103 pacientes do Grupo 1 e 124 do Grupo 2. Outras informações como idade, gênero, duração da doença no momento da avaliação e número de repetições do CAG expandido também foram coletadas.

Delineamento de Equalização

A equalização pode ser feita a partir de grupos equivalentes ou de grupos não equivalentes. Os delineamentos de grupos equivalentes (em inglês, *Random Groups* - RG e *Single Groups* - SG) partem do princípio de que todos os indivíduos respondentes pertencem a uma mesma população. Por outro lado, se essa proposição não for atendida, ainda assim é possível lidar com grupos não equivalentes (em inglês, *Common Item Nonequivalent Groups* - CINEG*)*, desde que existam itens comuns entre os instrumentos de medida, os quais irão servir para fazer uma conexão entre as diferentes populações (17). No contexto da SCA3/MJD, o delineamento mais adequado para a equalização das escalas foi o de grupos não equivalentes com itens comuns (CINEG). Os itens comuns devem ser representativos e não há uma regra estabelecida a respeito da quantidade, mas alguns autores recomendam que 40% do total de itens sejam comuns entre os testes (18). Além disso, foram utilizados apenas os itens da NESSCA cerebelar, ou seja, foram descartados os itens que avaliam sintomas não cerebelares da NESSCA, de forma a garantir a unidimensionalidade das escalas (11). A

SARA e a NESSCA cerebelar compartilham de dois itens muito semelhantes que avaliam os sintomas Marcha (SARA)/Ataxia de Marcha (NESSCA) e Coordenação da Fala (SARA)/Disartria (NESSCA). A quantidade de categorias foi adaptada e encontra-se detalhada no Anexo E, Quadros 1 e 2. Com estas alterações, o escore máximo da SARA foi reduzido para 34 pontos ao invés dos 40 pontos originais (redução de quatro pontos no item Marcha e dois pontos no item Coordenação da Fala). Além disso, outro ajuste feito na SARA foi nos itens que avaliam os lados direito e esquerdo do corpo, separadamente, permitindo pontuações não inteiras para a média dos dois lados. Os valores 0,5; 1,5; 2,5 e 3,5 foram arredondados para 1; 2; 3 e 4, respectivamente, do contrário, na etapa computacional, tais pontuações não inteiras configurariam mais categorias de resposta. A NESSCA cerebelar passou a somar de 0 a 15 pontos devido a exclusão dos itens não cerebelares.

Análise Estatística

A primeira etapa da análise estatística teve por objetivo a calibração dos parâmetros dos itens, para isso, existem três métodos de calibração que podem ser empregados no delineamento CINEG: calibração dos parâmetros fixos, simultânea ou separada. Para este exercício foi escolhida a calibração separada por ser o método mais utilizado e que abrange mais etapas do processo (17,18). Este método consiste em calibrar separadamente os dados de cada instrumento de medida, ou seja, os parâmetros dos itens da NESSCA cerebelar foram estimados separadamente dos parâmetros dos itens da SARA. Para isso, foram utilizadas as avaliações da NESSCA cerebelar do Grupo 1 e as avaliações da SARA do Grupo 2 – descartando momentaneamente as avaliações da NESSCA cerebelar do Grupo 2. Os modelos TRI adequados para modelar os dados das escalas são o GRM (em inglês, *Graded Response Model)* e o GPCM (em inglês, *Generalized Partial Credit Model),* pois são modelos apropriados para dados politômicos possuem categorias ordenadas, não necessariamente na mesma quantidade. As fórmulas dos modelos encontram-se no Anexo A do artigo. Para a equalização as duas escalas devem ser ajustadas pelo mesmo modelo, o mais adequado foi escolhido através dos critérios de seleção AIC e BIC.

A utilização do GRM e do GPCM requer que duas suposições sejam satisfeitas: a independência local (tomando como base o nível do traço latente, os itens não devem ser correlacionados uns com os outros) e unidimensionalidade (somente um traço latente está sendo avaliado). Se a suposição de unidimensionalidade estiver atendida, então a independência local também estará satisfeita (19,20). A unidimensionalidade foi avaliada através da análise fatorial – autores sugerem que o primeiro fator deve explicar no mínimo

20% da variância total (21). A contribuição dos itens foi avaliada observando as estimativas dos parâmetros, a Curva Característica do Item (CCI), que descreve a relação ente a probabilidade e o traço latente, e a Curva de Informação do Item (CII), que permite avaliar a discriminação dos itens em modelos politômicos (5).

O resultado da calibração separada são dois traços latentes estimados e parâmetros dependentes dos grupos e, por isso, não estão na mesma métrica. Logo, é necessária uma etapa de transformação linear, a qual permite criar uma relação entre as escalas através dos parâmetros dos itens comuns (17,18). Conforme definições encontradas em Kolen e Brennan (17), pela propriedade de invariância dos parâmetros dos modelos TRI, existe uma relação linear de forma que:

$$\theta_S = A\theta_N + B, \tag{1}$$

onde A é o coeficiente angular, B é o intercepto e θ_S e θ_N são os traços latentes estimados para cada um dos instrumentos de medida, SARA e NESSCA cerebelar, respectivamente.

Para os modelos GRM e GPCM as estimativas dos parâmetros dos itens comuns se relacionam da seguinte forma (18):

$$a_{iS} = a_{iN}/A \tag{2}$$

e

$$b_{iS} = Ab_{iN} + B \tag{3}$$

e

$$c_{iS} = c_{iN} \tag{4}$$

onde a_{iS}, b_{iS} e c_{iS} são os parâmetros dos itens para o item i na escala SARA e a_{iN}, b_{iN} e c_{iN} são os parâmetros dos itens para o item i na escala NESSCA cerebelar.

Existem diferentes métodos de estimação das constantes A e B, os principais são os métodos dos momentos (*Mean/Mean* e *Mean/Sigma*), conhecidos pela simplicidade, e os métodos de curva característica (*Haebara* e *Stocking-Lord*), que são empregados quando se busca por mais robustez (22). Existem estudos com dados simulados comparando a eficiência dos quatro métodos, onde os métodos da curva característica obtiveram, no geral, melhor desempenho, embora os autores ressaltem que os resultados podem não ser os mesmos para dados reais (22,23). Assim, os quatro métodos foram apurados para comparação das estimativas A e B que relacionam θ_S e θ_N.

No entanto, o resultado em termos do traço latente pode ser de difícil compreensão, geralmente o público está habituado a valores de escore absolutos, como na TCT (17,18). Em função disto, foram desenvolvidas duas técnicas que podem relacionar, de fato, um escore

absoluto na NESSCA a um escore absoluto na SARA, mesmo que tenham sido avaliadas sob a perspectiva da TRI, facilitando a interpretação. Estes métodos são conhecidos por equalização do verdadeiro escore (em inglês, *True Score Equating* - TSE) e equalização do escore observado (em inglês, *Observed Score Equating* – OSE), cujo resultado é uma relação direta que associa um escore na SARA a um escore equivalente na NESSCA, em termos de θ (18). Autores obtiveram resultados semelhantes para os métodos OSE e TSE quando utilizado o delineamento CINEG (24), embora o OSE possa levar a resultados fora dos limites da escala. Neste exercício foi abordado o método TSE.

Em um segundo momento, o objetivo foi avaliar o resultado do procedimento do TSE sob a perspectiva dos diferentes métodos de transformação linear. Os indivíduos do Grupo 2, com medidas reais para as duas escalas, possibilitaram essa avaliação. Utilizando como base o escore SARA observado e, através da tabela resultante da TSE, foi encontrado o escore NESSCA equivalente, viabilizando a comparação com o escore NESSCA observado. Os métodos de transformação linear foram avaliados utilizando medidas descritivas (média, quartis, mínimo e máximo), bem como a análise gráfica de *Bland-Altman*.

Aspectos Computacionais

As análises foram realizadas no software R versão 3.4.4 (25). A análise dos fatores foi feita utilizando o pacote *nFactors* (26), os parâmetros foram calibrados com funções do pacote *ltm* (27) e a transformação linear e a equalização do verdadeiro escore foram feitas utilizando o pacote *plink* (28). Os gráficos de *Bland-Altman* foram feitos com o auxílio do pacote *blandr* (29). No Anexo D do artigo constam algumas das etapas mais relevantes da sintaxe.

RESULTADOS

Características dos pacientes estão apresentadas na Tabela 1. Os grupos são homogêneos em termos das características clínicas inerentes a doença. Após as modificações da etapa descrita nos métodos, o intervalo de escore possível para a NESSCA cerebelar passou a ser (0,15) e, para a SARA ajustada, (0,34).

O resultado da análise fatorial sugere que a suposição de unidimensionalidade está atendida (21) por que há um fator preponderante que explica 59,6% e 48,1% da variância total para os dados da SARA e NESSCA cerebelar, respectivamente.

Tabela 1. Características da amostra.

Características	**Grupo 1** **n = 103**	**Grupo 2** **n = 124**
Gênero*		
Feminino	51 (49,5)	69 (55,6)
Masculino	51 (49,5)	55 (44,4)
Dados faltantes	1 (1,0)	
Características clínicas†		
Idade no início da doença, em anos	34,4 (9,9) (7 - 57)	33,4 (11,8) (5 - 65)
Duração da doença no momento da avaliação, em anos	9,5 (5,7) (0 - 29)	8,3 (4,7) (1 - 26)
CAG expandido, em número de repetições	74,2 (2,4) (69 - 79)	75,3 (3,6) (68 - 91)
Escalas de medida†		
NESSCA cerebelar	7,4 (2,7) (1 - 14)	7,0 (2,4) (2 - 15)
SARA ajustada	-	12,8 (6,1) (3 - 34)

*** Avaliados em valores absolutos (percentual).**

† **Avaliados em média (desvio padrão) (mínimo - máximo).**

Foi escolhido o modelo GPCM para ajuste da NESSCA cerebelar e da SARA, em virtude dos resultados que minimizam os valores de AIC e BIC – o resultado detalhado está no Anexo B do artigo. A Tabela 2 apresenta o resultado da calibração dos parâmetros dos itens da escala NESSCA cerebelar pelo GPCM - as CCIs e as CIIs encontram-se no Anexo E do artigo, Figuras 1 a 10. Avaliando as estimativas dos parâmetros, em conjunto com as CIIs, podem ser considerados itens com maior poder de discriminação: Ataxia de Marcha, Ataxia nos Membros, Disartria e Disfagia. A Tabela 3, por sua vez, apresenta as estimativas dos parâmetros para os itens da SARA ajustada pelo modelo GPCM – as CCIs e as CIIs estão no Anexo F do artigo, Figuras 11 a 27. No geral, todos os itens da SARA apresentaram bom poder de discriminação. Estão destacados na cor cinza, nas Tabelas 2 e 3, os itens comuns entre as escalas. Nessa etapa, ainda que as estimativas dos parâmetros sejam semelhantes para os itens comuns, os mesmos ainda são dependentes da amostra de respondentes.

Tabela 2. Estimativas para os parâmetros dos itens da NESSCA cerebelar.

Item		a	b_1	b_2	b_3	b_4
1	Ataxia de Marcha	4,530	-1,567	0,192	1,131	
2	Ataxia nos Membros	0,741	-2,533	-0,534	1,036	
3	Nistagmo	0,529	-1,841	4,960		
4	Disartria	1,796	-1,427	0,889	1,802	1,779
5	Disfagia	1,623	-0,955	0,857		

Tabela 3. Estimativas para os parâmetros dos itens da SARA ajustada.

Item		a	b_1	b_2	b_3	b_4	b_5	b_6
1	Marcha	4,976	-1,259	0,623	2,155			
2	Equilíbrio de Pé	1,750	-1,406	-1,049	0,833	1,754	0,837	1,784
3	Equilíbrio Sentado	1,959	0,801	1,634	2,084	2,437		
4	Coordenação da Fala	1,477	-1,765	0,699	1,769	2,395		
5	Teste de Perseguição do Dedo	1,246	-2,893	-0,063	1,815	2,246		
6	Teste Dedo-Nariz	1,244	-0,827	0,903	2,292	2,136		
7	Diadococinesia	0,983	-1,217	0,454	-0,282	3,225		
8	Teste Calcanhar-Joelho-Canela	1,049	-2,868	-0,107	0,461	1,924		

Após a calibração dos parâmetros dos itens, foi feita a transformação linear para que os traços latentes, θ_N e θ_S, fiquem na mesma escala. Os resultados com as estimativas para as constantes A e B dos quatro métodos avaliados – *Mean/Mean*, *Mean/Sigma*, *Haebara* e *Stocking-Lord* – estão no Anexo G do artigo e foram estimados com base nas equações (4) e (5). A partir deste resultado, foi calculada a relação de verdadeiro escore sob a perspectiva dos quatro métodos. Para fins de ilustração, será utilizado o resultado do método *Mean/Sigma*, o qual está apresentado na Tabela 4 para duas situações: partindo do escore SARA para obter o escore NESSCA equivalente e partindo do escore NESSCA para obter o escore SARA equivalente - a relação é simétrica. Os valores de θ_N e θ_S se relacionam conforme equação (3) utilizando as constantes A e B estimadas.

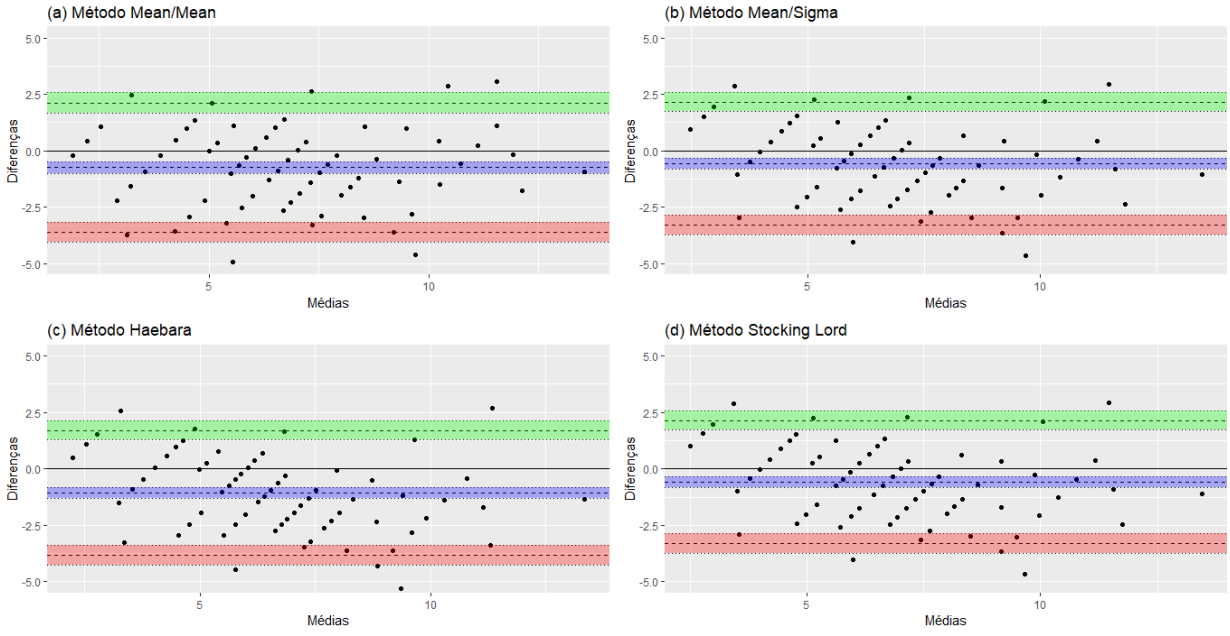

Figura 1. Gráficos *Bland-Altman* das diferenças entre o escore NESSCA estimado pela SARA e o escore NESSCA observado: (a) Método *Mean/Mean*; (b) Método *Mean/Sigma*; (c) Método *Haebara*; (d) Método *Stocking-Lord*.

Tabela 4. Equalização do verdadeiro escore utilizando o método *Mean/Sigma*[†]

TSE: SARA para NESSCA		
Escore SARA observado	**Escore NESSCA estimado pela SARA**	θ_S
0	0,000	-61,37
1	0,829	-3,00
2	1,332	-2,30
3	2,097	-1,81
4	3,053	-1,49
5	3,595	-1,29
6	3,976	-1,11
7	4,384	-0,88
8	4,812	-0,60
9	5,177	-0,33
10	5,484	-0,09
11	5,813	0,13
12	6,242	0,32
13	6,704	0,48
14	7,081	0,62
15	7,395	0,75
16	7,696	0,89
17	8,003	1,03
18	8,310	1,15
19	8,631	1,27
20	8,993	1,38
21	9,410	1,49
22	9,858	1,61
23	10,272	1,74
24	10,617	1,86
25	10,899	1,97
26	11,140	2,08
27	11,360	2,18
28	11,582	2,29
29	11,827	2,42
30	12,107	2,59
31	12,429	2,82
32	12,794	3,24

[†] Devido a pequena quantidade ou a ausência de pacientes para os escores mais altos, só foi possível estimar a relação até o escore 32 para a SARA e até o escore 13 para a NESSCA.

Com esta relação de "de-para" estabelecida na Tabela 4 e as avaliações da NESSCA do Grupo 2 (as quais não foram utilizadas na etapa de calibração), foi possível avaliar a acurácia de cada um dos métodos. Utilizando o resultado presente nas duas primeiras colunas da Tabela 4 e os dados dos pacientes do Grupo 2, foi calculada a diferença entre o escore NESSCA estimado pela SARA via equalização, e o escore NESSCA observado para o método *Mean/Sigma*. O mesmo foi realizado para os demais métodos. A Figura 1 ilustra, através dos gráficos de *Bland-Altman*, a diferença entre o escore NESSCA estimado pela SARA e o escore NESSCA observado, para os quatro métodos de transformação linear.

Na Tabela 5 estão as medidas descritivas das diferenças apresentadas entre o escore NESSCA estimado pela SARA e o escore NESSCA observado para a amostra de 124 pacientes do Grupo 2. Pela análise gráfica e descritiva, os métodos *Mean/Sigma* e *Stocking-Lord* apresentaram melhores resultados, a mediana das diferenças foi de 0,78 e 0,81, respectivamente. O método *Mean/Sigma* foi escolhido para ilustrar a relação de equivalência entre as escalas, conforme Tabela 4, por ter média e mediana das diferenças mais próximas de zero.

Tabela 5. Medidas descritivas das diferenças entre o escore NESSCA estimado pela SARA e o escore NESSCA observado.

Método de transformação	Mínimo	Q1	Mediana	Média	Q3	Máximo
Mean/Mean	0,061	0,401	0,939	1,240	1,783	4,779
Mean/Sigma	0,023	0,366	0,823	1,144	1,740	4,604
Haebara	0,022	0,581	1,248	1,465	2,116	5,392
Stocking-Lord	0,006	0,378	0,869	1,173	1,828	4,679

DISCUSSÃO

A equalização de escalas, no contexto da área da saúde, é uma técnica com muito potencial e ainda pouco explorada (2,3). Através da equalização é possível comparar resultados de publicações científicas que utilizam diferentes escalas de avaliação, para qualquer segmento da saúde que faça uso de escalas de medida. Neste trabalho, foi possível explorar os conceitos e as etapas da equalização de escalas trazendo para o contexto da área da saúde ao utilizar, de forma ilustrativa, dados de pacientes portadores da doença de Machado-Joseph. Como resultado da aplicação da técnica, foi possível estabelecer uma sugestão de relação direta entre a NESSCA cerebelar e a SARA. Com o método de transformação linear *Mean/Sigma* metade das estimativas tiveram erro abaixo de 0,82 pontos no escore NESSCA cerebelar estimado pela SARA em comparação com o escore NESSCA observado nos pacientes. Além disso, no nosso conhecimento, a avaliação da SARA pela perspectiva da TRI é inédita, e nessa análise as estimativas para os parâmetros de todos os itens mostraram-se, no geral, bastante consistentes e os itens são representativos do comprometimento pela doença, corroborando com a consistência do instrumento (9).

O delineamento de grupos não equivalentes, CINEG, é um dos mais utilizados na prática (17,18). Isso ocorre devido a flexibilidade e possível redução nos custos de aplicação, pois nem sempre é possível submeter os pacientes a todas as escalas, como acontece em outros delineamentos (17,18). Para estabelecer uma relação de equalização, é recomendado que os testes sejam construídos com este propósito, e no caso do CINEG, atendendo a requisitos a respeito da estrutura dos itens comuns, como a ordem em que os itens aparecem e a semelhança dos enunciados – tais recomendações evitam a presença de erros sistemáticos (17). Outro erro que pode estar presente na equalização é erro aleatório, este pode ser reduzido a medida que se aumenta o tamanho da amostra, até que se torne desprezível. Amostras pequenas, abaixo de 100 indivíduos, podem levar a erros altos (17). Além disso, a quantidade de itens comuns deve ser representativa em relação ao total de itens do

instrumento, não há uma regra estabelecida, porém alguns autores sugerem que 40% do total de itens sejam comuns entre os testes (18).

Na área da saúde, satisfazer a todas estas recomendações pode ser uma tarefa bastante desafiadora, visto que na maioria dos casos os instrumentos de medida foram construídos sem o propósito de equalizar, e, consequentemente, não possuem itens idênticos e que aparecem na mesma ordem em ambos os testes. A técnica exige uma série de cuidados e é fácil entender porque sua aplicação é frequente na avaliação educacional, com literatura majoritariamente oriunda das ciências sociais e comportamentais, e tão incomum na área da saúde (17,18). No entanto, aqui é importante ressaltar que, na educação, por exemplo, a prova anual do SAEB (Sistema de Avaliação da Educação Básica), tem por finalidade realizar um diagnóstico da educação básica brasileira, buscando avaliar os alunos em uma mesma escala de conhecimento, mesmo que sejam submetidos a diferentes cadernos de prova (1). Nesse caso, a equalização tem papel fundamental na avaliação dos alunos, por isso é extremamente importante que todas as condições de construção e aplicação dos testes sejam atendidas, com o objetivo de obter uma equalização bem-sucedida. Enquanto que, na área da saúde, a possibilidade de comparar indivíduos aferidos por diferentes escalas pode ser um fator complementar na avaliação do paciente, que, no caso da SCA3/MJD, poderá auxiliar na administração dos tratamentos para contornar as manifestações clínicas com base em publicações científicas, proporcionando melhor qualidade de vida aos pacientes.

Existem limitações nos resultados da equalização, visto que a SARA e a NESSCA não foram construídas sob a mesma perspectiva e não possuíam itens idênticos, por isso foram necessárias adaptações nas categorias, de forma a forçar a presença de itens comuns. Ainda que, para a NESSCA cerebelar, a recomendação de itens comuns tenha sido atingida (40% de itens comuns), no caso da SARA, apenas 25% dos itens eram comuns, o que constitui uma possível limitação deste resultado. Para estudos futuros, uma alternativa para aumentar a quantidade de itens comuns seria relacionar o item Ataxia nos Membros da NESSCA cerebelar com uma combinação dos itens Teste de Perseguição do Dedo, Teste Dedo-nariz e Diadococinesia da SARA.

Vale lembrar que, na prática, dois pacientes com o mesmo escore SARA podem ter escores NESSCA diferentes, isso ocorre devido à variabilidade fenotípica da doença e aos itens não comuns. A relação obtida só seria perfeita em todos os casos se os instrumentos de medida equalizados fossem idênticos (17,18). Uma das propriedades da equalização, em inglês, *first-order equity property*, esclarece que é esperado que os pacientes obtivessem o mesmo escore equalizado caso fossem submetidos a outra escala, em média, mas não exige

que as distribuições de probabilidade condicionais dos escores sejam iguais entre as escalas (18). Assim, é importante enxergar a equalização como uma sugestão de relação entre as escalas de medida, que viabilize a comparação de resultados científicos.

Este artigo teve por objetivo ilustrar o potencial da equalização de escalas para a área da saúde e demonstrar seu uso, através dos dados de pacientes portadores da doença de Machado-Joseph avaliados por duas escalas diferentes, NESSCA e SARA. Estudos futuros podem ampliar o uso da técnica para lidar com outras doenças e desfechos clínicos avaliados por diferentes escalas de medida.

A ser enviado a Revista Brasileira de Epidemiologia.

REFERÊNCIAS

1. de Andrade DF, Tavares HR, Valle R da C. Teoria da Resposta ao Item: Conceitos e Aplicações. São Paulo, SP: ABE; 2000. 154 p.

2. Chen W-H, Revicki DA, Lai J-S, Cook KF, Amtmann D. Linking pain items from two studies onto a common scale using item response theory. J Pain Symptom Manage. 2009;38(4):615–28.

3. McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. Med Care. 2000;38(9 Suppl):II43-59.

4. Coluci MZO, Alexandre NMC, Milani D. Construção de instrumentos de medida na área da saúde. Ciência & Saúde Coletiva. 2015;20(3):925–36.

5. Castro SMJ, Trentini C, Riboldi J. Item response theory applied to the Beck Depression Inventory. Revista Brasileira de Epidemiologia. 2010;13(3):487–501.

6. Moreira Jr. FJ. Aplicações da teoria da resposta ao item (TRI) no Brasil. Revista Brasileira de Biometria. 2010;28(4):137–70.

7. Kieling C, Prestes PR, Saraiva-Pereira ML, Jardim LB. Survival estimates for patients with Machado–Joseph disease (SCA3). Clinical genetics. 2007;72(6):543–545.

8. Saute JAM, Jardim LB. Machado Joseph disease: clinical and genetic aspects, and current treatment. Expert Opinion on Orphan Drugs. 2015;3(5):517–535.

9. Schmitz-Hübsch T, du Montcel ST, Baliko L, Berciano J, Boesch S, Depondt C, et al. Scale for the assessment and rating of ataxia: development of a new clinical scale. Neurology. 2006 Jun 13;66(11):1717–20.

10. Kieling C, Rieder CRM, Silva ACF, Saute JAM, Cecchin CR, Monte TL, et al. A neurological examination score for the assessment of spinocerebellar ataxia 3 (SCA3). European journal of neurology. 2008;15(4):371–376.

11. Maciel TH. Aplicação da Teoria da Resposta ao Item ao escore NESSCA de avaliação da progressão da Doença de Machado Joseph. [Porto Alegre]: Universidade Federal do Rio Grande do Sul; 2013.

12. Jardim LB, Hauser L, Kieling C, Saute JAM, Xavier R, Rieder CRM, et al. Progression Rate of Neurological Deficits in a 10-Year Cohort of SCA3 Patients. The Cerebellum. 2010;9(3):419–28.

13. Saute JAM, de Castilhos RM, Monte TL, Schumacher-Schuh AF, Donis KC, D'Ávila R, et al. A randomized, phase 2 clinical trial of lithium carbonate in Machado-Joseph disease: Lithium Trial in Machado-Joseph Disease. Movement Disorders. 2014;29(4):568–73.

14. Oliveira CM, Reckziegel ER, Augustin MC, Rocha AG, Bolzan G, Santos JA, et al. Causal factors behind early- and late-onset Machado-Joseph disease patients do not interfere with the rate of neurological deterioration. In Pisa; 2017. Available from: http://www.iarc2017.com/wp-content/uploads/2017/09/IARC-Abstract-Book.pdf

15. Saute JAM, da Silva ACF, Souza GN, Russo AD, Donis KC, Vedolin L, et al. Body Mass Index is Inversely Correlated with the Expanded CAG Repeat Length in SCA3/MJD Patients. The Cerebellum. 2012;11(3):771–4.

16. Donis KC, Saute JAM, Krum-Santos AC, Furtado GV, Mattos EP, Saraiva-Pereira ML, et al. Spinocerebellar ataxia type 3/Machado-Joseph disease starting before adolescence. neurogenetics. 2016;17(2):107–13.

17. Kolen MJ, Brennan RL. Test equating, scaling, and linking: methods and practices. Third Edition. New York: Springer; 2014. 566 p. (Statistics for Social and Behavioral Sciences).

18. Nering ML, Ostini R. Handbook of polytomous item response theory models. New York, NY: Routledge; 2010. 296 p.

19. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park, Calif: Sage Publications; 1991. 174 p. (Measurement methods for the social sciences series).

20. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38(9 Suppl):II28-42.

21. Hattie J. Methodology Review: Assessing Unidimensionality of Tests and ltems. Applied Psychological Measurement. 1985;9(2):139–64.

22. Kim S, Lee W-C. An Extension of Four IRT Linking Methods for Mixed-Format Tests. Journal of Educational Measurement. 2006;43(1):53–76.

23. Hanson BA, Béguin AA. Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate Versus Concurrent Estimation in the Common-Item Equating Design. Applied Psychological Measurement. 2002;26(1):3–24.

24. Lord FM, Wingersky MS. Comparison of IRT True-Score and Equipercentile Observed-Score "Equatings." Applied Psychological Measurement. 1984;8(4):453–61.

25. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: https://www.R-project.org/

26. Raiche G. **nFactors**: An *R* package for parallel analysis and non graphical solutions to the Cattell scree test. 2010; Available from: http://CRAN.R-project.org/package=nFactors

27. Rizopoulos D. **ltm** : An R package for Latent Variable Modelling and Item Response Theory Analyses. Journal of Statistical Software. 2006;17(5):1–25.

28. Weeks JP. **plink**: An R package for Linking Mixed-Format Tests Using IRT-Based Methods. Journal of Statistical Software. 2010;35(12):1–33.

29. Datta D. **blandr**: A Bland-Altman Method Comparison package for *R*. 2017; Available from: https://github.com/deepankardatta/blandr

7. CONCLUSÕES E CONSIDERAÇÕES FINAIS

A equalização de escalas, no contexto da área da saúde, é uma técnica bastante interessante e ainda pouco explorada (Chen et al., 2009; McHorney e Cohen, 2000). Através da equalização é possível comparar resultados de publicações científicas que utilizam diferentes escalas de avaliação. A técnica tem muito potencial pois são muitas as oportunidades de aplicação, devido à quantidade de diferentes escalas que são desenvolvidas e utilizadas em diversos desfechos clínicos. Este trabalho teve por objetivo explorar a aplicabilidade da técnica de equalização via TRI no contexto da saúde, para difundir e ampliar o uso da técnica, além de estabelecer uma relação de equivalência entre os escores das escalas NESSCA e SARA na avaliação do comprometimento pela doença de Machado-Joseph.

Assim como a TRI, a equalização de escalas também está bastante difundida na avaliação educacional, de onde derivam a maioria dos livros e publicações. Os exemplos, as regras, as recomendações e interpretações estão voltados, geralmente, para testes educacionais. Por isso, existe uma dificuldade em adaptar as situações encontradas na área da saúde à realidade já estabelecida pela técnica na literatura.

Como qualquer técnica estatística, a equalização pode apresentar erros. De fato, a relação obtida só será perfeita em todos os casos se os instrumentos de medida equalizados forem idênticos, e nesse caso o uso da técnica perde o sentido. Dessa forma, é importante enxergar a equalização como uma sugestão de relação entre as escalas de medida, que viabilize a comparação de resultados científicos.

Do ponto de vista estatístico, seria interessante um estudo de simulação comparando os métodos de transformação linear, os tipos de calibração e as abordagens de escore verdadeiro e escore observado, bem como revisar métodos não detalhados neste trabalho, como a equalização via *kernel*. Neste trabalho optou-se por calibrar as escalas utilizando os modelos GRM e GPCM devido a característica dos itens, mas outros modelos TRI podem ser avaliados.

8. ANEXOS DA DISSERTAÇÃO

ANEXO A

Escala SARA

Quadro 1. Escala SARA em português.

Item	Provas	Gravidade	Categoria de Resposta
1 - Marcha	O paciente é solicitado a andar em uma distância segura paralela a uma parede e dar uma meia-volta (meia-volta para direção oposta da marcha) e a andar pé-ante-pé sem apoio.	Normal, sem dificuldade para andar, virar-se ou andar na posição pé-ante-pé (até um erro aceito).	0
		Discretas dificuldades, somente visíveis quando anda 10 passos consecutivos na posição pé-ante-pé.	1
		Claramente anormal, marcha na posição pé-ante-pé impossível com 10 ou mais passos.	2
		Consideravelmente cambaleante dificuldades na meia-volta, mas ainda sem apoio.	3
		Mareadamente cambaleante, necessitando de apoio intermitente da parede.	4
		Gravemente cambaleante, apoio permanente com uma bengala ou apoio leve de um braço.	5
		Marcha > 10m somente possível com apoio forte (2 bengalas especiais ou um andador ou um acompanhante).	6
		Macha < 10m somente possível com um apoio forte (2 bengalas especiais ou um andador ou um acompanhante).	7
		Incapaz de andar mesmo com apoio.	8

Item	Provas	Gravidade	Categoria de Resposta
2 – Equilíbrio de Pé	O paciente é solicitado a: permanecer na posição natural, permanecer com os pés juntos e em paralelo (dedões juntos) e a permanecer em pé-ante-pé (ambos os pés em uma linha, sem espaço entre os tornozelos e os dedos); Deve-se retirar os sapatos e os olhos devem permanecer abertos. Para cada condição, três tentativas são permitidas. A melhor resposta é considerada.	Normal, consegue permanecer em pé na posição pé-ante-pé por > 10s.	0
		Capaz de permanecer em pé com os pés juntos sem desvios, mas não na posição pé-ante-pé por > 10s.	1
		Capaz de permanecer em pé com os pés juntos por > 10s, mas somente com desvios.	2
		Capaz de permanecer em pé por > 10s sem apoio na posição natural, mas não com os pés juntos.	3
		Capaz de permanecer em pé por > 10s na posição natural somente com apoio intermitente.	4
		Capaz de permanecer em pé por > 10s na posição natural somente com apoio constante de um braço.	5
		Incapaz de permanecer em pé por > 10s mesmo com apoio constante do braço.	6
3 – Equilíbrio Sentado	O paciente é solicitado a sentar na cama de exame sem apoio dos pés, olhos abertos e braços esticados na frente.	Normal, sem dificuldades em sentar > 10s.	0
		Discretas dificuldades, desvios leves.	1
		Desvios constantes, mas capaz de sentar > 10s sem apoio.	2
		Capaz de sentar > 10s somente com apoio intermitente.	3
		Incapaz de sentar > 10s sem um apoio constante.	4
4 – Coordenação da Fala	A fala é avaliada durante uma conversação normal.	Normal	0
		Sugestivo de alteração na fala.	1
		Alteração na fala, mas fácil de entender.	2
		Ocasionalmente palavras difíceis de entender.	3
		Muitas palavras difíceis de entender.	4
		Somente palavras isoladas compreensíveis.	5
		Fala ininteligível/anartria.	6

Item	Provas	Gravidade	Categoria de Resposta
5 – Teste de Perseguição do Dedo (cada lado avaliado isoladamente)	O paciente permanece confortavelmente sentado. Se necessário, é permitido o apoio dos pés e do tronco. O examinador senta em frente do paciente e realiza 5 movimentos consecutivos inesperados e rápidos de apontar em um plano frontal, a mais ou menos 50% do alcance do paciente. Os movimentos deverão ter uma amplitude de 30cm e uma frequência de 1 movimento a cada 2s. O paciente é solicitado a seguir os movimentos com o índex, o mais preciso e rápido possível. É considerada a execução dos 3 últimos movimentos.	Ausência de dismetria.	0
		Dismetria, não atingir ou ultrapassar o alvo < 5cm.	1
		Dismetria, não atingir ou ultrapassar o alvo < 15cm.	2
		Dismetria, não atingir ou ultrapassar o alvo > 15cm.	3
		Incapaz de realizar os 5 movimentos.	4
6 – Teste Dedo-Nariz (cada lado avaliado isoladamente)	O paciente permanece confortavelmente sentado. Se necessário, é permitido o apoio dos pés e do tronco. É solicitado que o paciente aponte repetidamente seu índex em seu nariz para o dedo do examinador, que está a cerca de 90% do alcance do paciente. Os movimentos são realizados a uma velocidade moderada. A execução do movimento é graduada de acordo com a amplitude do tremor de ação.	Ausência de tremor.	0
		Tremor com uma amplitude < 2cm.	1
		Tremor com uma amplitude < 5cm.	2
		Tremor com uma amplitude > 5cm.	3
		Incapaz de realizar os 5 movimentos.	4
7 – Diadococinesia (cada lado avaliado isoladamente)	O paciente deve permanecer confortavelmente sentado. Se necessário, é permitido o apoio dos pés e do tronco. É solicitado que o paciente realize 10 ciclos com alternação pronação e supinação em suas coxas o mais rápido e preciso possível O tempo exato para execução do movimento deverá ser obtido.	Normal, sem irregularidades (realiza < 10s).	0
		Discretamente irregular (realiza < 10s).	1
		Claramente irregular, difícil de distinguir movimentos individuais ou interrupções relevantes, mas realiza < 10s.	2
		Muito irregular, difícil de distinguir movimentos individuais ou interrupções relevantes, realiza > 10s.	3
		Incapaz de completar 10 ciclos.	4

Item	Provas	Gravidade	Categoria de Resposta
8 – Teste Calcanhar-Joelho-Canela (cada lado avaliado isoladamente)	O paciente deita na cama de exame, sem conseguir visualizar suas pernas. É solicitado que levante uma perna, aponte com o calcanhar no outro joelho, deslize pela tíbia até o tornozelo e retorne a perna em repouso na cama. A tarefa é realizada 3 vezes. O movimento de deslizamento deverá ser feito em 1s. Se o paciente deslizar sem o contato com a tíbia em todas as três tentativas gradue como 4.	Normal	0
		Discretamente anormal, contato com a tíbia mantido.	1
		Claramente anormal, saída da tíbia mais do que 3 vezes durante 3 ciclos.	2
		Gravemente anormal, saída da tíbia 4 ou mais vezes durante 3 ciclos.	3
		Incapaz de realizar a tarefa.	4

Fonte: adaptado de Schmitz-Hübsch et al. (2006).

ANEXO B

Escala NESSCA

Quadro 2. Escala NESSCA em português.

Item	Provas	Gravidade	Categoria de Resposta
1 – Ataxia de Marcha	- Andar espontaneamente, 10 passos, paralelos a uma parede e incluindo uma meia volta; - Andar na ponta dos pés, com os calcanhares e em conjunto.	Ausente.	0
		Mínima: apenas ao andar na ponta dos pés, com os calcanhares e em conjunto.	1
		Moderado: autonomia de marcha preservada.	2
		Incapacidade de caminhar sem ajuda.	3
		Cadeira de rodas ou acamados.	4
2 – Ataxia nos Membros (bilateral)	- Teste dedo-nariz; - Diadococinésia (movimentos rápidos e alternados das mãos com os cotovelos fixos); - Rebote de Gordon Holmes (rebote do membro superior).	Ausente.	0
		Mínima: uma prova alterada.	1
		Moderado: duas provas alteradas.	2
		Importante: três provas alteradas.	3
		Provas: (a) dismetria, (b) movimentos rápidos e alternados das mãos e (c) rebote do membro inferior. Resultados positivos podem ser uni ou bilaterais.	
3 – Nistagmo		Ausente.	0
		No olhar fixo ou circular, depois de sacadas.	1
		Permanente.	2
4 – Oftalmoplegia Externa Progressiva		Ausente.	0
		Supranuclear: síndrome do fascículo longitudinal medial ou limitação ao olhar para cima ou para a convergência.	1
		Oftalmoplegia nuclear, com estrabismo.	2
5 – Achados Piramidais	- Reflexo nos membros, incluindo teste de clônus patelar e tornozelo; - Reflexo plantar; - Exame de tônus muscular; - Teste de força motor: braços estendidos e teste de Mingazzini (60 segundos cada).	Ausente.	0
		Poucos reflexos rápidos.	1
		Hiperreflexia geral ou clônus ou sinal de Babinski	2
		Três achados: (a) hiperreflexia geral; (b) espasticidade; (c) clônus; (d) sinal de Babisnki ou (e) paresia.	3
		Quatro ou cinco dos acima mencionados.	4

Item	Provas	Gravidade	Categoria de Resposta
6 – Disartria		Ausente.	0
		Leve: Dificuldade de fala, mas fácil de entender.	1
		Moderado: discurso compreensível, mas com dificuldade.	2
		Grave: discurso de difícil compreensão.	3
		Anartria.	4
7 – Disfagia		Ausente.	0
		Leve.	1
		Importante: ocorrendo todos os dias.	2
8 – Fasciculações		Ausente.	0
		Contração de fasciculação no rosto.	1
		Difusa ou em outras partes do corpo.	2
9 – Perda Sensorial	(a) Sensação vibratória nos dedos dos pés: normal < 11 segundos; (b) Discriminação entre estímulos táteis e algésicos usando uma agulha; 10 tentativas por pé; (c) Discriminação entre água fria (10ºC) e água quente (40ºC - 60ºC); 10 tentativas por pé.	Ausente.	0
		Uma prova alterada: redução em (a) ou (b) ou (c), de dois a quatro erros, em média, nos dois pés.	1
		Duas provas alteradas.	2
		Perda total do sentido de vibração nos dedos dos pés; cinco ou mais erros em uma das provas de discriminação ou três provas alteradas.	3
10 – Distonia		Ausente.	0
		Leve, acionada por movimentos voluntários.	1
		Moderada, prejudicando, em algum grau, movimentos voluntários.	2
		Quase constantes, prejudicando severamente os movimentos voluntários.	3
11 – Rigidez		Ausente.	0
		Moderada: não impede mobilização total, mobilização passiva.	1
		Importante: impede total, mobilização passiva.	2

Item	Provas	Gravidade	Categoria de Resposta
12 – Bradicinésia	- O paciente é solicitado a realizar 10 ciclos de repetição (extensão e flexão) do segundo dedo contra o polegar.	Ausente.	0
		Movimentos lentos, com redução de amplitude.	1
		Movimentos dificilmente podem ser feitos.	2
13 – Retração Palpebral		Ausente.	0
		Presente.	1
14 – Blefarospasmo		Ausente.	0
		Presente.	1
15 – Amiotrofia Distal	- Inspeção dos músculos interósseos, tenar e hipotênar.	Ausente.	0
		Presente.	1
16 – Função do Esfíncter		Normal	0
		Urgência	1
		Incontinência.	2
17 – Câimbra		Ausente.	0
		Presente.	1
18 – Vertigem		Ausente.	0
		Presente.	1

Fonte: adaptado de Maciel (2013).

ANEXO C

Métodos Clássicos de Equalização

Métodos Clássicos de Equalização

A descrição dos métodos foi baseada em Kolen e Brennan (2014).

Random Groups

O delineamento de grupos aleatórios é um dos mais simples, uma vez que a única preocupação do pesquisador será com o tamanho da amostra e homogeneidade entre os grupos sorteados. Os três principais métodos para equalização utilizando este delineamento são os seguintes:

a) *Mean Equating* (equalização pela média)*:* este método parte do princípio que os instrumentos X e Y diferem por uma constante que pode ser obtida através da média dos instrumentos. Por exemplo, se o grupo que respondeu o instrumento X obteve, em média, escore x e o grupo que respondeu o instrumento Y, obteve, em média, escore y, este método considera que a diferença de pontos entre os questionários deve ser adicionada ou deduzida do escore obtido pelo instrumento X para fazer a transformação para o escore no instrumento Y.

Seja X a variável que representa o escore obtido no instrumento X e Y a variável que representa o escore obtido no instrumento Y, e sejam x e y escores particulares obtidos nos em X e Y, respectivamente. Seja $\mu(X)$ a média obtida pelo grupo que respondeu ao instrumento X, e equivalentemente, $\mu(Y)$, a média obtida pelo grupo que respondeu ao instrumento Y. Tem-se:

$$m_Y(x) = y = x - \mu(X) + \mu(Y), \tag{41}$$

onde $m_Y(x)$ se refere ao escore de x no instrumento X transformado para a escala do instrumento Y, utilizando o método da média.

b) *Linear Equating* (equalização linear): ao invés de considerar apenas uma constante, como no método da média, esse método possibilita a existência de diferenças que variam ao longo da escala. Por exemplo, entre os escores mais baixos é possível ter maior variabilidade do que entre os indivíduos que obtiveram escores mais altos. Se o desvio padrão entre os escores de cada instrumento de medida for igual, isto é, $\sigma(X) = \sigma(Y)$, este método irá se equiparar ao método da média. No entanto, quando há diferença na variabilidade entre os escores obtidos em cada instrumento, isto é, $\sigma(X) \neq \sigma(Y)$, é obtida uma equação com base nos dados:

$$l_Y(x) = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right] \tag{42}$$

onde $l_Y(x)$ é a conversão do escore obtido em X para a escala Y.

c) *Equipercentile Equating* (equalização pelo equipercentil): este método é ainda mais flexível que o linear, onde uma curva é utilizada para descrever as diferenças entre os instrumentos de medida. É verificada a função de distribuição acumulada dos escores abaixo de um percentil específico para obter escores equivalentes. Por exemplo: dentre os respondentes do instrumento X, 20% obtiveram até x pontos, enquanto que dentre os respondentes do instrumento Y, 20% obtiveram até y pontos, ou seja, o escore x no instrumento X equivale ao escore y no instrumento Y. Este método funciona bem para variáveis contínuas, para utilizar em variáveis discretas é preciso utilizar intervalos de percentil. Seja G^{-1} a função de distribuição acumulada inversa dos escores obtidos no instrumento Y e $F(x)$ a função de distribuição acumulada de X. Então:

$$e_Y(x) = G^{-1}[F(x)], \tag{43}$$

onde $e_Y(x)$ é a conversão do escore obtido em X para a escala Y.

Mesmo quando o tamanho da amostra é grande, a estimativa por percentis amostrais pode não ser suficientemente precisa devida ao erro amostral. Uma estratégia para contornar essa situação é fazer uso de métodos de suavização, os quais não serão abordados neste trabalho e podem ser consultados em González e Wiberg (2017) e Kolen e Brennan (2014).

Single Group

Para o delineamento de um único grupo (sem balanceamento), os métodos de equalização são iguais aos já descritos para o delineamento RG. No entanto, conforme já mencionado, os resultados podem ser prejudicados devido a fatores relacionados à ordem de aplicação dos instrumentos, por isso recomenda-se utilizar a versão com balanceamento.

Single Group with Counterbalancing

Para este delineamento existem duas formas de conduzir a equalização:

a) Equalizar os instrumentos X e Y considerando o delineamento RG para aqueles que responderam primeiro o instrumento X e aqueles que responderam primeiro o instrumento Y, utilizando um dos métodos de equalização já descritos para o delineamento RG;

b) Equalizar os instrumentos X e Y considerando o delineamento RG para aqueles que responderam em um segundo momento o instrumento X e aqueles que responderam em um segundo momento o instrumento Y, utilizando um dos métodos de equalização já descritos para o delineamento RG;

Comparar os resultados obtidos nas instruções a) e b). Se os resultados forem muito diferentes, significa que os instrumentos X e Y são diferentemente afetados quando respondidos em segundo lugar. Nesse caso, utilizar apenas os dados obtidos no instrumento que foi aplicado primeiro para cada respondente, pois assim não terá um efeito de ordem de aplicação. Se os resultados de a) e b) forem semelhantes, então os formulários são igualmente sensibilizados quando aparecem em segundo lugar na ordem de aplicação. Nesse caso, pode-se fazer a equalização considerando todos os dados disponíveis, utilizando os métodos clássicos apresentados para o delineamento RG.

Common Item Nonequivalent Groups

Este delineamento geralmente é utilizado quando não é possível aplicar dois instrumentos de medida a um mesmo indivíduo, por limitações de tempo ou de confidencialidade, bem como quando não há a garantia de homogeneidade entre os grupos de respondentes. Para conduzir a equalização, os instrumentos devem ter itens em comum, que podem ser internos ou externos. São ditos internos quando os itens em comum contribuem para o escore final do respondente, e são ditos externos quando os itens em comum possuem o único propósito de viabilizar a equalização, sendo desconsideradas as respostas obtidas nestes itens nos resultados finais da análise.

Em geral, este delineamento é usado para ajustar funções para diferentes populações, por isso os métodos de equalização precisam atender a mais premissas e são mais complexos. Os métodos lineares estão listados a seguir (Davier, 2011):

a) *Tucker Method*: este método se baseia em uma regressão do escore total a partir do escore obtido nos itens comuns. Também supõe que a variância condicional pode ser estimada a partir do escore dos itens comuns;

b) *Levine Score Method*: o método de Levine faz uso da suposição que diz que o escore de um respondente pode ser decomposto em duas partes, o verdadeiro escore (que não varia mesmo com repetições) e um erro aleatório;

c) *Chained Linear Equating*: este método parte do princípio que a função simétrica linear que conecta o escore total obtido no instrumento X ao escore total obtido no instrumento Y não varia de acordo com a população.

Para mais detalhes e exemplos ilustrativos, consultar Davier (2011).

Os métodos que utilizam os equipercentis são mais robustos porque utilizam mais informação do que apenas a média e os desvios, uma vez que a equalização ocorre ao longo de toda a escala de escore. Os principais estão listados abaixo:

a) *Frequency Estimation Method*: este método retorna a média da estimativa da distribuição acumulada dos escores do instrumento X e do instrumento Y para uma população sintética, que é uma ponderação dos dois grupos;

b) *Modified Frequency Estimation Method*: este método é baseado na alteração do método de estimação da frequência, com o objetivo de corrigir vieses de equalização;

c) *Chained Equipercentile Method*: neste método, escores obtidos através do instrumento X são convertidos para os escores dos itens comuns utilizando respondentes do grupo 1, enquanto escores do itens comuns são convertidos para os escores do instrumento Y utilizando respondentes do grupo 2. Esses dois resultados são encadeados para obter uma conversão dos escores do instrumento X para escores do instrumento Y.

Em termos computacionais, já existem alguns pacotes com funções implementadas que conduzem métodos clássicos de equalização. Para o software R, por exemplo, os pacotes *equate* (Albano, 2016) e *SNSequate* (González, 2017) desempenham a maioria das técnicas clássicas de equalização. Paralelamente, outra forma de estabelecer uma relação de equalização entre escalas é através do método de suavização de Kernel (González e Wiberg, 2017; Kolen e Brennan, 2014), o qual está implementado no pacote *kequate* (Andersson et al. 2013).

ANEXO D

Equalização do escore observado

Equalização do escore observado

Seja $f(x|\theta)$ e $f(x)$ a distribuição condicional e a marginal dos escores do instrumento X, e, equivalentemente $f(y|\theta)$ e $f(y)$ as distribuições dos escores do instrumento Y. Para obter a distribuição condicional do escore observado de indivíduos com um certo valor de θ é necessário calcular, entre os itens e categorias, todas as possíveis combinações de respostas que levam a um escore específico. Lord e Wingersky (1984) propuseram um algoritmo que resolve esse problema. Seja $f_r(x|\theta)$ a distribuição de probabilidade condicional dos escores dos primeiros r itens de indivíduos com traço latente θ e seja m_1 a quantidade de categorias do item 1, cujas probabilidades $P_{11}(\theta), P_{12}(\theta), \dots, P_{1m}(\theta)$ são, respectivamente, $f_1(x = U_{11}|\theta), f_1(x = U_{12}|\theta), \dots, f_1(x = U_{1m}|\theta)$. Logo, para $r > 1$, a fórmula para encontrar a probabilidade de respondentes com traço latente θ obter um escore x depois do r-ésimo item, é:

$$f_r(x|\theta) = \sum_{k=1}^{m_r} f_{r-1}(x - U_{rk}|\theta) P_{rk}(\theta) \tag{44}$$

com $min < x < max$, onde min é o escore mínimo depois de adicionar o r-ésimo item, equivalentemente para max.

Este procedimento determina a distribuição do escore observado para respondentes que possuem um certo valor de θ. A generalização desta função é uma integral, que pode ser escrita como uma aproximação de um finito número de θ:

$$f(x) = \sum_{g=1}^{G_X} f\left(x|\theta_g\right) W(\theta_g) \tag{45}$$

onde θ_g e $W(\theta_g)$ são os pontos de quadratura e seus respectivos pesos utilizados para representar a distribuição dos respondentes do instrumento X.

Se N_X são os respondentes do instrumento X com traços latentes $\theta_1, \theta_2, \dots, \theta_{N_X}$, a distribuição marginal dos escores observados pode ser expressa por

$$f(x) = \frac{1}{N_X} \sum_{g=1}^{N_X} f\left(x|\theta_g\right) \tag{46}$$

e, similarmente para Y,

$$f(y) = \frac{1}{N_Y} \sum_{g=1}^{N_Y} f\left(y|\theta_g\right) \tag{47}$$

Estas duas distribuições marginais são equalizadas utilizando o método do equipercentil – é verificada a função de distribuição acumulada dos escores abaixo de um percentil específico para obter escores equivalentes.

A Figura 12 ilustra uma relação fictícia entre os a distribuição dos equipercentis. Por exemplo, o equipercentil 90 no instrumento X corresponde ao escore 20, enquanto que o mesmo equipercentil no instrumento Y corresponde ao escore 22.

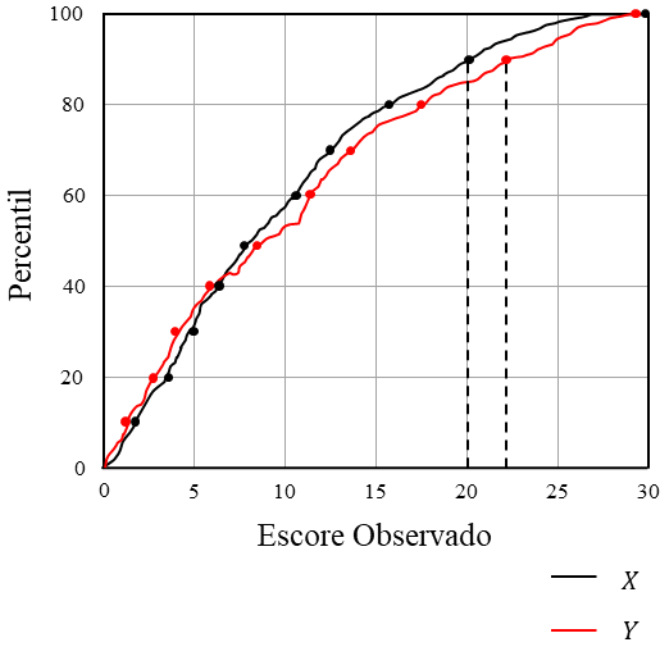

Figura 1. Equalização do escore observado utilizando o método do percentil.

Fonte: adaptado de Nering e Ostini (2010).

9. ANEXOS DO ARTIGO

ANEXO A

Modelos GRM e GPCM

Modelo de Resposta Gradual

O Modelo de Resposta Gradual (Samejima, 1969), diferentemente do NRM, assume que as categorias de um item tenham uma ordem. Conhecido por GRM (em inglês, *Graded-Response Model*), trata-se de uma generalização do 2PL e é considerado um modelo TRI "indireto" pois requer um procedimento adicional para calcular a probabilidade condicional de um indivíduo ter um determinado nível do sintoma. Uma vantagem do GRM é que os itens do instrumento não precisam ter a mesma quantidade de categorias de resposta, como ocorre na NESSCA e na SARA. Considerando que as possíveis categorias de um item sejam denotadas por $k = 0,1,\dots,m_i$ onde $m_i + 1$ é o número de categorias do item i, a probabilidade de um indivíduo j pertencer a uma particular categoria ou outra mais alta pode ser dada por:

$$P_{i,k}^{+}(\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_{i,k})}}, \quad (1)$$

com $i = 1,2,\dots,I, j = 1,2,\dots,n$ e $k = 0,1,\dots,m_i$, onde:

θ_j representa a intensidade do comprometimento pela SCA3/MJD (traço latente) do j-ésimo paciente;

a_i é o parâmetro de inclinação comum a todas as categorias de um mesmo item i;

$b_{i,k}$ é o parâmetro de posição da k-ésima categoria do item i, ou seja, representa o nível de comprometimento necessário para a escolha da categoria de resposta k, ou acima de k, com probabilidade igual a 0,50;

D é um fator de escala, constante e igual a 1. Utiliza-se 1,7 quando se deseja que a função logística forneça resultados semelhantes ao da função ogiva normal;

I é o número de itens no instrumento de medida;

n é o número de indivíduos respondentes.

Deverá existir uma ordenação entre as categorias de um dado item, ou seja:

$$b_{i,1} \leq b_{i,2} \leq \cdots \leq b_{i,m_i}$$

A probabilidade de um indivíduo j pertencer a categoria k no item i é dada pela expressão:

$$P_{i,k}(\theta_j) = P_{i,k}^{+}(\theta_j) - P_{i,k+1}^{+}(\theta_j) \quad (2)$$

onde $P_{i,0}^{+}(\theta_j) = 1$ e $P_{i,m_i+1}^{+}(\theta_j) = 0$, logo:

$$P_{i,k}(\theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_{i,k})}} - \frac{1}{1 + e^{-Da_i(\theta_j - b_{i,k+1})}} \quad (3)$$

O número de parâmetros, por item, será dado pelo número de categorias k do item i.

Por se tratar de modelos para itens politômicos, com mais de duas categorias, a equação (3) gera as curvas de categoria de resposta (CCR), as quais são simbolizadas por $P_{ik}(\theta)$. Estas curvas ilustram a relação entre a probabilidade de um indivíduo com comprometimento θ pertencer a categoria k do sintoma do item i. O conjunto de todas as CCRs de um teste resulta na curva característica do teste (CCT) que representa a probabilidade de obtenção de um escore total em função de θ, pode ser interpretado como a média para um dado valor de θ. Para um teste com I itens, a CCT é dada por:

$$T(\theta) = \sum_{i=1}^{I} \sum_{k=1}^{m_i} U_{ik} P_{ik}(\theta) \quad (4)$$

onde U_{ik} é uma função do escore do item, ou seja, são os valores de escore possíveis de obter no item i.

Geralmente utiliza-se $U_{ik} = k - 1$ (quando uma resposta associada a primeira categoria recebe escore zero, que é o caso da NESSCA e da SARA) ou $U_{ik} = k$ (quando uma resposta associada a primeira categoria representa escore igual a 1) (Kolen e Brennan, 2014).

Modelo de Crédito Parcial Generalizado

O modelo de crédito parcial generalizado foi desenvolvido por Muraki, em 1992, e consiste de uma generalização do PCM, relaxando a hipótese de poder de discriminação igual para todos os itens (Muraki, 1992). Ou seja, permite que os itens dentro de uma escala tenham diferentes parâmetros de inclinação, o que é interessante no contexto da NESSCA e da SARA. Também conhecido como *Generalized Partial Credit Model* (GPCM), supondo que o item i possui $m_i + 1$ categorias de resposta ordenadas $(k = 0, 1, \dots, m_i)$, temos que o modelo é dado por:

$$P_{i,k}(\theta_j) = \frac{exp\left[\sum_{u=0}^{k} Da_i(\theta_j - b_{i,u})\right]}{\sum_{u=0}^{m_i} exp\left[\sum_{v=0}^{u} Da_i(\theta_j - b_{i,v})\right]}, \quad (5)$$

com $i = 1,2, \dots, I, j = 1,2, \dots, n$ e $k = 0,1, \dots, m_i$, onde:

θ_j representa a intensidade do comprometimento pela SCA3/MJD (traço latente) do j-ésimo paciente;

$P_{i,k}(\theta_j)$ é a probabilidade de um indivíduo com nível de comprometimento θ_j ter um

sintoma na categoria k dentre as $m_i + 1$ categorias do item i;

a_i é o parâmetro de inclinação do item i;

$b_{i,k}$ é o parâmetro do item que regula a probabilidade do sintoma ser k ao invés da categoria adjacente $(k-1)$ no item i. Cada parâmetro $b_{i,k}$ corresponde ao valor do traço latente no qual o indivíduo tem a mesma probabilidade de ser classificado nas categorias k e $(k-1)$, isto é, onde $P_{i,k}(\theta_j) = P_{i,k-1}(\theta_j)$. Pode ser interpretado como um parâmetro de interseção entre as categorias de resposta do item i;

D é um fator de escala, constante e igual a 1. Utiliza-se 1,7 quando se deseja que a função logística forneça resultados semelhantes ao da função ogiva normal;

I é o número de itens no instrumento de medida;

n é o número de indivíduos respondentes.

É importante observar que o parâmetro de inclinação (a_i) presente neste modelo não deve ser interpretado diretamente, da mesma forma como nos modelos dicotômicos. Nos modelos politômicos, a discriminação do item depende da combinação de a_i com a distribuição dos parâmetros $b_{i,k}$. Os parâmetros $b_{i,k}$ são os pontos na escala onde as curvas das categorias se cruzam, em qualquer ponto da escala θ_j. No geral, define-se $b_{i,0} = 0$. Além disso, frequentemente os parâmetros $b_{i,k}$ são decompostos em um parâmetro de posição b_i e nos parâmetros para as categorias, $d_{i,k}$, onde:

$$b_{i,k} = b_i - d_{i,k} \tag{6}$$

Para avaliar a contribuição de um item politômico pode-se observar a Curva de Informação do Item (CII). Essa curva indica a quantidade de informação que um determinado sintoma contribui para a medida do traço latente e em qual intervalo esse sintoma é mais informativo (Castro et al., 2010).

ANEXO B

Resultado do ajuste dos modelos

Tabela 1. Resultado do ajuste dos modelos para a NESSCA cerebelar.

Modelo	AIC	BIC
GPCM	1042,66	1092,72
GRM	1038,08	1088,14

Tabela 2. Resultado do ajuste dos modelos para a SARA.

Modelo	AIC	BIC
GPCM	2106,07	2221,71
GRM	2114,17	2229,80

ANEXO C

Adaptação das categorias da SARA

Quadro 3. Adaptação das categorias da SARA para o item Marcha.

NESSCA		SARA	
Gravidade	**Categoria**	**Gravidade**	**Categoria**
Ausente.	0	Normal, sem dificuldade para andar, virar-se ou andar na posição pé-ante-pé (até um erro aceito).	0
Mínima: apenas ao andar na ponta dos pés, com os calcanhares e em conjunto.	1	Discretas dificuldades, somente visíveis quando anda 10 passos consecutivos na posição pé-ante-pé.	1
		Claramente anormal, marcha na posição pé-ante-pé impossível com 10 ou mais passos.	2
Moderado: autonomia de marcha preservada.	2	Consideravelmente cambaleante dificuldades na meia-volta, mas ainda sem apoio.	3
		Mareadamente cambaleante, necessitando de apoio intermitente da parede.	4
Incapacidade de caminhar sem ajuda.	3	Gravemente cambaleante, apoio permanente com uma bengala ou apoio leve de um braço.	5
		Marcha > 10m somente possível com apoio forte (2 bengalas especiais ou um andador ou um acompanhante).	6
		Macha < 10m somente possível com um apoio forte (2 bengalas especiais ou um andador ou um acompanhante).	7
Cadeira de rodas ou acamados.	4	Incapaz de andar mesmo com apoio.	8

Quadro 4. Adaptação das categorias da SARA para o item Coordenação da Fala.

NESSCA		SARA	
Gravidade	**Categoria**	**Gravidade**	**Categoria**
Ausente.	0	Normal	0
Leve: Dificuldade de fala, mas fácil de entender.	1	Sugestivo de alteração na fala.	1
		Alteração na fala, mas fácil de entender.	2
Moderado: discurso compreensível, mas com dificuldade.	2	Ocasionalmente palavras difíceis de entender.	3
Grave: discurso de difícil compreensão.	3	Muitas palavras difíceis de entender.	4
		Somente palavras isoladas compreensíveis.	5
Anartria.	4	Fala ininteligível/anartria.	6

ANEXO D

Sintaxe

Este anexo apresenta algumas partes da sintaxe consideradas mais relevantes para este trabalho.

Análise Fatorial

```
N<- cor(nessca[,6:10])
componentAxis(N, nFactors=3)
S<-cor(sara[6:13])
componentAxis(S, nFactors=3)
```

Calibração dos Parâmetros – NESSCA

```
t2_nessca <- gpcm(nessca[,6:10])
```

Calibração dos Parâmetros – SARA

```
t1_sara <- gpcm(sara[,6:13])
```

Curva Característica do Item (Exemplo: NESSCA)

```
plot(t2_nessca, type = "ICC", items=1, main="Curva Característica do Item\nAtaxia de Marcha",xlab="θ",ylab="P(θ)")
plot(t2_nessca, type = "ICC", items=2, main="Curva Característica do Item\nAtaxia nos Membros",xlab="θ",ylab="P(θ)")
plot(t2_nessca, type = "ICC", items=3, main="Curva Característica do Item\nNistagmo",xlab="θ",ylab="P(θ)")
plot(t2_nessca, type = "ICC", items=4, main="Curva Característica do Item\nDisartria",xlab="θ",ylab="P(θ)")
plot(t2_nessca, type = "ICC", items=5, main="Curva Característica do Item\nDisfagia",xlab="θ",ylab="P(θ)")
```

Curva de Informação do Item (Exemplo: NESSCA)

```
plot(t2_nessca, type = "IIC", items = 1, ylim = c(0,7), main="Curva de Informação do Item\nAtaxia de Marcha", xlab="θ",ylab="Informação")
plot(t2_nessca, type = "IIC", items = 2, ylim = c(0,7), main="Curva de Informação do Item\nAtaxia nos Membros", xlab="θ",ylab="Informação")
plot(t2_nessca, type = "IIC", items = 3, ylim = c(0,7), main="Curva de Informação do Item\nNistagmo", xlab="θ",ylab="Informação")
plot(t2_nessca, type = "IIC", items = 4, ylim = c(0,7), main="Curva de Informação do Item\nDisartria", xlab="θ",ylab="Informação")
plot(t2_nessca, type = "IIC", items = 5, ylim = c(0,7), main="Curva de Informação do Item\nDisfagia", xlab="θ",ylab="Informação")
```

Preparando input para equalização

Trata-se de uma lista que contém todas as informações necessárias para a equalização: parâmetros dos itens, quantidade de categorias de cada item, modelos TRI utilizados e itens comuns.

```
lista=list(pars=list(
        group1=data.frame(matrix(c(coef(t2_nessca)$Gait[[4]],coef(t2_nessca)$Limb[[4]],coef(t2_nessca)$Nistag[[3]],
coef(t2_nessca)$Disart[[5]],coef(t2_nessca)$Dysphag[[3]],coef(t2_nessca)$Gait[[1]],coef(t2_nessca)$Limb[[1]],coef(t2_
nessca)$Nistag[[1]],coef(t2_nessca)$Disart[[1]],coef(t2_nessca)$Dysphag[[1]],coef(t2_nessca)$Gait[[2]],coef(t2_nessca)
$Limb[[2]],coef(t2_nessca)$Nistag[[2]],coef(t2_nessca)$Disart[[2]],coef(t2_nessca)$Dysphag[[2]],coef(t2_nessca)$Gait
[[3]],coef(t2_nessca)$Limb[[3]],NA,coef(t2_nessca)$Disart[[3]],NA,NA,NA,NA,coef(t2_nessca)$Disart[[4]],NA),5,5)),
        group2=data.frame(matrix(c(coef(t1_sara)$Gait_A_cod[[4]],coef(t1_sara)$Stance[[7]],coef(t1_sara)$Sit[[5]],c
oef(t1_sara)$Speech_cod[[5]],coef(t1_sara)$Chase[[5]],coef(t1_sara)$Nosefing[[5]],coef(t1_sara)$Disdia[[5]],coef(t1_sa
ra)$HeelShin[[5]],coef(t1_sara)$Gait_A_cod[[1]],coef(t1_sara)$Stance[[1]],coef(t1_sara)$Sit[[1]],coef(t1_sara)$Speech
_cod[[1]],coef(t1_sara)$Chase[[1]],coef(t1_sara)$Nosefing[[1]],coef(t1_sara)$Disdia[[1]],coef(t1_sara)$HeelShin[[1]],
coef(t1_sara)$Gait_A_cod[[2]],coef(t1_sara)$Stance[[2]],coef(t1_sara)$Sit[[2]],coef(t1_sara)$Speech_cod[[2]],coef(t1_s
ara)$Chase[[2]],coef(t1_sara)$Nosefing[[2]],coef(t1_sara)$Disdia[[2]],coef(t1_sara)$HeelShin[[2]],coef(t1_sara)$Gait_
A_cod[[3]],coef(t1_sara)$Stance[[3]],coef(t1_sara)$Sit[[3]],coef(t1_sara)$Speech_cod[[3]],coef(t1_sara)$Chase[[3]],coe
f(t1_sara)$Nosefing[[3]],coef(t1_sara)$Disdia[[3]],coef(t1_sara)$HeelShin[[3]],NA,coef(t1_sara)$Stance[[4]],coef(t1_sa
ra)$Sit[[4]],coef(t1_sara)$Speech_cod[[4]],coef(t1_sara)$Chase[[4]],coef(t1_sara)$Nosefing[[4]],coef(t1_sara)$Disdia[[
4]],coef(t1_sara)$HeelShin[[4]],NA,coef(t1_sara)$Stance[[5]],NA,NA,NA,NA,NA,NA,NA,coef(t1_sara)$Stance[[6]],N
A,NA,NA,NA,NA,NA),8,7))),
                cat = list(group1 = c(4,4,3,5,3), group2 = c(4,7,5,5,5,5,5,5)),
                items = list(group1 = list(grm = c(1:5)),
                        group2 = list(gpcm = c(1:8))),
                common = matrix(c(1,1,4,4),2,2,byrow = TRUE))
```

Definição de objetos auxiliares

```
pm1 <- as.poly.mod(5,"grm",lista$items$group1)
pm2 <- as.poly.mod(8,"gpcm",lista$items$group2)
x <- as.irt.pars(lista$pars,lista$common,lista$cat,list(pm1,pm2))
```

Equalização (Exemplo: NESSCA – Método *Mean/Mean*)

```
aux_mm_N <- plink(x, method="MM", rescale="MM", base.grp = 2)
tse_mm_N <- equate(aux_mm_N, D=1.7, method="TSE", base.grp = 2)
colnames(tse_mm_N) <- c("theta","SARA","NESSCA_MM")
```

Comparação das estimativas com o escore observado (Exemplo: NESSCA)

```
tse1_N <- merge(tse_mm_N[,c("SARA", "NESSCA_MM")], tse_ms_N[,c("SARA", "NESSCA_MS")], by = "SARA")
tse2_N <- merge(tse_hb_N[,c("SARA", "NESSCA_HB")], tse_sl_N[,c("SARA", "NESSCA_SL")], by = "SARA")
tse_N <- merge(tse1_N[,c("SARA", "NESSCA_MM", "NESSCA_MS")], tse2_N[,c("SARA", "NESSCA_HB",
"NESSCA_SL")], by = "SARA")
valida_tse_N <- merge(valida[,c("Fonte", "id_bdorig", "SARA", "NESSCA")], tse_N[,c("SARA", "NESSCA_MM",
"NESSCA_MS", "NESSCA_HB", "NESSCA_SL" )], by = "SARA")
```

Apuração das diferenças (Exemplo: NESSCA)

```
valida_tse_N$dif_mm <- abs(valida_tse_N$NESSCA_MM-valida_tse_N$NESSCA)
valida_tse_N$dif_ms <- abs(valida_tse_N$NESSCA_MS-valida_tse_N$NESSCA)
valida_tse_N$dif_hb <- abs(valida_tse_N$NESSCA_HB-valida_tse_N$NESSCA)
valida_tse_N$dif_sl <- abs(valida_tse_N$NESSCA_SL-valida_tse_N$NESSCA)
```

Gráfico Bland-Altman (Exemplo: NESSCA – Método *Mean/Mean*)

```
mm_bland_plot.1_N <- blandr.draw(valida_tse_N$NESSCA_MM, valida_tse_N$NESSCA,plotTitle="Bland-Altman")
mm_bland_plot_N <- mm_bland_plot.1_N +
        ggplot2::coord_cartesian( ylim=c(-20,20)) +
        labs(title="(a) Método Mean/Mean",
        y="Diferenças",
        x="Médias") +
        theme(plot.title = element_text(hjust = 0))
```

ANEXO E

Curvas Característica do Item e Curvas de Informação do Item para cada item da NESSCA

Este anexo apresenta as curvas de categoria de resposta e curvas de informação do item para cada item da NESSCA. Para as curvas de informação do item delimitou-se o eixo y até 7,0 para auxiliar na interpretação e comparação.

1 – Ataxia de Marcha

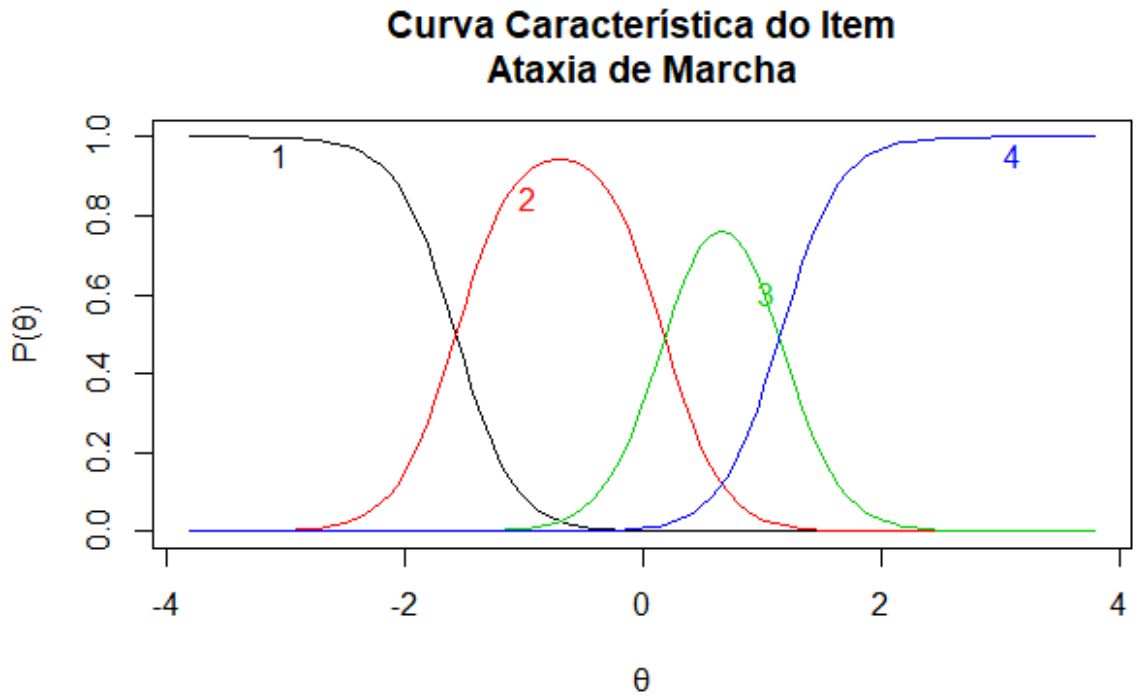

Figura 2. Curva Característica do Item para o item Ataxia de Marcha.

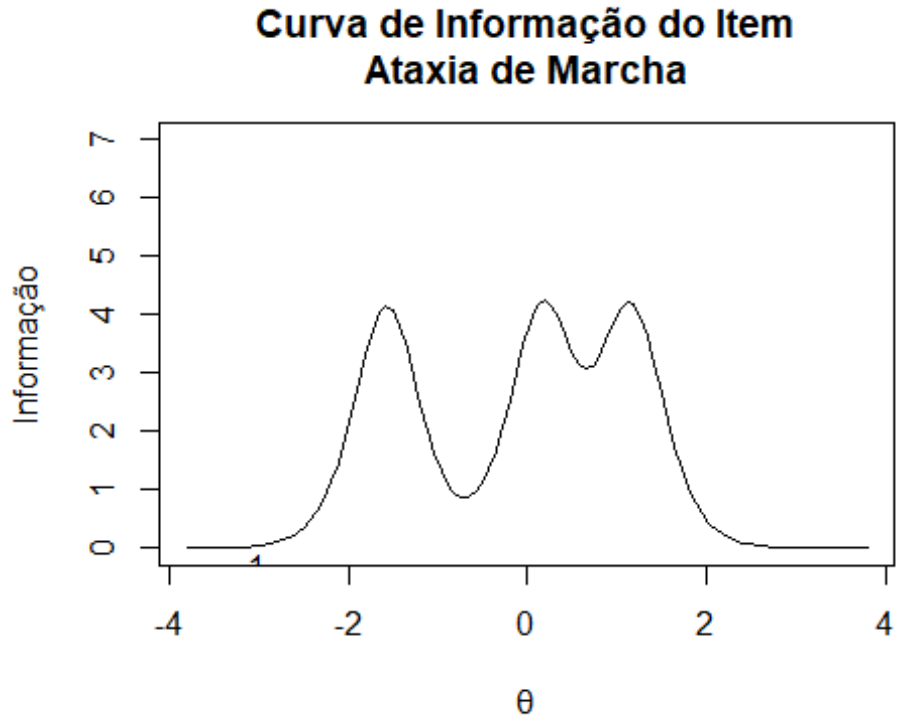

Figura 3. Curva de Informação do Item para o item Ataxia de Marcha.

2 – Ataxia nos Membros

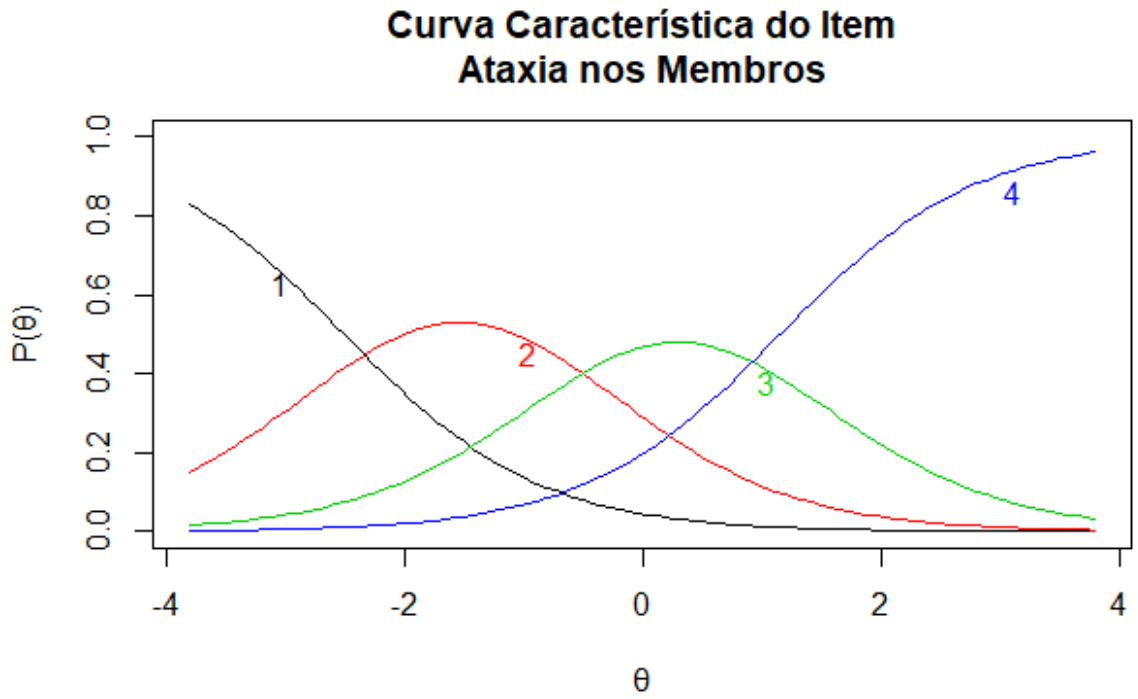

Figura 4. Curva Característica do Item para o item Ataxia nos Membros.

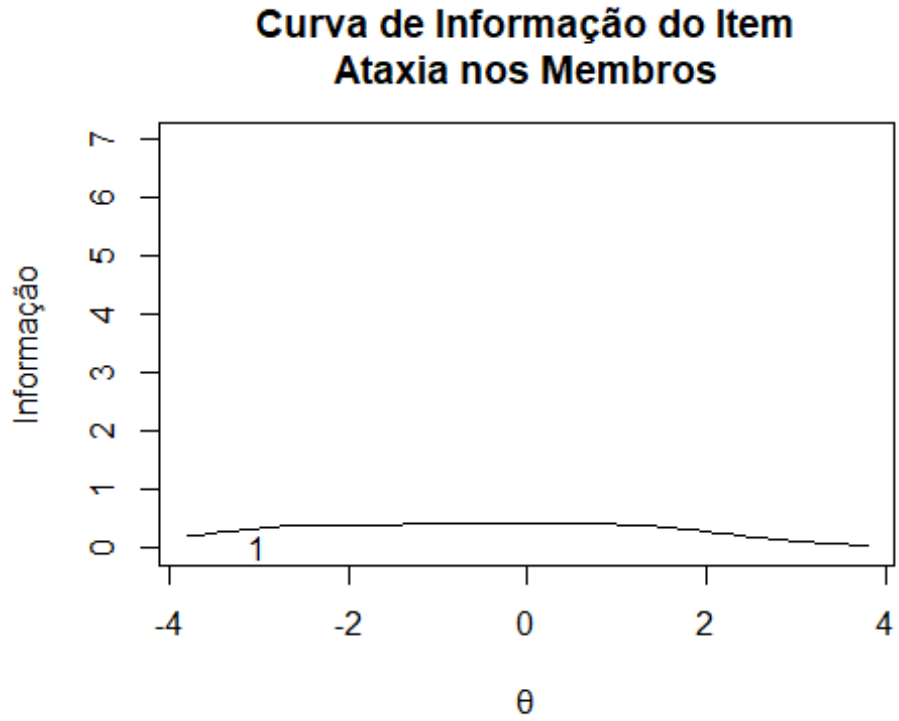

Figura 5. Curva de Informação do Item para o item Ataxia nos Membros.

3 – Nistagmo

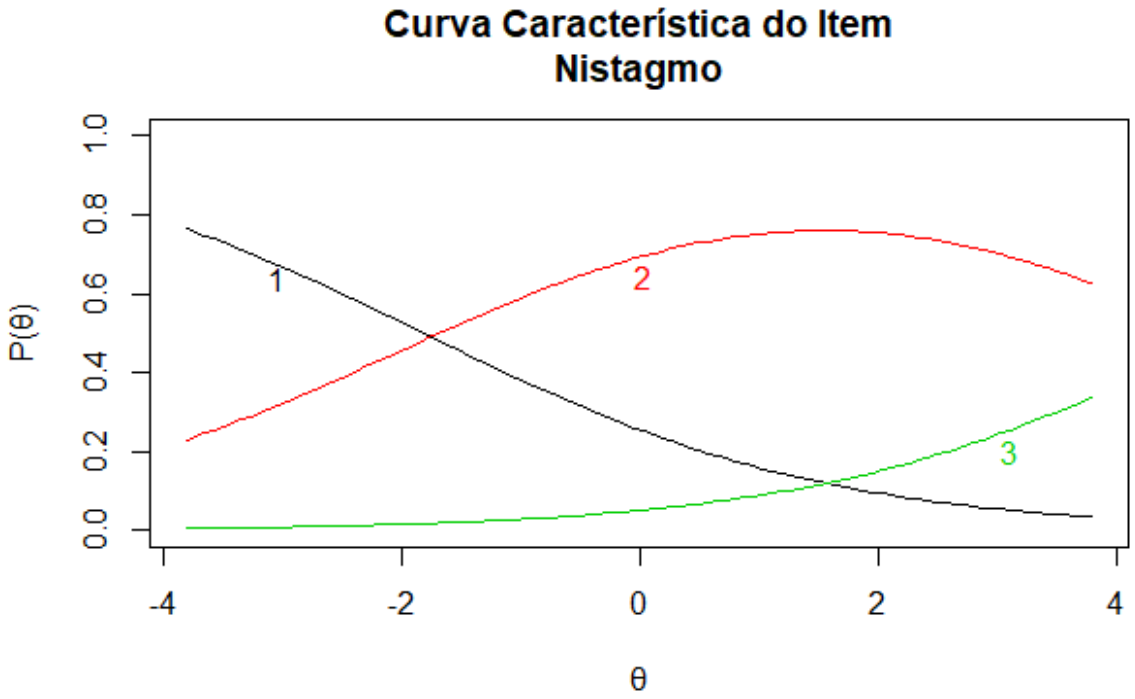

Figura 6. Curva Característica do Item para o item Nistagmo.

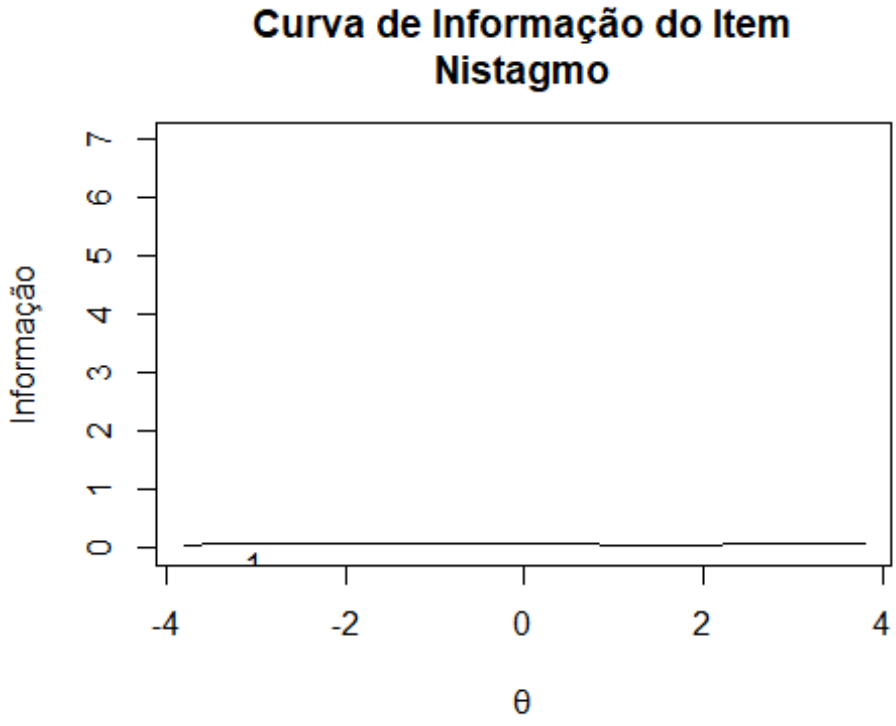

Figura 7. Curva de Informação do Item para o item Nistagmo.

4 – Disartria

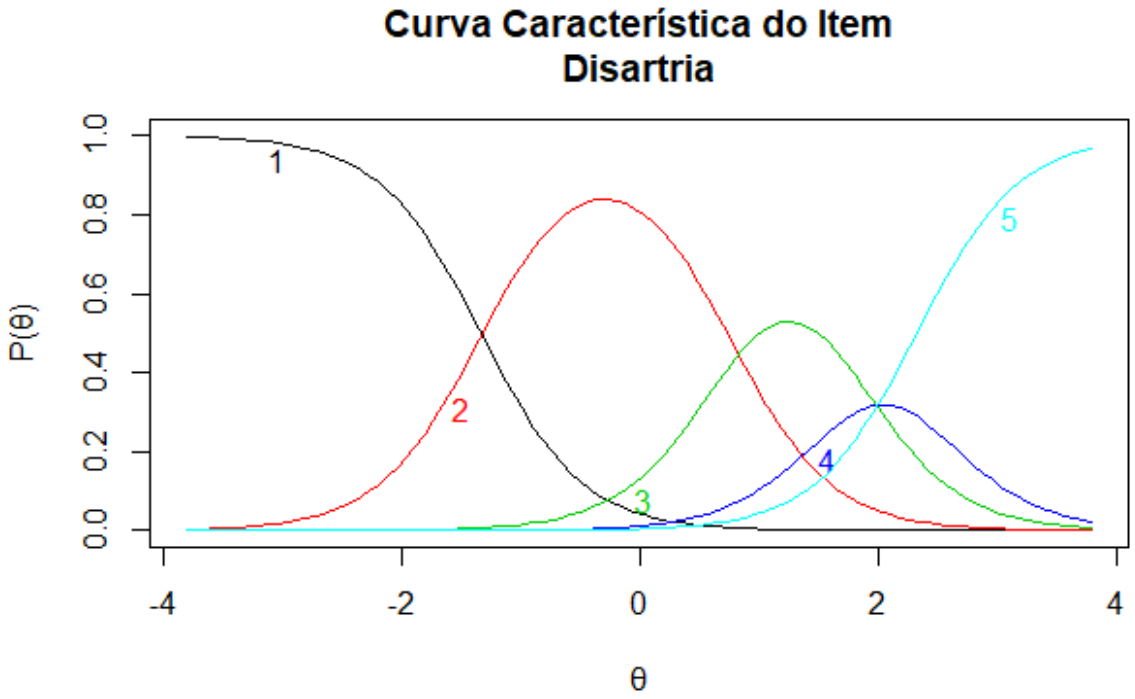

Figura 12. Curva Característica do Item para o item Disartria.

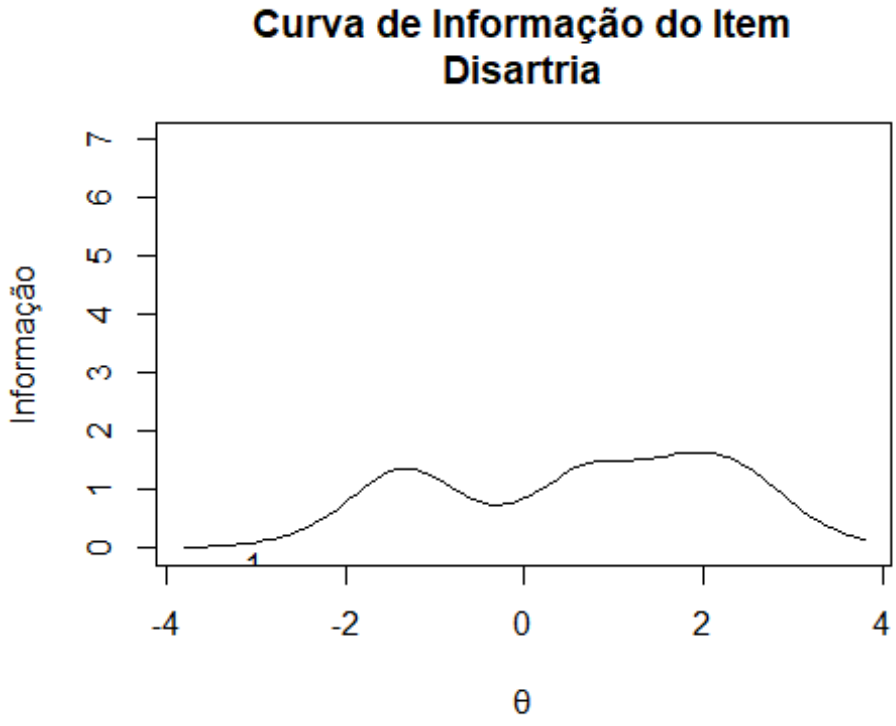

Figura 13. Curva de Informação do Item para o item Disartria.

5 – Disfagia

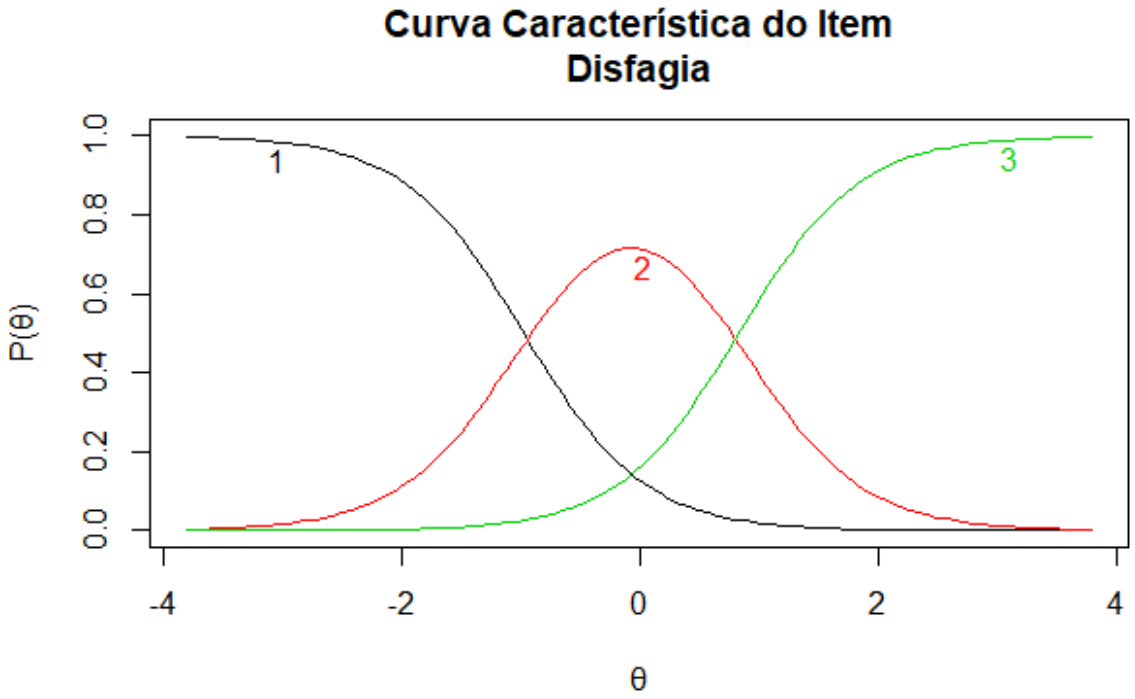

Figura 14. Curva Característica do Item para o item Disfagia.

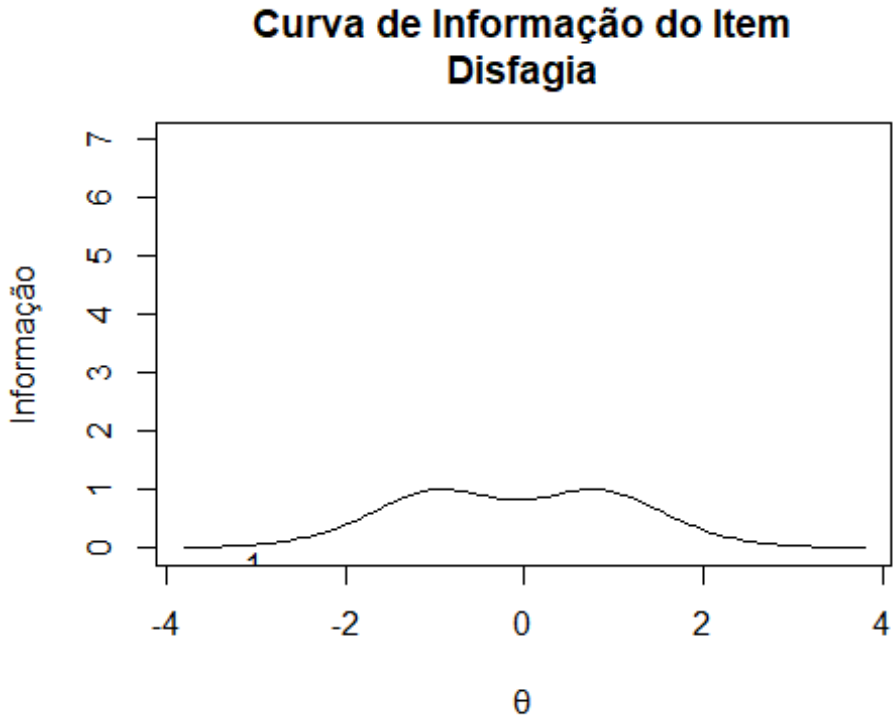

Figura 15. Curva de Informação do Item para o item Disfagia.

ANEXO F

Curvas Característica do Item e Curvas de Informação do Item para cada item da SARA

Este anexo apresenta as curvas de categoria de resposta e curvas de informação do item para cada item da SARA. Para as curvas de informação do item delimitou-se o eixo y até 7,0 para auxiliar na interpretação e comparação.

1 – Marcha

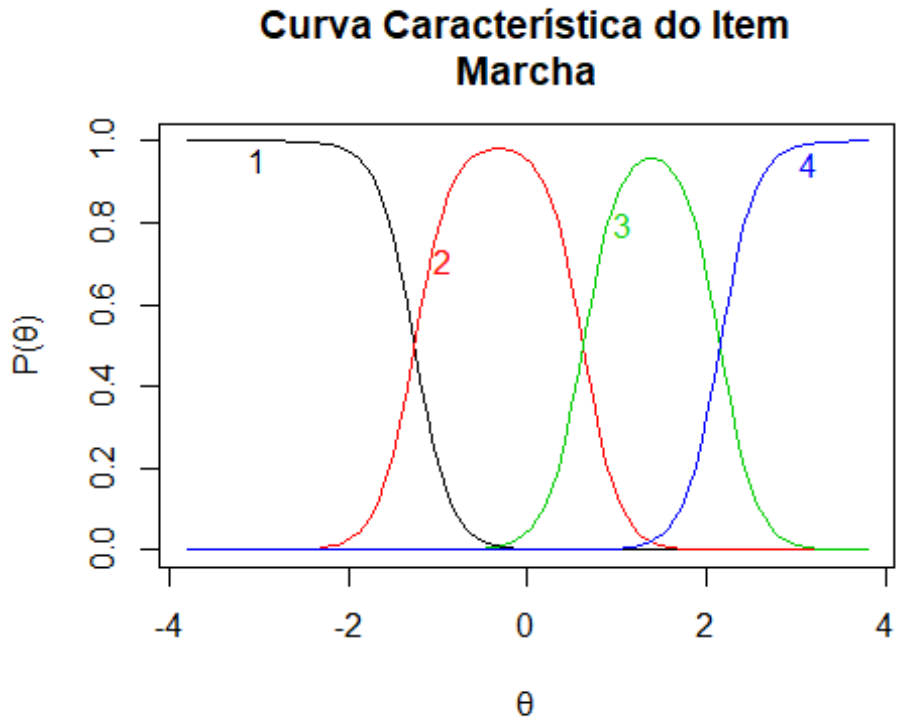

Figura 38. Curva Característica do Item para o item Marcha.

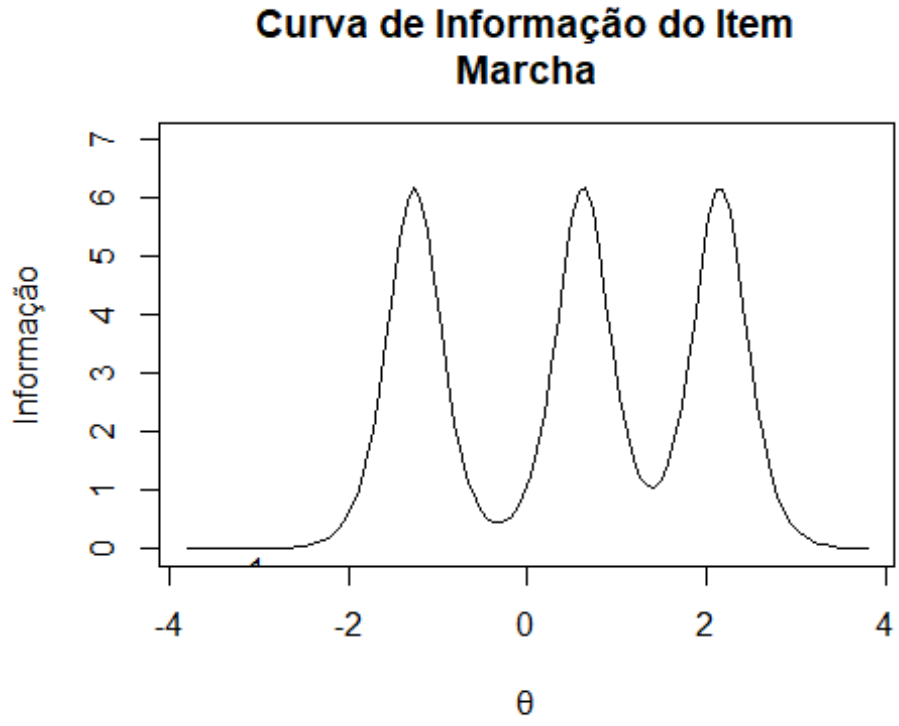

Figura 39. Curva de Informação do Item para o item Marcha.

2 – Equilíbrio de Pé

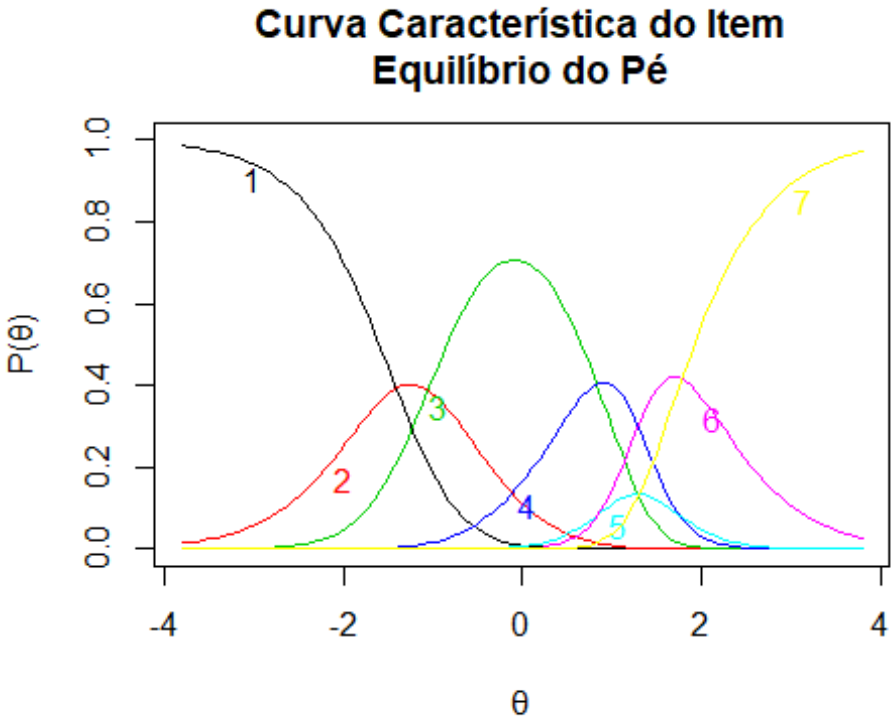

Figura 40. Curva Característica do Item para o item Equilíbrio de Pé.

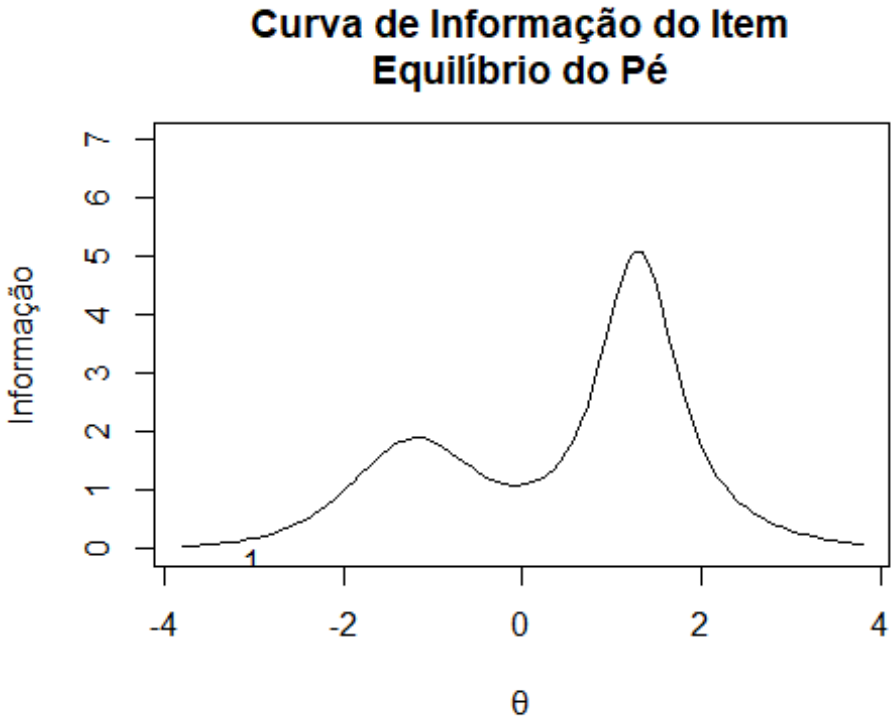

Figura 41. Curva de Informação do Item para o item Equilíbrio de Pé.

3 – Equilíbrio Sentado

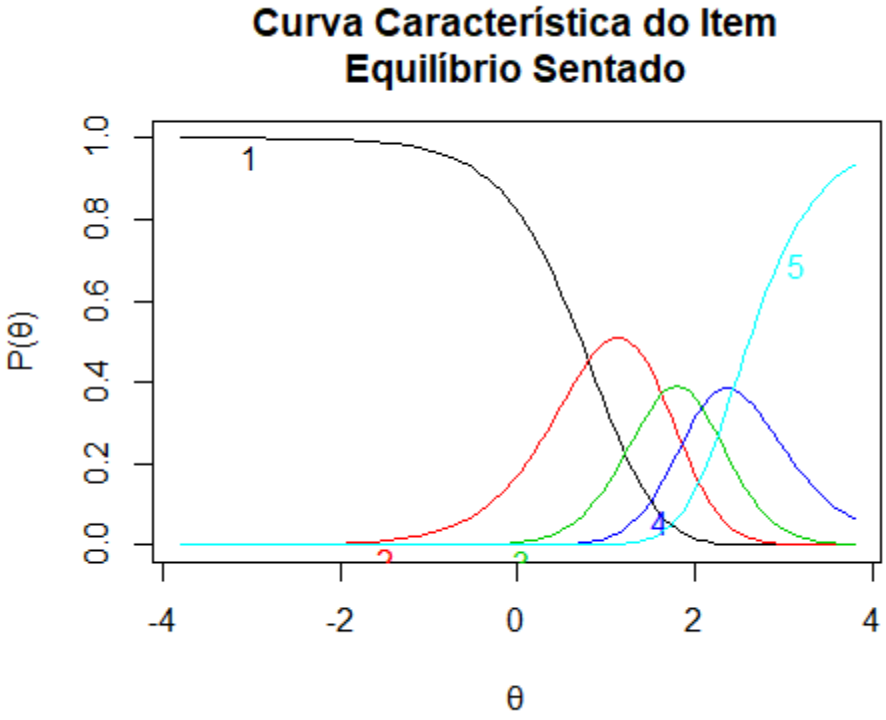

Figura 42. Curva Característica do Item para o item Equilíbrio Sentado.

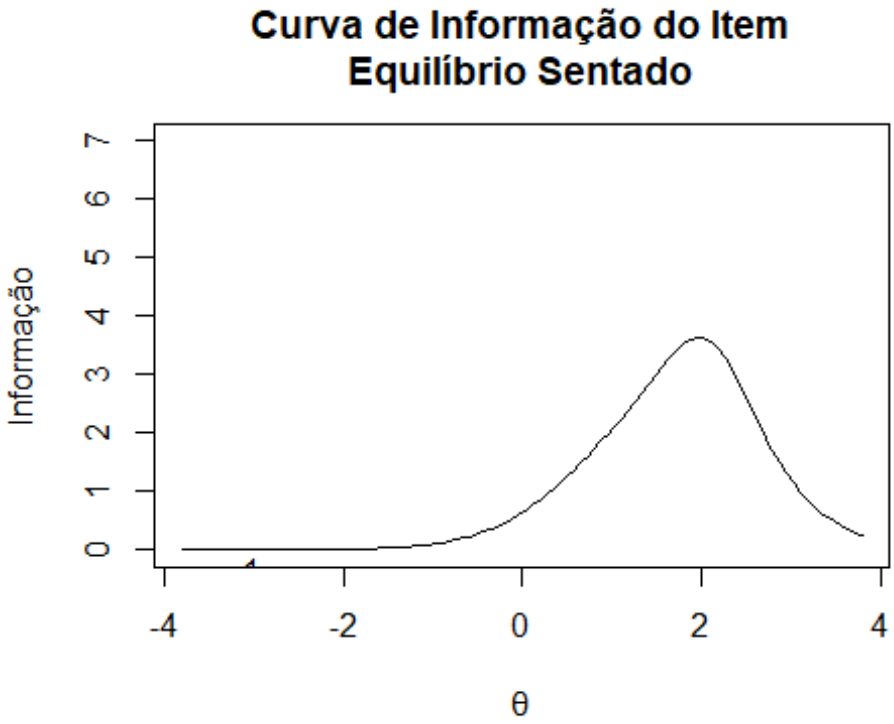

Figura 43. Curva de Informação do Item para o item Equilíbrio Sentado.

4 – Coordenação da Fala

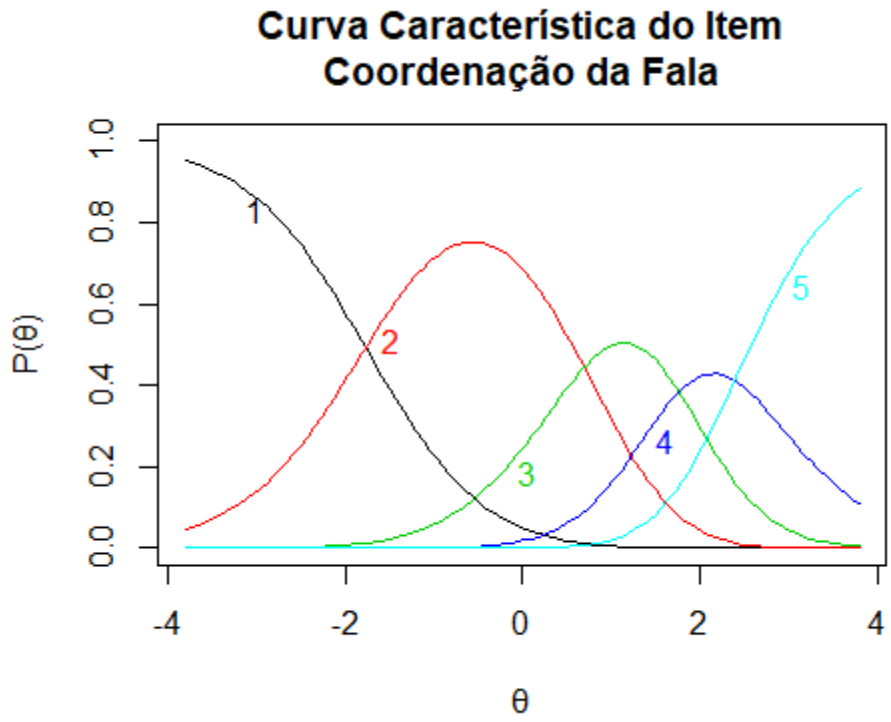

Figura 44. Curva Característica do Item para o item Coordenação da Fala.

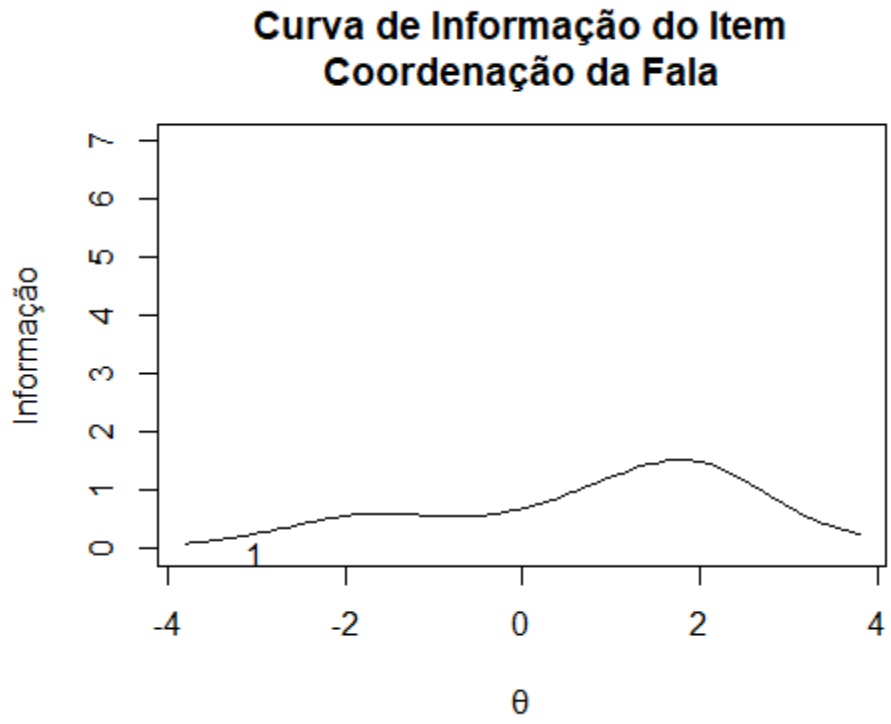

Figura 45. Curva de Informação do Item para o item Coordenação da Fala.

5 – Teste de Perseguição do Dedo

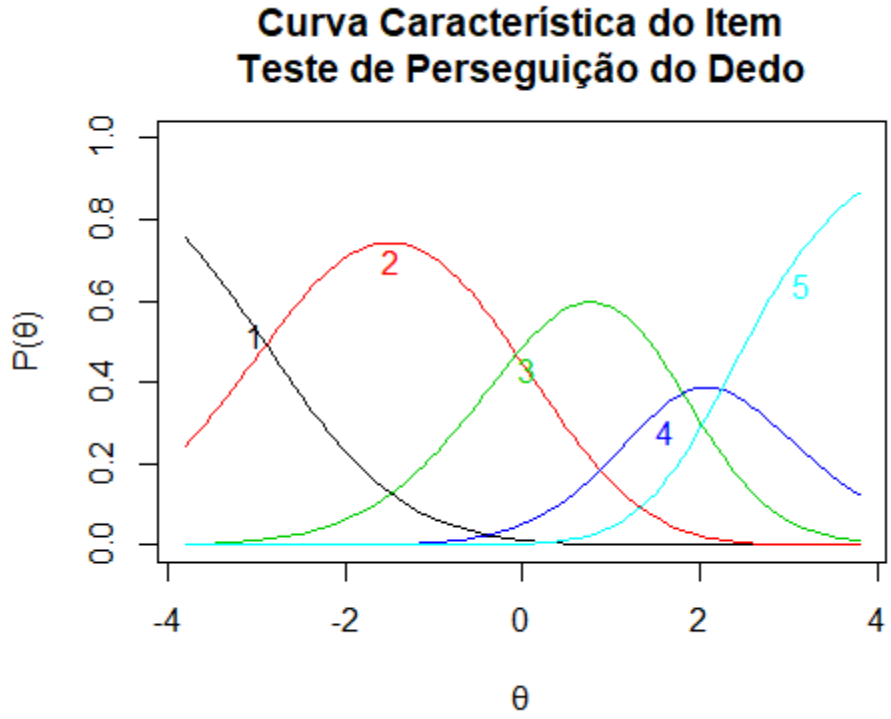

Figura 46. Curva Característica do Item para o item Teste de Perseguição do Dedo.

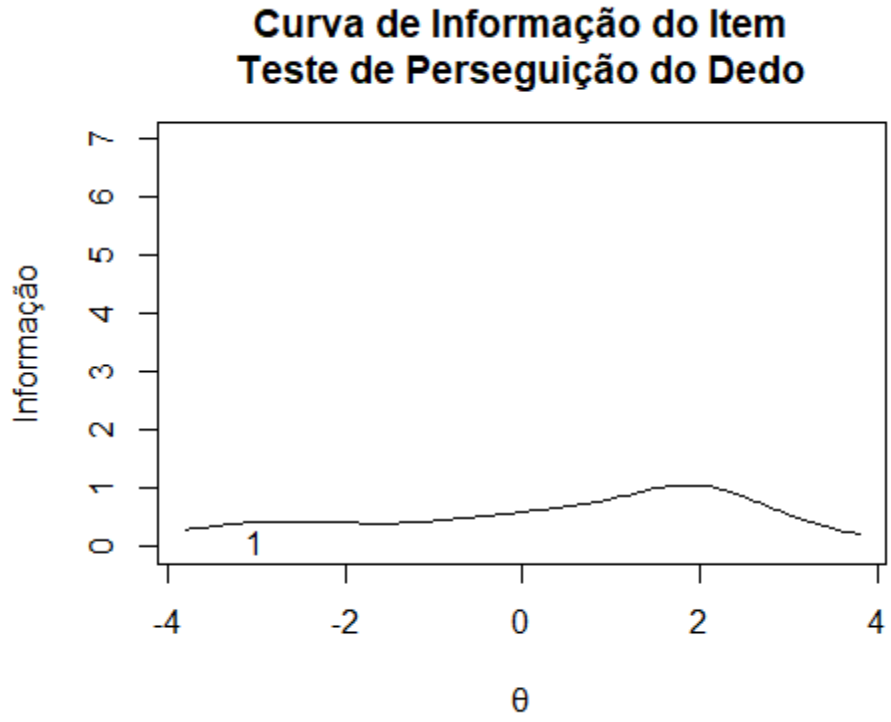

Figura 47. Curva de Informação do Item para o item Teste de Perseguição do Dedo.

6 – Teste Dedo-Nariz

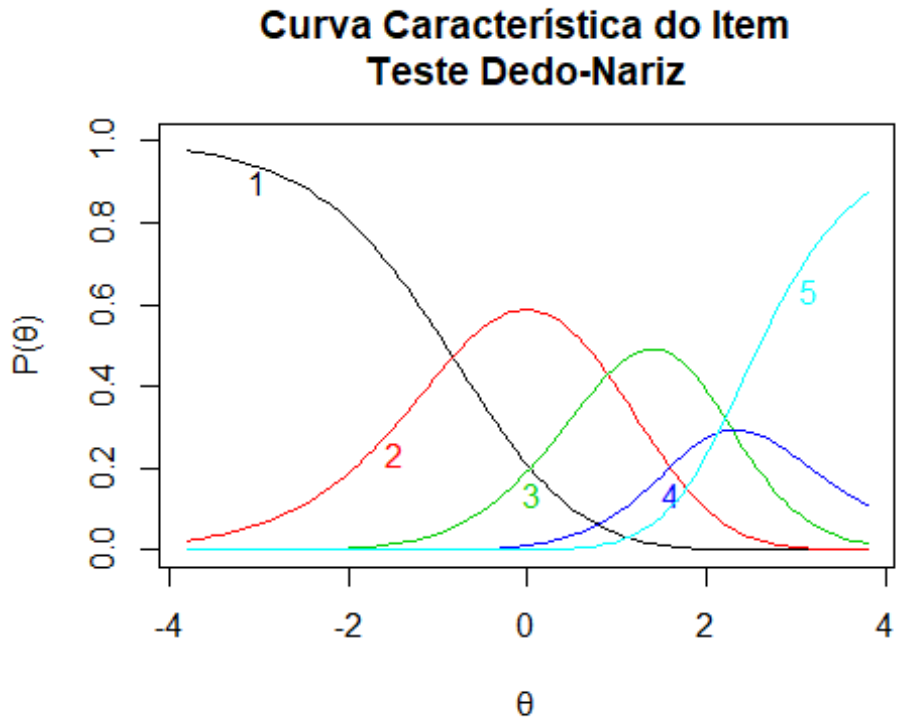

Figura 48. Curva Característica do Item para o item Teste Dedo-Nariz.

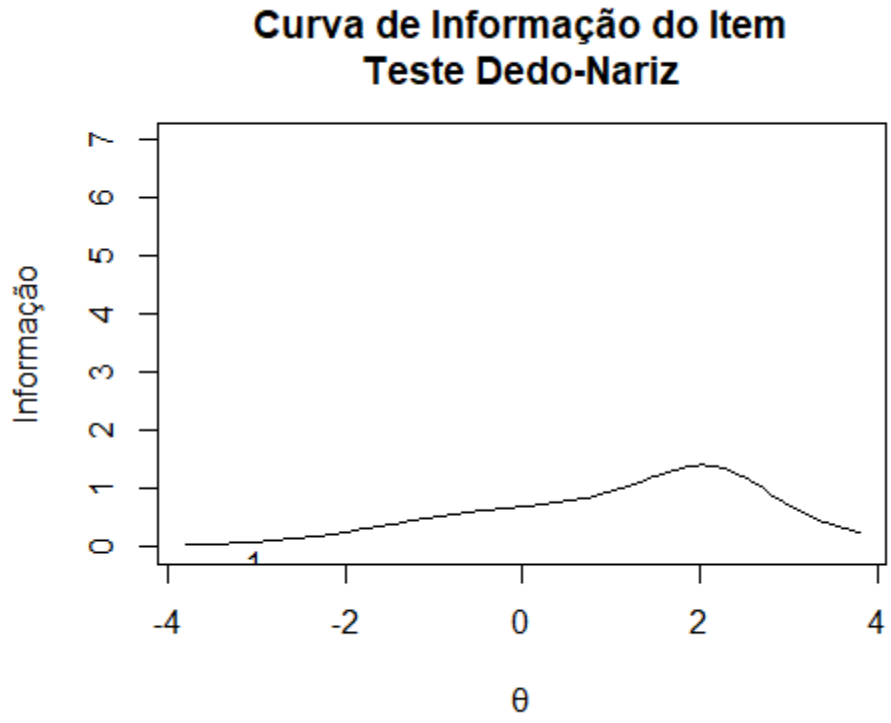

Figura 49. Curva de Informação do Item para o item Teste Dedo-Nariz.

7 – Diadococinesia

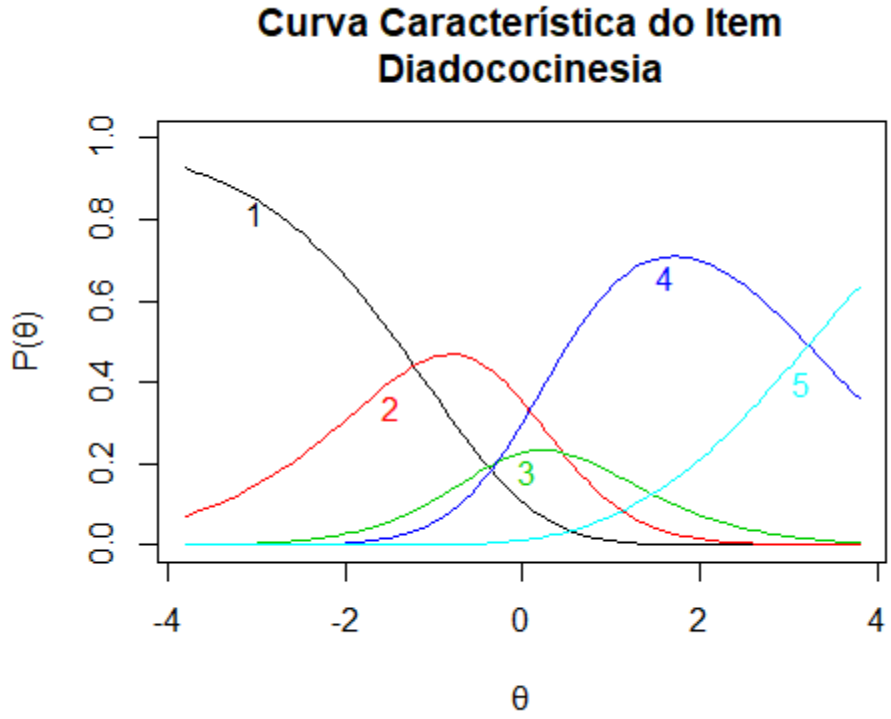

Figura 50. Curva Característica do Item para o item Diadococinesia.

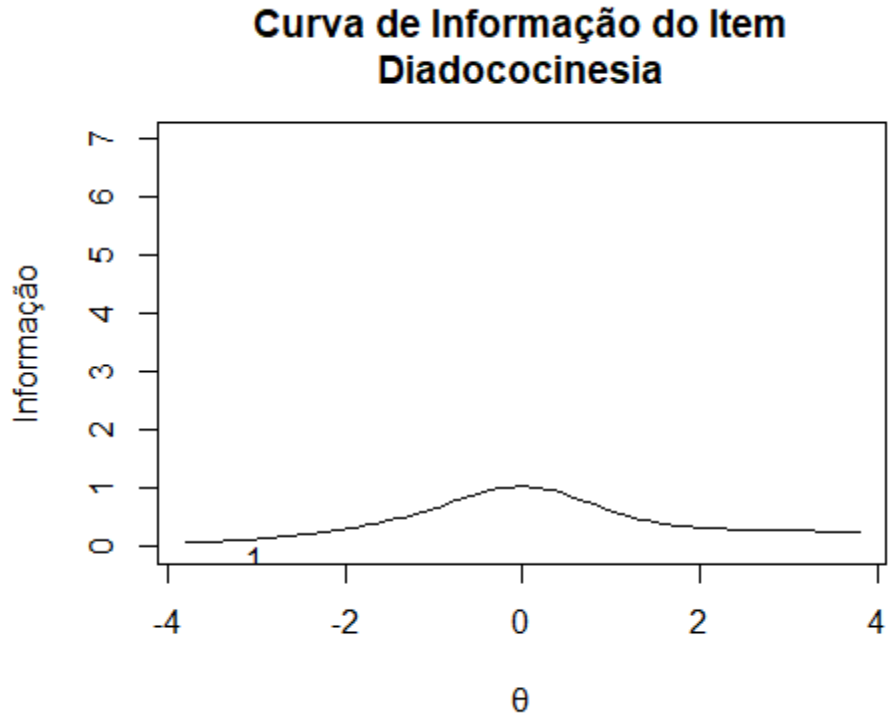

Figura 51. Curva de Informação do Item para o item Diadococinesia.

8 – Teste Calcanhar-Joelho-Canela

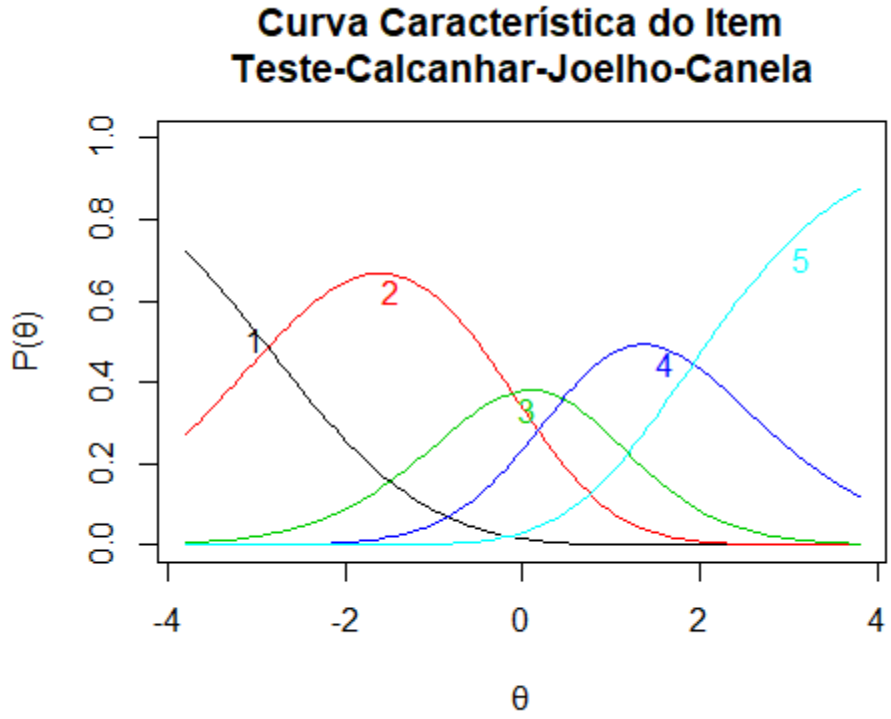

Figura 52. Curva Característica do Item para o item Teste Calcanhar-Joelho-Canela.

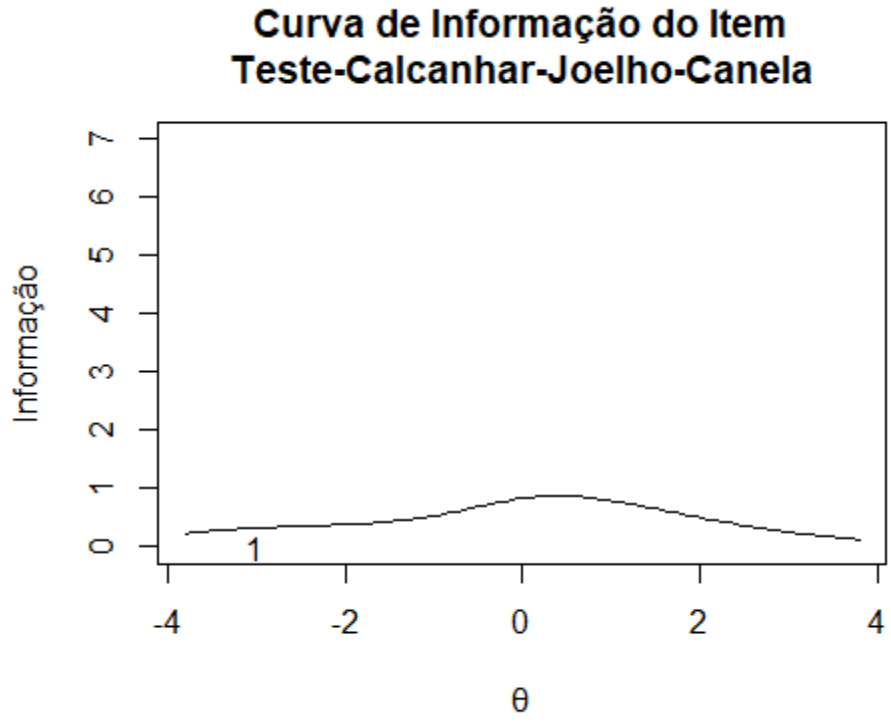

Figura 53. Curva de Informação do Item para o item Teste Calcanhar-Joelho-Canela.

ANEXO G

Resultado da transformação linear

Tabela 1. Resultado da transformação linear.

Método de transformação	$\theta_N = A\theta_S + B$	
	A	**B**
Mean/Mean	0,9803	0,2675
Mean/Sigma	1,1614	0,1951
Haebara	1,2354	0,4814
Stocking-Lord	1,1576	0,2301