KEEP CALM AND APPLY THE IRT

José Manuel Pacheco Miguel

Teoria de Resposta ao Item

Representação e utilidade do modelo logístico de traço latente na psicometria actual

Dissertação de Doutoramento em Psicologia, área de especialização em Avaliação Psicológica, apresentada à Faculdade de Psicologia e de Ciências da Educação da Universidade de Coimbra, sob orientação dos Professores Doutores José Tomás da Silva e Gerardo Prieto

Coimbra, Dezembro de 2013

Universidade de Coimbra
Faculdade de Psicologia e de Ciências da Educação

Teoria de Resposta ao Item

Representação e utilidade do modelo logístico de traço latente na psicometria actual

José Manuel Pacheco Miguel

Dissertação de Doutoramento em Psicologia, área de especialização
em Avaliação Psicológica, apresentada à Faculdade de Psicologia
e de Ciências da Educação da Universidade de Coimbra,
sob orientação dos Professores Doutores José Tomás da Silva e Gerardo Prieto

Coimbra, Dezembro de 2013

Para a Cristina, a Carolina ...
... e para a Constança

À memória de meu Pai

Agradecimentos

A presente Dissertação constitui a materialização de um projecto cuja realização não teria sido possível sem o contributo de um grupo restrito de pessoas às quais é devido um reconhecimento que a amizade concretiza no quotidiano e que a formalidade do momento permite que agora seja tornado público.

Agradecimentos são devidos, desde logo, ao Doutor José Tomás da Silva, da Universidade de Coimbra, e ao Doutor Gerardo Prieto, da Universidade de Salamanca, pela competência e pelo conhecimento que partilharam de forma generosa e com um entusiasmo contagiante. Em ambos se homenageiam, agora publicamente, as destacadas qualidades intelectuais e humanas que a modéstia pessoal dissipa, mas às quais o privilégio da amizade permite agora que se faça justiça. Umas e outro foram determinantes na consecução dos propósitos subjacentes ao projecto que agora se conclui.

São igualmente devidos agradecimentos à Doutora Maria Paula Paixão que partilha com os dois colegas anteriores as qualidades intelectuais, a competência científica e os atributos humanos. Embora sem as responsabilidades científicas inerentes à orientação do projecto de Doutoramento que agora se encerra, impõe-se realçar o inexcedível entusiasmo e o encorajamento permanente que foram uma constante durante a concretização de todo o projecto.

Impõe-se, neste momento, a manifestação pública de apreço à Doutora Lígia Mexia Leitão. Livre das responsabilidades científicas e académicas, nem por isso deixou de acompanhar este percurso de forma atenta e, não raras vezes, particularmente apreensiva. Agradece-se a presença e disponibilidade incondicionais, frutos do carinho que os laços da amizade estreitam, naqueles momentos em que as circunstâncias da existência teimam em perverter os planos previamente traçados. Bem-haja, ainda,

pela inabalável confiança e incentivo com que tem estimulado este percurso académico, cujo desafio lançou há mais de duas décadas.

Aos colegas da Faculdade de Psicologia e de Ciências da Educação da Universidade de Coimbra, agradece-se o interesse e preocupação com que acompanharam o evoluir deste projecto, bem como o apoio e incentivo prestados, factores importantes na superação do desânimo decorrente das vicissitudes do quotidiano.

Aos órgãos de gestão das Escolas Secundárias, pelo acolhimento, bem como aos professores e psicólogos que colaboraram nas tarefas inerentes ao projecto que culmina no presente trabalho é devido o agradecimento que decorre do entusiasmo e empenho com que generosamente responderam a essas solicitações. Um gesto de apreço particular a todos os alunos que, de forma anónima e voluntária, prestaram a sua colaboração na qualidade de sujeitos experimentais nos estudos empíricos que fundamentam a Dissertação.

O último agradecimento é devido aquele grupo de pessoas que abnegadamente prescindem de si em prol dos projectos pessoais que, apesar de não serem da sua autoria, adoptam altruisticamente como seus. À esposa e filha agradece-se a desculpabilização de todos aqueles contributos atitudinais capazes de legitimar novos paradigmas, no mínimo a reformulação dos existentes, da terapia familiar; à neta, agradece-se a forma generosa como, pese embora a idade, aceita a intermitência dos afectos. À mãe e aos sogros agradece-se a compreensão que só a idade consegue explicar. A todos, bem-hajam pelo monólogo quando o papel pressupunha diálogo.

Resumo

Enquadramento

O modelo de Rasch é o modelo mais simples da teoria de resposta ao item (TRI) e constitui uma abordagem potencialmente útil ao nível da construção e refinamento psicométrico de instrumentos de avaliação psicológica. Trata-se de um modelo logístico de um parâmetro da TRI no qual a quantidade de traço latente existente na pessoa e a quantidade do mesmo traço latente reflectido nos vários itens do instrumento podem ser estimados de forma independente e comparados directamente entre si, uma vez que sujeitos e itens foram medidos numa mesma métrica comum, a escala logit. Esta, definida em unidades de probabilidades logarítmicas, é uma escala intervalar na qual a unidade dos intervalos entre as localizações conjuntas pessoas-itens têm um valor ou um significado consistente. O objectivo da presente dissertação visou a apresentação das potencialidades do modelo de Rasch, comparativamente à abordagem da teoria clássica dos testes (TCT), quando aplicado ao caso específico de instrumentos de avaliação psicológica que recorrem a itens politómicos aos quais os sujeitos respondem numa escala do tipo de Likert, designado *Rating Scale Model* (RSM).

Metodologia

Para concretizar este objectivo, relativo à aplicação do Rasch RSM, foram conduzidos quatro estudos psicométricos com amostras independentes de alunos do ensino secundário. O primeiro estudo, com alunos do 12º ano de escolaridade (N = 265), foi realizado com a *Career Decision Self-Efficacy Scale – Short Form* (CDSE-SF). No segundo e terceiro estudos, com alunos dos 10º, 11º e 12º anos de escolaridade, procedeu-se à análise da *Positive and Negative Affect Schedule* (PANAS), numa abordagem TCT com recurso à análise factorial exploratória/confirmatória (AFE/

AFC) (N = 528) e à modelação Rasch (N = 519), respectivamente. Finalmente, no quarto e último estudo, também com alunos dos 10º, 11º e 12º anos de escolaridade (N = 508), fez-se a análise Rasch da *Rosenberg Self-Esteem Scale* (RSES). Em cada um dos estudos, para além das medidas dos itens, foram ainda recolhidos dados para caracterização sociodemográfica das amostras.

Resultados

Dos resultados obtidos, salientam-se aqueles que se revestem de maior importância para os objectivos estipulados. No âmbito dos três estudos realizados com o Rasch RSM, os resultados alcançados permitiram reunir evidência psicométrica que corrobora, para a CDSE-SF, a PANAS e a RSES a respectiva: (1) validade de conteúdo (i.e., o bom ajustamento dos dados ao modelo permitiu que a parametrização dos sujeitos e a calibração dos itens tenha sido feita com elevada precisão na); (2) validade estrutural (i.e., unidimensionalidade e ausência de DIF), e (3) validade substantiva (i.e., as categorias da escala de resposta funcionam de forma adequada). Os resultados do único estudo baseado na TCT, com a PANAS, revelaram dificuldade de ajustamento do modelo aos dados que só parcialmente conseguiram replicar a estrutura factorial original do instrumento.

Conclusões

Os resultados sugerem que o RSM proporciona um quadro de referência útil para o refinamento psicométrico dos instrumentos de avaliação psicológica. Da sua aplicação resulta que, no caso específico das versões Portuguesas da CDSE-SF, da PANAS e da RSES, os seus respectivos itens são representativos e relevantes para os domínios dos construtos avaliados (e.g., validade de conteúdo), têm correspondência com os construtos definidos em cada um dos instrumentos (e.g., validade estrutural) e são avaliados através de escalas de resposta cujo diagnóstico ao funcionamento empírico das respectivas categorias revelou adequação das mesmas (e.g., validade substantiva).

Palavras-chave: dimensionalidade, Análise Rasch, Rating scale model, auto-eficácia, afecto, auto-estima.

Abstract

Framework

Rasch model is the simplest Item Response Theory (ITR) model and constitutes a potentially useful approach at both the construction and refinement psychometric levels of psychological assessment instruments. It is an ITR one parameter logistic model in which the quantity of the latent trait existing in the person and the quantity of the same latent trait reflected in the various items of the instrument can be independently estimated and compared, because both items and subjects are measured using the same common metric, the logit scale. This scale, defined in logarithmic probabilities units, is an interval scale, in which the interval units between the response categories have a consistent value or significance. The goal of this dissertation is focused on the presentation of the Rasch model potentiality as compared with the classical test theory (CTT) approach when applied to the specific case of psychological assessment instruments that recur to polytomous items to which the subjects respond using a Likert type rating scale - the *Rating Scale Model* (RSM).

Methodology

In order to achieve this goal, related to the application of the Rasch RSM, four psychometric studies were carried out with independent samples of students attending secondary education. The first study, with a sample of 12th grade students ($N = 265$), was carried out using the *Career Decision Self-Efficacy Scale – Short Form* (CDSE-SF). In the second and third studies, comprising students attending respectively the 10th, 11th and 12th grades, we ran an analysis of the *Positive and Negative Affect Schedule* (PANAS), using both exploratory/confirmatory factor analysis (EFA/ CFA), and Rasch modeling ($N = 519$), respectively. Finally, in the fourth and last study, also comprising 10th, 11th and 12th grade students ($N =$

508), a Rasch analysis of the *Rosenberg Self-Esteem Scale* (RSES) was made. In each of these four studies, in addition to the items' measurements, data needed for the sociodemographic characterization of the samples were collected.

Results

From the results obtained we highlight those that are more relevant to the goals previously stated. Within the three studies carried out with the RSM the results obtained gathered psychometric evidence that confirms, concerning the CDSE-SF, the PANAS and the RSES, their respective: 1) content validity (i.e. the good adjustment of the data to the model allowed both the parameterization of the subjects and the calibration of the items to be made with high precision); 2) structural validity (i.e. unidimensionality and lack of DIF) and; 3) substantive validity (i.e. the categories of the response scale function adequately). The results of the only study which was based on CTT, and in which the PANAS was used, revealed some adjustment difficulties of the model to the data, which only partially replicated the original factorial structure of the instrument.

Conclusions

The results suggest that the RSM provides a reference framework useful for the psychometric refinement of psychological assessment instruments. In fact, from its application it can be concluded that, in the specific case of the Portuguese versions of the CDSE-SF, the PANAS and the RSES, their respective items are representative and relevant for the domains of the constructs under assessment (e.g. content validity), have correspondence with the constructs defined in each of the instruments (e.g. structural validity), and are assessed using response scales whose diagnostic to the empirical functioning of their respective categories revealed their adequacy (e.g. substantive validity).

Key-words: dimensionality, Rasch analysis, Rating scale model, self-efficacy, affect, self-esteem.

Teoria de Resposta ao Item

Representação e utilidade do modelo logístico de traço latente na psicometria actual

José Manuel Pacheco Miguel

Índice

Introdução

É consensual afirmar que o progresso científico em geral, e da Psicologia em particular, se encontra intimamente associado à medição (Wright, 1999). De facto, uma teoria científica consiste num sistema de relações entre constructos e factos empíricos em que os primeiros revestem categorias conceptuais elaboradas pelos investigadores no sentido de explicar os segundos. Nesse sentido, a descrição dos constructos psicológicos remete para a necessidade de neles incorporar aquelas características que variam entre os sujeitos e/ou entre os estímulos, circunstâncias em que o processo de medição reveste a função de representar numericamente tal variabilidade (Tabachnick & Fidell, 2007).

O propósito fundamental da psicometria consagra a formulação de modelos, bem como o estabelecimento de procedimentos analíticos, passíveis de facilitar a obtenção e a interpretação das medidas psicológicas (Nunnally & Bernstein, 1995). Enquanto processo de aquisição destas medidas, a psicometria assume uma dupla função. Uma, de índole teórica, relativa à adequação da medida para legitimar a qualidade com que esta representa a variabilidade de cada constructo. A outra, de índole prática, relacionada com a qualidade da medida em termos da sua utilidade preditiva porquanto a prática psicológica apela de forma recorrente para o uso de técnicas de medida com propósitos diagnósticos.

É por isso compreensível que apesar da premente mudança que caracteriza o mundo actual os testes psicológicos permaneçam como o estandarte da psicologia aplicada (Rust & Golombok, 1999). De facto, apesar das modificações ocorridas no âmbito das determinações legais e das aplicações específicas inerentes ao seu uso, os testes mais conhecidos e utilizados têm permanecido relativamente estáveis ao longo

das diversas revisões a que têm sido submetidos, pelo menos no contexto da realidade norte-americana (Embretson & Hershberger, 1999). Esta estabilidade não caracteriza, porém, os princípios subjacentes ao desenvolvimento dos testes, uma vez que as bases psicométricas em que assentam as provas de avaliação psicológica que têm vindo a ser desenvolvidas, para responder às necessidades contemporâneas da psicologia aplicada, se alteraram profundamente. De facto, apesar do processo de desenvolvimento dos testes ter radicado na Teoria Clássica dos Testes (TCT) ao longo de várias décadas (Kline, 1998, 2000), a Teoria de Resposta ao Item (TRI) tem vindo a assumir-se como a principal referência teórica neste domínio (Harvey & Hammer, 1999), realidade não alheia ao incremento da capacidade computacional das novas gerações de processadores informáticos, mas que nem por isso fragiliza o potencial teórico dos seus princípios de mensuração, nem tão pouco a instrumentalidade de que os mesmos se revestem para a resolução de problemáticas afins no processo sempre complexo de aquisição e interpretação de medidas psicológicas.

O modelo do traço latente, designação pela qual também é conhecida a TRI, reveste a singularidade de apresentar duas linhas autónomas de desenvolvimento, embora simultâneas (Hambleton & Swaminathan, 1985). A linha norte-americana tem início com a publicação da obra de Lord e Novick, *Statistical Theories of Mental Test Scores*, em 1968, e que inclui quatro capítulos sobre a TRI da autoria de Allan Birnbaum. A linha europeia enraíza no trabalho do matemático dinamarquês Georg Rasch cujo interesse nas propriedades científicas dos modelos de medida resultou no desenvolvimento de numa família de modelos TRI que, ao conseguirem separar totalmente os parâmetros do sujeito e do item, permitiram elaborar métodos de estimação eficazes para estas medidas. Com efeito, ao estimar o nível do traço do sujeito a partir das suas respostas aos itens do teste, os modelos TRI conseguem especificar como é que o nível do traço e simultaneamente as propriedades do item estão relacionados com as respostas dadas por um sujeito ao item.

Calcular o nível do traço no contexto de um modelo TRI, enquanto processo de procura de estimativas optimais para modelar o comportamento, faz da TRI um modelo de medida passível de facilitar a quantificação de constructos psicológicos a partir dos indicadores manifestos (Embretson & Hershberger, 1999). Com efeito, os constructos constituem a conceptualização de variáveis latentes subjacentes ao comportamento e que por isso revestem entidades inobserváveis que influenciam

as variáveis manifestas, designadamente os resultados de um teste ou as respostas a um item, as quais são os indicadores da localização dos sujeitos nas variáveis latentes (Miguel, 2006). Mas estas medidas dos constructos psicológicos, por serem indirectas, não definem completamente as variáveis latentes já que a medida destas últimas resulta precisamente da observação do comportamento em tarefas ou em itens relevantes, isto é, as propriedades das pessoas e dos itens numa dimensão psicológica são inferidas a partir do comportamento. Por isso, a teoria da medida em psicologia deve radicar numa fundamentação capaz de permitir relacionar os comportamentos com os constructos psicológicos que lhes estão subjacentes, isto é, um modelo do comportamento.

No caso da medida psicológica, o protótipo que melhor operacionaliza a relação do comportamento com os constructos consagra um modelo matemático em que variáveis independentes (VIs) são numericamente combinadas para de forma optimal permitirem predizer a variável dependente (VD). São várias as características que definem um modelo matemático em particular (Tabachnick & Fidell, 2001). Desde logo, a escala usada para medir as observações que mais não é do que a VD e cuja medida pode ser um resultado, as respostas ao item ou mesmo uma matriz de relações entre itens ou respostas. Para além disso, um modelo especifica uma ou mais VIs cuja medida, a exemplo da escala, podem ser resultados, respostas ao item ou matrizes de relações entre itens ou respostas. Finalmente, um modelo matemático especifica o modo como as VIs se combinam numericamente para predizerem a VD sendo que os coeficientes de ponderação de cada uma destas variáveis constituem os parâmetros do modelo; em alternativa a esta combinação aditiva de VIs, do tipo daquela que operacionaliza o modelo linear geral, o modelo pode especificar relações de maior complexidade entre as VIs e a VD, a exemplo das funções de distribuição probabilísticas. A TCT constitui um caso elucidativo da modelação por combinação aditiva ao postular que o resultado total no teste (VD) é estimado em função do somatório directo de duas VIs, o resultado verdadeiro no traço e o erro cometido pela pessoa aquando da ocasião em que foi avaliada. Dos pressupostos que estabelece visando controlar o erro subjacente à medida do traço, cujo cumprimento nem sempre é fácil de assegurar, decorrem duas limitações. Por um lado, as duas VIs não são realisticamente decomponíveis para um resultado individual, razão pela qual o modelo é utilizado em alternativa para justificar estimativas de estatísticas populacionais que

quando combinadas com outros pressupostos permitem fundamentar a estimação da variância verdadeira e da variância do erro. Por outro lado, ao omitir as propriedades do item no modelo, na medida em que este não considera a dificuldade e discriminação do estímulo, exige que as mesmas sejam justificadas à margem do modelo matemático pelo seu impacto nas estatísticas de variância e de fidelidade dos testes.

Em suma, a principal vantagem da TCT consiste nos pressupostos teóricos relativamente fracos que facilitam a sua aplicação em muitas situações de teste (Hambleton & Jones, 1993) e, embora o seu foco principal seja a informação ao nível do teste, as estatísticas do item (i.e., dificuldade e discriminação) também constituem uma parte importante do modelo da TCT. Ao nível do item, a TCT é relativamente simples, não invocando um modelo teórico complexo para relacionar a capacidade do sujeito avaliado com o sucesso por ele obtido num determinado item. Em vez disso, considera o conjunto dos sujeitos avaliados e analisa empiricamente a sua taxa de sucesso num item, usando-a como o índice de dificuldade do item; na realidade, trata-se de um indicador inverso da dificuldade do item, com valores mais elevados a indicarem itens mais fáceis. A capacidade de um item distinguir entre sujeitos com nível elevado/baixo de atributo medido constitui o nível de discriminação e tende a ser operacionalizado, em termos estatísticos, como o coeficiente de correlação produto-momento de Pearson entre o resultado do item e o resultado total do teste; com itens dicotómicos, esta estimativa é calculada com base no coeficiente de correlação ponto bisserial.

A principal limitação da TCT pode ser sumariada como a dependência circular; a estatística de pessoa (i.e., resultado observado) depende da amostra de itens utilizados e as estatísticas do item (i.e., dificuldade e discriminação do item) dependem da amostra de sujeitos avaliados. Esta dependência circular das estatísticas do item e do sujeito coloca dificuldades teóricas na aplicação da TCT em algumas situações de medida que os investigadores têm procurado ultrapassar, propondo abordagens empíricas baseadas em procedimentos post hoc cujo propósito visa concretizar a medição invariante do item (Engelhard, 1992). No entanto, há questões cuja abordagem continua a não ser conseguida no âmbito do enquadramento teórico da TCT (Fan, 1998).

Alternativamente, ao incluir as propriedades do item no modelo, a TRI é um exemplo paradigmático de modelação da probabilidade com que o item é escolhido/

acertado em função da respectiva dificuldade e discriminação. Associando no processo de modelação o item e o comportamento que este avalia, a moderna teoria dos testes, designação cunhada por Embretson (1996) para designar a teoria do traço latente, substitui as VIs do modelo clássico pelo nível do traço e pelas propriedades do item.

O protagonismo crescente da TRI na teorização da medida psicológica decorre do poder dos seus métodos de modelação porquanto envolvem o cumprimento de dois pressupostos, igualmente robustos (Embretson & Reise, 2000). Um prende-se com a curva característica do item (ICC[1]), também designada função de resposta ao item (IRF[2]). A forma da relação causal que se presume existir entre o constructo de interesse (e.g., nível do traço, representado por θ) e a resposta observada a cada item constitui é uma das mais importantes na TRI porquanto é a responsável pelas principais diferenças existentes entre os vários modelos TRI; representada através de um diagrama de dispersão, indica a probabilidade de resposta que seria esperada num conjunto de sujeitos caracterizados pela homogeneidade de resultados no construto que o item avalia. O outro pressuposto requer a unidimensionalidade dos itens em análise, uma vez que os modelos TRI se baseiam na assunção dos conjuntos de itens analisados serem efectivamente unidimensionais. O cumprimento deste requisito não constitui uma restrição prática *a priori* dado que as técnicas TRI tendem a ser habitualmente aplicadas com instrumentos cujas estruturas dimensionais já foram previamente estudadas de uma forma significativa através de métodos analíticos factoriais, razão pela qual as diferentes subescalas de um instrumento podem ser analisadas separadamente através da utilização de um modelo IRT unidimensional (Ackerman, 2005). Contudo, embora em termos práticos, nenhuma escala que integre um número razoável de itens se tenda a revelar verdadeiramente unidimensional (Tabachnick & Fidell, 2007), a investigação realizada para avaliar o impacto das violações do pressuposto relativo à unidimensionalidade sugerem que os modelos TRI unidimensionais se revelam relativamente robustos perante o incumprimento moderado da unidimensionalidade estrita (Drasgow, Levine & McLaughlin, 1987).

[1] Opta-se por manter o acrónimo relativo à expressão cunhada pela literatura anglo-saxónica, *Item Characteristic Curve*.

[2] Opta-se por manter o acrónimo relativo à expressão cunhada pela literatura anglo-saxónica, *Item Response Function*.

Embora todos os modelos TRI unidimensionais convirjam no pressuposto de que só um constructo latente é o principal determinante causal das respostas que se observam em cada um dos itens do teste, diferem relativamente ao modo como cada um presume θ como a causa dessa resposta. Entre os modelos mais difundidos contam-se aqueles em que as respostas ao item podem ser expressas de um modo dicotómico, sendo que as diferenças fundamentais que os caracterizam decorrem do número de parâmetros que cada um deles requer para a modelação das respostas a cada um dos itens. Em todos os casos, porém, o parâmetro ou parâmetros do item logra definir a forma da relação causal que existe entre o constructo e a resposta observada (e.g., escolha/acerto) no item, isto é, a relação entre o constructo e a probabilidade de escolha/acerto do item é presumida variar ao longo do intervalo de valores possíveis de θ enquanto função do parâmetro ou parâmetros do item (Bolt, 2005).

1. Modelos TRI unidimensionais

Assumindo que um único traço latente é suficiente para caracterizar as diferenças pessoais, estes modelos são apropriados para analisar dados nos quais um só factor comum se encontra subjacente à resposta a um item. Os modelos logísticos TRI baseiam-se na distribuição logística que expressa a probabilidade de uma resposta através de um expressão simples; se ϖ_{is} representar a combinação dos parâmetros da pessoa e do item no modelo, essa probabilidade vem dada por:

$$P(X_{is} = 1|\varpi_{is}) = \frac{e^{(\varpi_{is})}}{1 + e^{(\varpi_{is})}}$$

em que ***e*** é a base do logaritmo natural (2.718) elevada ao expoente ϖ_{is}. A probabilidade de sucesso (escolha/acerto) no item obtém-se calculando o antilogaritmo de ϖ_{is}, de acordo com a expressão anterior.

Atendendo ao número de parâmetros do item envolvidos na modelação da probabilidade com que este pode ser escolhido/acertado por uma população homogénea de sujeitos, a literatura referencia três modelos de resposta dicotómica (Ellis, 2007).

Modelo logístico de um parâmetro [Modelo 1PL]

Também designado modelo de Rasch, é o modelo mais simples e, tal como a designação pressagia, assume que basta um só parâmetro do item para representar o

processo de resposta. O parâmetro em causa é designado dificuldade e, representado pela letra β, é operacionalmente definido como o resultado no constructo que se encontra associado a 50% de probabilidade do item ser escolhido ou de receber uma resposta correcta. É de referir que o parâmetro β e o constructo θ utilizam a mesma escala (logit) em virtude do primeiro ser definido directamente em termos do segundo, facto que representa uma importante característica dos modelos TRI porquanto permite que as características do item do teste (β) e as características do sujeito com ele avaliado (θ) partilhem uma escala de medida comum. Assim sendo, num modelo 1PL todos os itens apresentam curvas características (ICC) com a mesma forma, mas que se distinguem entre si pela sua localização no contínuo do constructo, isto é, a forma da relação funcional entre o constructo e a resposta observada (e.g., forma da ICC) é constante ao longo de todos os itens, diferindo apenas no nível de θ associado à probabilidade de uma resposta de escolha/acerto.

O modelo de Rasch prediz a probabilidade de sucesso de uma pessoa *s* no item *i* (i.e., $P(X_{is} = 1)$) do seguinte modo:

$$P(X_{is} = 1|\theta_s, \beta_i) = \frac{e^{(\theta_s - \beta_i)}}{1 + e^{(\theta_s - \beta_i)}}$$

O logit desta equação $(\theta_s - \beta_i)$, que substitui ϖ_{is} na equação anterior, é dado pela diferença entre o nível do traço medido e o nível de dificuldade o item que foi utilizado para o medir.

Modelo logístico de dois parâmetros [Modelo 2PL]

Foi proposto com o propósito de resolver um inconveniente que os críticos do modelo de Rasch lhe tendem a apontar, relativo ao pressuposto segundo o qual todos os itens de um teste partilham curvas características com forma idêntica. De facto, apesar de tal se revestir de plausibilidade em conjuntos de itens cuidadosamente seleccionados a partir de pools iniciais mais alargados de itens, a sua ocorrência real revela-se rara na esmagadora maioria das situações concretas de avaliação (Bolt, 2005). Visando ultrapassar esta desvantagem, o modelo 2PL inclui um parâmetro adicional, designado discriminação e representado pela letra α, que permite que as curvas características de itens diferentes apresentem declives igualmente distintos.

$$P(X_{is} = 1|\theta_s, \beta_i, \alpha_i) = \frac{e^{[\alpha_i(\theta_s - \beta_i)]}}{1 + e^{[\alpha_i(\theta_s - \beta_i)]}}$$

Este parâmetro, representando as diferenças de discriminação dos itens, permite que estes tenham diferentes magnitudes de relação com o constructo avaliado, sendo aqui que reside a sua importância na TRI, uma vez que permite determinar de forma directa a quantidade de informação proporcionada por um determinado item. De facto, quando todos os restantes factores se igualam são os itens cujos α apresentam valores mais elevados aqueles que veiculam mais informação acerca do constructo. Neste sentido, o modelo 2PL é adequado a medidas cujos itens não se encontram igualmente relacionados com o traço latente que eles avaliam, isto é, medidas cujos itens não são igualmente indicativos da localização da pessoa nesse traço latente.

Modelo logístico de três parâmetros [Modelo 3PL]

Apesar do modelo lograr equipar o poder de discriminação dos itens do teste, ainda assim não permite explicar porque é que a forma das curvas características também pode diferir entre os itens ao nível da assimptota inferior (i.e., percentagem esperada de respostas correctas/escolhidas que são esperadas em sujeitos com valores diminutos do constructo) (Bolt, 2005). Para o efeito, o modelo 3PL complementa os parâmetros β e α com a inclusão do parâmetro γ — pseudo escolha — num esforço para reflectir o facto da assimptota inferior da curva característica do item poder assumir valores não nulos enquanto valores mínimos efectivos, situação que não se verifica nos modelos 1PL e 2PL já que nestes o valor da assimptota inferior da ICC se encontra fixado em zero.

$$P(X_{is} = 1|\theta_s, \beta_i, \alpha_i, \gamma_i) = \gamma_i + (1 - \gamma_i)\frac{e^{[\alpha_i(\theta_s - \beta_i)]}}{1 + e^{[\alpha_i(\theta_s - \beta_i)]}}$$

Um efeito inevitável que se encontra associado a este acréscimo do parâmetro γ tem a ver com a redução do poder efectivo de discriminação de um item e, por isso, de diminuição do nível de informação proporcionado por esse estímulo. Com efeito, quanto mais fácil é adivinhar a resposta correcta de um item, menos elucidativo este se torna em termos da informação útil que disponibiliza para a estimação do resultado inerente ao constructo, em termos do sujeito que responde ao teste. Para evitar estes problemas de estimação, é prática comum estimar uma assimptota inferior para todos os itens ou para grupos de itens semelhantes (Embretson & Reise, 2000).

A principal razão justificativa de um resultado diferente de zero na assimptota inferior prendeu-se, inicialmente quando os modelos TRI foram desenvolvidos no

contexto de testes de resposta dicotómica, com a necessidade de controlar os acertos devidos ao acaso sempre que tais modelos passavam a ser aplicados a testes com itens de escolha múltipla (Embretson & Reise, 2000). De facto, mesmo com grupos de sujeitos homogéneos em termos dos seus resultados no constructo em avaliação, desde que esses valores sejam extremamente baixos, é possível antecipar índices não nulos de respostas correctas a itens difíceis devido ao acaso e não à presença do constructo. Posteriormente, quando os modelos TRI passaram a ser aplicados a testes com itens de resposta não dicotómica, a necessidade de uma assimptota inferior diferente de zero manteve-se, não para ponderar as escolhas ou os acertos devidos ao acaso, mas antes para controlar a desejabilidade social de resposta ou a natureza relativamente extrema de alguns itens (Bolt, 2005).

2. Modelo de Rasch: a mensuração invariante

A medição nas ciências sociais e do comportamento define-se como o processo de localização de pessoas e de itens numa linha que representa uma variável latente, com esta a representar o construto ou atributo que a escala foi concebida para medir (Engelhard, 2008). Esta concepção da medição reflecte a ideia de que as respostas de uma pessoa a um conjunto de itens são principalmente uma função da localização da pessoa (nível de realização) e dos itens (dificuldade dos itens) na variável latente, traduzindo os pontos de vista do tipo ideal de medida nas ciências físicas.

Definindo a medição como a atribuição de números através de regras a um conjunto de pessoas ou de objectos, a chave para a mensuração invariante consiste em desenvolver um conjunto de regras e de requisitos que permitam desenvolver escalas com um conjunto de propriedades desejáveis. Com o propósito de concretizar esta tarefa, Rasch definiu os requisitos daquilo que considera o tipo ideal de escalas, regras estas que são definidas *a priori* e que, por isso, não são derivadas a partir de qualquer conjunto específico de dados. Tal como referiu Andrich (1988) Rasch apresentou dois princípios de invariância para fazer comparações que num sentido importante precedem, embora inevitavelmente conduzem, à medição: "the comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should be also independent of which stimuli within the considered class were or might also have been compared. Symmetrically, a comparison between two individuals should be independent of which particular

stimuli within the class considered were instrumental for the comparison; and it should be independent of which other individuals were also compared, on the same or on some other occasion" (Rasch, 1961, pp. 331-332).

Resulta evidente que Rasch reconhecia a importância da calibração do item ser invariante relativamente à amostra de sujeitos (itens vistos como estímulos) e da medição destes ser invariante relativamente aos itens usados para os avaliar. Este conceito constitui um ponto importante na distinção do modelo de Rasch face a outras abordagens de medida baseadas em enquadramentos estatísticos centrados no ajustamento dos modelos aos dados (Engelhard, 1992). A concepção de Rasch baseia-se numa abordagem filosófica da mensuração que salvaguarda os modelos face aos caprichos inerentes às imperfeições dos dados a modelar, já que o ajustamento é dos dados ao modelo e não do modelo aos dados (Prieto & Delgado, 2007). De facto, a abordagem psicométrica nela fundamentada permite criar itens e recolher observações que têm o potencial de cumprir com os requisitos de modelos que impõem pressupostos mais restritivos, designadamente o modelo de Rasch, no sentido de concretizar as propriedades desejáveis da mensuração invariante (Engelhard, 2008).

O modelo de Rasch permite estimar, de forma precisa e detalhada, as qualidades psicométricas de uma escala de medida, aferindo o nível do traço existente nos sujeitos que respondem aos seus itens, bem como a adequação destes ao atributo que a escala pretende medir. A calibração que faz dos itens do instrumento de medida permite aferir acerca da sua utilidade face à avaliação e interpretação das diferenças no construto, considerando-se que a escala de medida possui boas qualidades metrológicas quando, para além de cumprir com os dois princípios fundamentais da TRI (unidimensionalidade e independência local), apresenta um índice de dificuldade dos itens ajustado ao nível do traço existente nos sujeitos avaliados (Bonde & Fox, 2007; Embretson & Reise, 2000; McDonald, 1999).

As vantagens do modelo Rasch relativamente à TCT, e a outros modelos TRI, encontra-se amplamente documentada (Andrich, 1988; Bond & Fox, 2007; Embretson & Reise, 2000). Com base nelas, Prieto e Delgado (2003) elencam um conjunto de características do modelo que, em sua opinião, são as mais relevantes, designadamente, *medição conjunta* (e.g., os parâmetros das pessoas e dos itens são expressos nas mesmas unidades e localizam-se no mesmo contínuo, em função da quantidade de construto que cada um deles mede, facto que confere riqueza diagnós-

tica ao modelo, uma vez que a interpretação das pontuações individuais deixa de se basear nas normas do grupo, para em alternativa se passar a centrar na identificação dos itens em que o sujeito revela uma probabilidade elevada/baixa de resolver correctamente); *objectividade específica* (e.g., a validade e generalização da medida subentende independência face às das condições específicas em que a mesma foi obtida, ou seja, pressupõe que as comparações entre pessoas são independentes dos itens administrados e que os parâmetros dos itens não são influenciados pela distribuição amostral usada para a respectiva calibração); *especificidade do erro padrão de medida* (e.g., quantifica a precisão com que é medido cada ponto da dimensão que a escala avalia, permitindo seleccionar os itens que medem com maior precisão aspectos do atributo previamente seleccionados[3]); e *propriedades de intervalo* (e.g., a interpretação das diferenças na escala é a mesma ao longo de atributo medido, isto é, a diferenças iguais entre θ [atributo no sujeito] e β [dificuldade do item] correspondem probabilidades idênticas de uma resposta ser acertada/escolhida, razão pela qual a escala logit possui propriedades de intervalo muito importantes porque, para além de assegurarem a invariância das diferentes pontuações ao longo do contínuo da variável que a escala mede, constituem condição necessária para a utilização rigorosa de análises paramétricas frequentemente empregues no domínio das ciências sociais).

3. Escala logit

A compreensão do modelo Rasch pressupõe o conhecimento subjacente das unidades usadas na escala logit. No essencial, os logits são as unidades usadas para marcar os instrumentos de medida, definindo as unidades do modelo de Rasch. A designação resulta da abreviatura da expressão "log-odds unit", em língua inglesa, que traduzida significa unidade de probabilidades logarítmicas. Trata-se de uma transformação não linear de proporções que é usada com o propósito de criar uma escala linear com maior probabilidade de ter unidades iguais (Cramer, 2003). Esta transformação não linear é usada com o propósito de criar uma escala linear com

[3] A objectividade específica não implica que a precisão das estimações dos parâmetros seja a mesma para diferentes conjuntos de itens e de sujeitos. De facto, contrariamente à TCT que pressupõe que os testes medem todas as regiões da variável com a mesma precisão, o modelo de Rasch permite estimar com maior precisão os parâmetros dos sujeitos com nível baixo doa tributo, se os itens são fáceis; de igual modo, se os itens são difíceis, medem com maior precisão os sujeitos com nível alto do atributo.

maior possibilidade de ter unidades iguais que a escala original; em essência, a transformação logística visa proporcionar uma transformação não linear da proporção de resultados das pessoas e dos itens de modo a que os valores daí resultantes tenham uma maior probabilidade de terem unidades iguais.

Esta transformação logística produz valores que podem variar entre -∞ e +∞, embora a maioria dos valores possa ser encontrada no intervalo ente -5.00 e +5.00 logits, sendo definida do seguinte modo:

$$\textbf{Logit} = \Psi[x] = ln\left[\frac{x}{1-x}\right]$$

em que o símbolo **Ψ[x]** é a representação simbólica da transformação logística para *x* e *ln* representa o logaritmo natural.

Os logits para as pessoas podem ser definidos substituindo a proporção correcta, *p*, operacionalizada, a partir da equação anterior, como a razão entre o número de itens correctos e o número de itens para *x*, tal como consta na equação seguinte:

$$\textbf{Logit}_{\boldsymbol{Pessoa}} = \Psi[p] = ln\left[\frac{p}{1-p}\right]$$

De igual modo, no caso dos logits para os itens, estes podem ser definidos substituindo *x* na primeira equação pelo valor de ***p*** que se obtém dividindo o número de resposta correctas pelo número total de pessoas que responderam ao item (*p-value*), tal como consta na equação seguinte:

$$\textbf{Logit}_{\boldsymbol{Item}} = \Psi[p\text{–}value] = ln\left[\frac{p}{1-(p-value)}\right]$$

Estas medidas logit, quer para as pessoas quer para os itens, possuem um conjunto de características desejáveis que Wright (1993) destaca referindo que "when any pair of logit measurements have been made with respect to the same origin on the same scale, the difference between them is obtained merely by subtraction and is also in logits [...] the logit scale is unaffected by variations in the distribution of measures that have been previously made, or by which items [...] may have been used to construct and calibrate the scale. The logit scale can be made entirely independent of the particular group of items that happened to be included in a test this time, or the particular samplings of persons that happened to have been used to calibrate these items" (p. 288).

Há um conjunto alargado de razões que podem ser invocadas para justificar a transformação logística das proporções *(p-value)* em logits (Cramer, 2003). Primeiro, a transformação logística constitui uma boa aproximação à distribuição normal após divisão por uma constante (1.7). Segundo, a transformação logística integra a família das distribuições exponenciais que possuem muitas propriedades estatísticas desejáveis. Terceiro, a transformação logística pode ser concebida como representando variação intra-sujeitos aleatória relacionada com os processos de resposta latentes. Finalmente, os logits podem ser encarados uma das várias transformações possíveis das proporções que permitem criar uma escala linear.

Na modelação de Rasch, os logits podem ainda ser definidos do seguinte modo:

$$\text{Logit} = ln\left[\frac{P_{i1}}{P_{i0}}\right] = \theta - \delta_i$$

em que P_{i1} é a probabilidade condicional de pontuar 1 no item *i*, P_{i0} é a probabilidade condicional de pontuar 0 no item *i*, θ a localização da pessoa na variável latente e δ é o nível de dificuldade do item *i*. A razão P_{i1}/P_{i0} (Probabilidade de Sucesso)/(Probabilidade de Fracasso) representa a probabilidade de sucesso.

Demonstrada que ficou a importância do modelo logístico de traço latente para a teorização da medida em Psicologia, constitui propósito abrangente da presente Dissertação a aplicação dos princípios da Teoria de Resposta ao Item, em termos dos respectivos métodos e procedimentos computacionais, no âmbito da investigação e da prática psicológica em Portugal. Especificamente, e tendo em consideração as propriedades do modelo de Rasch previamente apresentadas, optou-se por centrar todo o trabalho na análise de Rasch.

O modelo de Rasch (1960) é um modelo de medida unidimensional que calcula a relação entre a dificuldade do item e a capacidade da pessoa como a razão das escolhas positivas ou negativas de um item, expressando a diferença em probabilidades logarítmicas (logits) (Embretson & Reise, 2000). Segundo o modelo, a probabilidade de uma pessoa escolher um item encontra-se logisticamente relacionada com a diferença entre o nível do construto latente presente na pessoa e o nível de dificuldade subjacente à escolha do item. Por outras palavras, os construtos latentes a serem medidos são determinados pela probabilidade dos sujeitos que respondem concordarem ou discordarem com os itens em diferentes graus de aprovação ao longo do

construto que os itens avaliam. Os dados são ajustados ao modelo através da transformação matemática dos resultados brutos dos itens em logaritmos, a que se segue a colocação simultânea dos sujeitos e das respostas por eles dadas aos itens numa mesma escala logit. Procedendo assim, as percentagens não relacionadas de itens muito e pouco escolhidos são transformadas numa escala linear que pode estimar, em termos probabilísticos, como é que os sujeitos que responderam aos itens do instrumento irão provavelmente responder quando vierem a ser confrontados com itens semelhantes numa futura implementação do instrumento.

Quando os itens do instrumento apresentam apenas duas alternativas de resposta (e.g., sim/não, verdadeiro/falso, correcto/incorrecto, etc.), utiliza-se o modelo de Rasch original, concebido pelo autor para itens dicotómicos:

$$ln\left[\frac{P_{ni}}{1-P_{ni}}\right] = B_n - \delta_i$$

em que P_{ni} é a probabilidade da pessoa n responder ao item i, B_n é o nível de capacidade da pessoa n e δ é o nível de dificuldade do item i (Wright, 1999).

Quando, em alternativa, a escala de resposta ao item inclui três ou mais categorias ordenadas (e.g., discordo totalmente, discordo, em dúvida, concordo, concordo totalmente), as respostas dos sujeitos produzem dados politómicos cuja idiossincrasia tem que ser tomada em consideração pelo modelo (Ostini & Nering, 2006). O modelo de Rasch estipula que as categorias ordenadas da escala de classificação tipo de Likert possuam intervalos iguais entre os limiares das categorias adjacentes de resposta. No caso específico em que todos os itens possuem a mesma escala com três ou mais alternativas de resposta, a modelação dos dados deverá ser realizada com o Rating Scale Model (RSM) que constitui uma extensão do modelo de Rasch para itens politómicos (Andrich, 1978), razão pela qual é designado frequentemente como o Rasch Rating Scale Model. De acordo com Linacre (2002), é dado por:

$$ln\left[\frac{P_{nik}}{P_{ni(k-1)}}\right] = B_n - D_i - F_k$$

onde P_{nik} representa a probabilidade da pessoa n responder na categoria k do item i, $P_{ni(k-1)}$ é a probabilidade da pessoa n responder na categoria k-1 do item i, B_n é a medida da pessoa n no traço avaliado, D_i é a dificuldade do item i e F_k é a dificuldade do passo da categoria k-1 para a categoria k (i.e., calibração do passo); esta ca-

libração do passo (F_k) é um limiar da escala de classificação definido como sendo a localização correspondente à equiprobabilidade de observação das categorias adjacentes *k-1* e *k*.

Tendo em consideração a singularidade dos instrumentos de avaliação psicológica que foram utilizados na presente Dissertação, todos eles com itens politómicos aos quais os sujeitos são solicitados a responder com a mesma escala de resposta, decidiu-se modelar os respectivos dados implementando o RSM. A utilidade subjacente à utilização deste modelo de Rasch na análise de escalas de atitudes justifica-se por várias razões:

— apesar dos dados obtidos a partir dos instrumentos de avaliação psicológica tenderem a ser analisados como se fossem representativos de uma escala intervalar, de facto tais dados são ordinais. A utilização de resultados brutos oriundos de dados categoriais de tipo Likert em análises de correlação pode conduzir a conclusões potencialmente erróneas (Bond & Fox, 2007; Wolfe & Smith, 2007a);

— as análises tradicionais (e.g., análise factorial exploratória e correlação) tendem a tratar cada um dos itens de um instrumento de avaliação como se todos eles contribuíssem de mesma forma para a medição do construto, independentemente de existirem alguns itens que são escolhidos mais facilmente que outros. A análise baseada no RSM permite demonstrar o peso relativo de cada um dos itens para a avaliação do construto através da utilização de mapas item-pessoa, onde se apresentam itens e pessoas na mesma escala logit (Wilson, 2005), de acordo com as estimativas de dificuldade do item;

— a análise de componentes principais (ACP) dos resíduos, controlando o factor Rasch, é útil na determinação da unidimensionalidade dos dados categoriais de tipo Likert, tais como aqueles que são obtidos através da aplicação de um instrumento de avaliação psicológica. Na ACP dos resíduos dos itens, os dados obtidos a partir de um instrumento multidimensional são analisados calculando uma solução não rodada dos resíduos dos itens que se obtém após a extracção de uma medida linear (Bond & Fox, 2007; Wright, 1996a). Os resíduos dos itens cuja variância não é explicada pelo modelo de Rasch podem apresentar uma correlação suficientemente forte ao ponto de formar factores espúrios, reduzindo assim a validade da análise factorial (Wright, 1996, p.10). Os resíduos dos itens com correlações elevadas (e.g., cargas factoriais superiores a .40) possuem variância não explicada pelo modelo Rasch, reque-

rendo por isso uma averiguação mais aprofundada para determinar a unidimensionalidade do construto (Bond & Fox, 2007; Linacre, 1998; Smith, 2004a);

— a análise Rasch proporciona indicadores de precisão, tanto para os itens do instrumento de medida como para as pessoas (e.g., respostas dos participantes no estudo), baseados no conceito estatístico de separação, com o propósito de medir, não apenas a precisão convencional das pessoas como indicação da consistência das suas respostas, mas também a precisão dos itens, para indicar como é que estes foram medidos na amostra populacional (Fisher, 1992). O índice de separação é o rácio entre a variância livre de erro e a variância observada e refere-se ao número de grupos de sujeitos que o instrumento de medida consegue distinguir na amostra populacional (Wilson, 2005; Wright, 1996b). Assim, ao explicar o erro de medida, a análise Rasch das estatísticas de precisão relativas à separação dos itens e dos sujeitos proporciona um cálculo mais preciso acerca do modo como os itens do instrumento medem a amostra populacional;

— a análise Rasch utiliza o conceito de ajuste do item, sem ficar na dependência do alfa de Cronbach, para demonstrar a qualidade dos itens medidos pelos hipotéticos construtos (Smith, 2004b). Os itens e as respostas das pessoas que não ajustam ao modelo podem ficar a dever-se a descuido na resposta ao item ou a enviesamento do item (Wolfe & Smith, 2007b). Bond e Fox (2007) também recomendam a utilização do o modelo de Rasch com dados categoriais de tipo Likert devido à capacidade com que o RSM consegue determinar o nível de escolha dos itens do instrumento de medida, bem como pelo grau de precisão com que os participantes são medidos. Através da combinação da ACP dos resíduos normalizados dos itens com a análise ao ajustamento dos itens, a modelação Rasch baseada na aplicação do RSM pode ser utilizada para satisfazer as reivindicações de generalização dos resultados dos instrumentos de medida em diferentes amostras de sujeitos (Wolfe & Smith, 2007a).

A aplicação do modelo de Rasch RSM foi feita com o propósito de proceder à validação dos instrumentos de avaliação psicológica incluídos na presente Dissertação, todas eles com escalas de resposta que produzem dados categoriais de tipo Likert. Utilizou-se, para o efeito, um dos *softwares* estatísticos especializados mais usados, o Winsteps (Linacre, 2011). A avaliação realizada com cada um dos instrumentos centrou-se nos aspectos da validade de conteúdo, da validade estrutural e da

validade substantiva descritos por Wolfe e Smith (2007b), bem como no quadro das evidências de validade de Messick (1995).

O aspecto da validade de conteúdo remete para a representatividade e relevância dos itens para o domínio do construto que o instrumento avalia. No caso presente, procedeu-se à análise da qualidade dos índices de ajuste de cada um dos itens das escalas em estudo (i.e., índices de *infit* e *outfit mean squared fit* [MNSQ] e correlações ponto-medida [r_{pm}]), indicadores que o RSM disponibiliza quando é implementado no programa Winsteps. As estatísticas de ajuste do item proporcionam uma indicação acerca do nível de concordância dos dados observados com o modelo unidimensional de Rasch (Bond & Fox, 2007). Podendo assumir valores num intervalo que varia entre zero e infinitoo MNSQ do item tem um valor esperado de 1.0. Embora tenham sido sugeridos vários intervalos para esta medida de ajuste do item (Wright & Linacre, 1994), optou-se por considerar que há ajuste dos itens sempre que os valores de MNSQ variam entre .5 e 1.5, seguindo-se a sugestão proposta por Linacre (2011); valores superiores a 1.5 indicam falta de homogeneidade com os restantes itens, enquanto que valores inferiores a .5 representam redundância face aos outros itens da escala. Analisou-se ainda a correlação ponto-medida (r_{pm}) porque valores nulos ou negativos desta medida indicam sentido invertido da escala de resposta aos itens (Smith, 2004).

O aspecto da validade estrutural refere-se à correspondência dos itens com o construto definido, sendo a sua avaliação realizada com recurso a técnicas de análise factorial. Nesse sentido, foram calculadas ACPsdos resíduos normalizados, após controlar a dimensão Rasch (Linacre, 2011). A ACP, decompondo a matriz de correlações dos itens baseadas nos resíduos normalizados, permite identificar as possíveis dimensões que possam estar latentes no padrão das respostas. Se os dados ajustarem bem ao modelo unidimensional de Rasch, os valores dos resíduos deverão representar ruído aleatório e a matriz de correlação dos resíduos deverá aproximar-se de zero (Linacre, 2011; Smith, 2004). Na ausência de consenso sobre os critérios a aplicar para identificar a existência de uma dimensão secundária quando se usam resíduos normalizados baseados na PCA, optou-se por aplicar o critério proposto por Smith e Miao (1994) e por Raîche (2005), resultante das conclusões dos seus estudos de simulação, que propõem valores de *eiganvalue* no intervalo [1.4—2.0] para o primeiro componente como indicação de unidimensionalidade.

Finalmente, o aspecto da validade substantiva refere-se ao diagnóstico do funcionamento empírico das categorias da escala de resposta, com o propósito de determinar se estas funcionam em conformidade com aquilo que era esperado pelos autores dos instrumentos quando desenvolveram os respectivos itens (Linacre, 1999, 2002). Para o efeito, procedeu-se a análise do funcionamento da escala de classificação e do nível de ajuste do sujeito em cada um dos instrumentos de avaliação psicológica que foram utilizados no âmbito da elaboração ada presente Dissertação. No enquadramento conceptual do RSM, Linacre (2004) propôs um conjunto de critérios para determinar a eficiência das categorias da escala de resposta ao item: (a) distribuição uniforme das frequências das respostas pelas diferentes categorias, com um mínimo de 10 observações em cada uma; (b) progressão monotónica da medida média observada — B_n — e das calibrações dos passos — F_k — ao longo das categorias de resposta; (c) *outfit* MNSQ, mais sensível a respostas não esperadas que o *infit*, das categorias de resposta inferior a 2.0; e (d) os limiares das categorias de resposta deverão aumentar à medida que vão aumentando as categorias de resposta.

A prática da avaliação psicológica, reconhecidamente uma tarefa básica no exercício profissional da Psicologia, depende da qualidade dos instrumentos de avaliação utilizados, a qual, por seu turno, é indissociável do conhecimento da informação relativa aos respectivos estudos psicométricos centrados na análise de itens, na precisão, na validade e na produção de normas (Simões, Almeida e Gonçalves, 1999). A precisão informa acerca do grau em que os resultados de um instrumento se encontram livres de erros sistemáticos de medida; se estes tiverem presentes, a precisão é reduzida e a possibilidade de generalização dos resultados torna-se questionável (Abad, Olea, Ponsoda & García, 2011). A validade, não sendo uma propriedade do instrumento em si e antes a qualidade com que os seus resultados reflectem de modo consistente o construto que ele pretende medir (Messick, 1995), refere-se à natureza apropriada, ao significado e à utilidade das inferências formuladas a partir dos resultados de um instrumento, constituindo o critério mais importante na análise da qualidade de uma prova de avaliação psicológica (Simões et al., 1999). De facto, segundo os autores, remete para questões de natureza técnica, relativas à análise das propriedades e evidências empíricas, e ética, em termos das consequências pessoais e sociais da avaliação. Por isso, e ainda segundo eles, o estudo psicométrico dos instrumentos de avaliação psicológica é indissociável das preocupações éticas e deontológicas.

A TRI em geral, e a modelação Rasch em particular, pese embora o nível de sofisticação matemática exigido, proporcionam um quadro de referência potencialmente útil para o refinamento psicométrico dos instrumentos de avaliação psicológica, para além de introduzirem novas regras de medida (e.g., obtém-se propriedades de escala de intervalo através da aplicação de modelo de medida justificáveis; o erro padrão da medida difere relativamente aos vários resultados, ou padrões de resposta, mas pode ser generalizado a várias populações; testes com menos itens podem ser mais precisos do que testes mais longos; estimativas não enviesadas das propriedades dos itens podem ser obtidas a partir de amostras não representativas) (Embretson & Reise, 2000). Em reposta às preocupações de Simões et al. (1999) e com o propósito de expor as potencialidades da modelação Rasch, optou-se por conduzir estudos das qualidades metrológicas de instrumentos de avaliação cuja utilização é frequente no domínio da prática psicológica com adolescentes do ensino secundário. Relativamente a cada um deles o Rasch RSM foi utlizado para conduzir análises sobre a qualidade dos itens e, uma vez demonstrado o ajustamento dos respectivos dados ao modelo (*infit*/*outfit* para pessoas e itens), para reunir informação relativa à respectiva precisão (*standard error*, para cada item; *item reliability separation index*; e *person reliability separation index*, este último um índice homólogo ao coeficiente alfa de Cronbach da TCT) e validade (*differencial item functioning* [DIF][4]; dimensionalidade; e funcionamento empírico da escala de resposta). Tendo em consideração a natureza exploratória dos estudos agora realizados, reserva-se a apresentação das normas relativas a cada um dos instrumentos para trabalhos de investigação ulteriores com amostras representativas da população portuguesa e de outros escalões etários. No Capítulo 1 da presente dissertação apresenta-se o estudo Rasch da versão Portuguesa da *Career Decison Self-Efficacy Scale — Short Form* (CDSE-SF; Betz, Klein & Tailor, 1996) com uma amostra de alunos do ensino secundário. O Capítulo 2 e o Capítulo 3 remetem para o estudo da versão Portuguesa *da Positive and Negative Affect Schedule* (PANAS; Watson, Clark & Tellegen, 1988), respecti-

[4] O modelo de Rasch examina o princípio da invariância que é um dos princípios fundamentais da TRI, através da análise do funcionamento diferencial dos itens no qual se assume que o desempenho do sujeito depende apenas das suas diferenças em função do construto medido, determinadas pelo nível de dificuldade dos itens, e não das características específicas do grupo a que a pessoa pertence (e.g., género, idade, entre outras).

vamente com o modelo de Rasch e com a abordagem da TCT, em duas amostras independentes de adolescentes do ensino secundário (10º, 11º e 12º anos de escolaridade). Finalmente, no Capítulo 4, que antecede a Conclusão/Discussão, procede-se à apresentação do estudo Rasch da versão Portuguesa da *Rosenberg Self-Esteem Scale* (RSES; Rosenberg, 1965).

References

Abad, F. J., Olea, J., Ponsoda, V., & García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Editorial Síntesis.

Ackerman, T. A. (2005). Multidimensional Item Response Theory Modeling. In A.M. Olivares & J. McArdle (Eds.), *Contemporary Psychometrics* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Andrich, D. A. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, *38*, 665-680.

Andrich, D. A. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Betz, N. E., Klein, K. L., & Taylor, K. M. (1996). Evaluation of a short-form of the Career Decision-Making Self-Efficacy Scale. *Journal of Career Assessment*, *4*, 47-57.

Bolt, D. M. (2005). Limited- and Full-Information Estimation of Item Response Theory Models. In A.M. Olivares & McArdle (Eds.), *Contemporary Psychometrics* (pp. 27-71). Mahwah, NJ: Lawrence Erlbaum Associates.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cramer, J. S. (2003). *Logit models from economics and other fields*. Cambridge: Cambridge University Press.

Drasgow, F., Levine, M. V. & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*, 59-79.

Ellis, M. V. (2007). *Item Response Theory: Introduction to the Theory and Methods of IRT*. Workshop realizado no Congresso Counseling Psychology at the Crossroads: Current research and future directions. Coimbra, 29-31 de Outubro.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*, 341-349.

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

Embretson, S. E. & Hershberger, S. L. (1999). *The New Rules of Measurement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*, 155-189.

Engelhard, G., Jr. (1992). Historical views of invariance: Evidence from the Measurement Theories of Thorndike, Thurstone, and Rasch. *Educational and Psychological Measurement*, *52*, 275-291.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*, 357-381.

Fisher, W. P., Jr. (1992). Reliability statistics. *Rasch Measurement Transactions*, *6*, 238.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*, 38-47.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston: Kleuwer.

Harvey, R. J. & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist*, *27*, 353-383.

Kline, P. (1998). *The New Psychometrics: Science, psychology and measurement*. London: Routledge.

Kline, P. (2000). *A Psychometric Primer*. London: Free Association Books.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-types work best? *Journal of Outcome Measurement*, *2*, 266-283.

Linacre, J. M. (1999). Investigating rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85-106.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258-278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2011). *Winsteps Rasch measurement computer program, version 3.73.0* [*computer program*]. Chicago, Il: Winsteps.com.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Messick, M. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.

Miguel, J.P. (2006). *Inventário de Interesses Vocacionais de Amesterdão: Adaptação e validação no ensino superior* [Dissertação de Mestrado]. Coimbra: FPCEUC.

Nunnally, J. C. & Bernstein, I. J. (1995). *Teoría Psicométrica* (2ª ed.) [Psychometric Theory (3rd ed.)]. Mexico: McGraw-Hill/Interamericana.

Ostini, R. & Nering, M. L. (2006). *Polytomous item response models*. Thousand Oaks, CA: Sage.

Prieto, G. & Delgado, A. R. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema*, *15*, 94-100.

Prieto, G., & Delgado, A. R. (2008). Measuring math anxiety (in Spanish) with the Rasch Rating Scale Model. *Journal of Applied Measurement*, *8*, 149-160.

Raîch, G. (2005). Critical eigenvalues sizes in standardized principal component analysis. *Rasch Measurement Transactions*, *19*, 1012.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Rasch, G. (1961). On general laws and meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley Symposium on mathematical statistics and probability* (pp. 321-333). Berkeley: University of California Press.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rust, J. & Golombok, S. (1999). *Modern Psychometrics: The science of psychological assessment* (2nd ed.). London: Routledge.

Simões, M. R., Almeida, L. S., & Gonçalves, M. M. (1999). Testes e provas psicológicas em Portugal: Roteiro de alguma questões que atravessam a utilização de instrumentos de/na avaliação psicológica. In M. R. Simões, M. M. Gonçalves, & L. S. Almeida (Eds.), *Testes e provas psicológicas em Portugal* (Vol. 2, pp. 1-12). Braga: APPORT/SHO.

Smith, E. V. (2004a). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 575-600). Maple Grove, MN: JAM Press.

Smith, E. V. (2004b). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measure perspective. In E. V. Smith, Jr. & R. M. Smith (Eds.), Introduction to *Rasch measurement* (pp. 93-122). Maple Grove, MN: JAM Press.

Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith, Jr. & R. M. Smith (Eds.), Introduction to *Rasch measurement* (pp. 73-92). Maple Grove, MN: JAM Press.

Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective measurement: theory into practice* (Vol. 2, pp. 316-327). Norwood, NJ: Ablex.

Tabachnick, B.G. & Fidell, L.S. (2001). *Computer-Assisted Research Design and Analysis*. Needham Heights, MA: Allyn & Bacon.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Boston, MA: Pearson Allyn & Bacon.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063-1070.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wolfe, E. W., & Smith, E. V., Jr. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I — Instrument development tools. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 202-242). Maple Grove, MN: JAM Press.

Wolfe, E. W., & Smith, E. V., Jr. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II — Validation activities. In E. V. Smith, Jr. & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 243-290). Maple Grove, MN: JAM Press.

Wright, B. D. (1993). Logits? *Rasch Measurement Transactions*, 7, 288.

Wright, B. D. (1996a). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, 3, 3-24.

Wright, B. D. (1996b). Reliability and separation. *Rasch Measurement Transactions*, 9, 472.

Wright, B. D. (1999). Fundamental Measurement for Psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The New Rules of Measurement* (pp. 65-104). Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean fit values. *Rasch Measurement Transactions*, *8*, 370.

Career Decision Self-Efficacy Scale — Short Form: A Rasch Analysis of the Portuguese Version

Abstract

The present study analyzes the psychometric properties of the Career Decision Self-Efficacy Scale-Short Form (CDSE-SF) in a sample of Portuguese secondary education students using the Rasch model. The results indicate that the 25 items of the CDSE-SF are well fitted to a latent unidimensional structure, as required by Rasch modeling. The response scale, containing 5 categories, showed proper functioning; therefore, the people and item parameters could be estimated with high precision (.89 and .97, respectively). Differential item functioning (DIF) analyses confirmed that there were no differences in the results of the CDSE-SF concerning gender. Finally, psychometric implications derived from the results of the present study are discussed, and suggestions are provided for future investigations.

Keywords: Career Decision Self-Efficacy Scale-Short Form; dimensionality; Rasch analysis; rating scale model

Published

Miguel, J. P., Silva, J. T. & Prieto, G. (2013). Career Decision Self-Efficacy Scale — Short Form: A Rasch analysis of the Portuguese version. *Journal of Vocational Behavior, 82,* 116–123.

Career Decision Self-Efficacy Scale — Short Form: A Rasch Analysis of the Portuguese Version

1. Introduction

The concept of self-efficacy in Bandura's social cognitive theory (1977, 1997) was introduced in a study of vocational behavior conducted by Hackett and Betz (1981) with the purpose of understanding and clarifying women's career choices and trajectories. In this foundational study, the authors outlined this concept's potentials and proposed its generalization to other groups of subjects and other aspects of vocational development. From that study, the concept of career self-efficacy has been applied to multiple areas of academic and professional development and different populations and groups (Betz, 2007; Hackett & Betz, 1995; Lent & Hackett, 1987). The concept has attracted the attention of many researchers (e.g., Betz & Luzzo, 1996) since its original definition by Taylor and Betz (1983) as an individual's belief that he or she is capable of successfully completing tasks and specific behaviors required in career decision making.

Career decision self-efficacy is based on two well-established psychological theories: one developed in the disciplines of social and personality psychology (the theory of self-efficacy) and the other originating from vocational psychology (the theory of career maturity). Taylor and Betz (1983) performed a synthesis of the two theoretical perspectives and constructed the Career Decision Self-Efficacy Scale (CDSE) based on this conceptualization. The CDSE assesses the expectations of self-efficacy in the domain of behaviors relevant to the process of career decisions. The construct is defined based on the behavioral indicators that characterize the five areas of competency for making career choices outlined in Crites' hierarchical model of ca-

reer maturity (1978). These five areas include accurate self-appraisal, gathering occupational information, goal selection, making plans for the future and problem solving, which constitute the set of subscales of the CDSE. Each of these five subscales was originally composed of 10 statements that described the tasks necessary for career decision-making, for a total of 50 items that comprise the global scale.

Betz, Klein and Taylor (1996) subsequently developed a short version of this measurement tool with the purpose of facilitating its use in applied and research contexts. In their review, the authors maintained the original theoretical structure but eliminated five items from each subscale. The Career Decision Self-Efficacy Scale-Short Form (CDSE-SF) has 25 items divided by the identical five scales of the long version. The answers are initially obtained through a continuum with 10 levels, ranging from 1 = not at all confident to 10 = totally confident. However, Betz, Hammond and Multon (2005) have more recently proposed shortening the response scale to five levels, ranging from 1 = not at all confident to 5 = totally confident. Betz et al. (2005) concluded that this change did not affect the psychometric quality of the short version of the scale, either in terms of the accuracy of the responses (i.e., internal consistency) or the estimates of convergent and discriminant validity (i.e., construct validity) of the results.

In terms of internal consistency (measured through Cronbach's alpha coefficient), Taylor and Betz (1983) initially reported values ranging from .86 to .89 for the five subscales of the CDSE and .97 for the full scale (long version). The alpha values for the subscales of the short version are typically observed to be lower, ranging from .73 to .83, with the total value for the CDSE-SF at .94 (Betz et al., 1996). In the empirical study with the new response scale (Betz et al., 2005), the reliability coefficients ranged from .78 to .87 for the five subscales, assuming a value of .94 or .95 in the full scale, depending on the sample used. These data are similar to the data obtained by Nilsson, Schmidt and Meek (2002) in their study on the generalization of reliability; these authors reported Cronbach's alpha coefficients for the full scale ranging from .92 to .97 (M = .94, SD = .01, n = 11) for samples relating to studies with the CDSE-SF. The mean values for the subscales ranged from .72 to .83.

When assessing career self-efficacy in a sample of high school students, the values of internal consistency of the CDSE-SF, for both the 25 items and the five subscales, were comparable to the values presented by Betz et al. (1996) for college

students. The cross-cultural studies of the CDSE-SF generally reported values of internal consistency similar to the values observed in the United States (Creed, Patton and Watson, 2002; Hampton, 2006). For the Australian students, the authors reported a reliability of .94 for the full scale with Cronbach's alphas ranging from .70 to .78 on the subscales; the results were practically identical for the South African students (.93 for the total result, ranging from .70 to .79 for the subscales). Similar results were obtained by Hampton (2006) in a sample of Chinese secondary students.

Studies conducted with the Portuguese version of the CDSE-SF have shown that the estimates of internal consistency have been lower than those observed in other countries. With higher education students, coefficients ranged from .53 to .71 for the subscales, with a value for the full scale ranging between .88 and .90 (Paixão, Leitão, Miguel & Borges, 2004; Kumar, Silva and Paixão, 2007); with high school students, coefficients are similar ranging from .41 to .73 for the subscales, with a value for the full scale of .89 or .90 (Silva & Paixão, 2005; Silva, Paixão & Albuquerque, 2009), respectively.

In short, this review of studies using the CDSE allows for the conclusion that the typical level of internal consistency is adequate when analyzing the pattern of responses for all items regardless of the cultural background or level of education of the subjects who were assessed using the scale. However, when examining the results at the subscale level, the reliability estimates are much more variable between the studies and are generally of lower quality (i.e., often below .7).

Despite the convergence of results from different studies, regarding the psychometric information relative to the concurrent reliability and validity indicators of the CDSE-SF, the performance of the subscales concerning dimensionality is inconsistent overall.

The initial study of the dimensionality of the CDSE (Taylor & Betz, 1983), with college students via exploratory Principal Components Analysis (PCA), led authors to report an unidimensional structure as the one that best fit the data. They concluded that "the measure (...) may be more appropriately viewed as a (...) general domain of career decision-making tasks and behaviors" (Taylor & Betz, 1983, pp. 79-80). An identical conclusion was reached by Hampton (2006), suggesting that the CDSE-SF is a "measure of generalized self-efficacy covering a domain of career decision-making behavior rather than an instrument that measured self-ef-

ficacy expectations for five career decision skills" (p. 151) and, later on, by Chaney, Hammond, Betz and Multon (2007), defending that the scale is primarily a general measure of career decision-making self-efficacy.

Studies using Confirmatory Factor Analysis (CFA), all conducted only with higher education samples from different countries (Gaudron, 2011; Hampton, 2005; Miller, Roy, Brown, Thomas and McDaniel, 2009; Watson, Brand, Stead & Ellis, 2001), also differ in their results. The only one that confirmed the original theoretical structure of the CDSE-SF was that of Miller et al. (2009). None of the studies of Watson et al. (2001), Hampton (2005) or Gaudron (2011) was able to adequately fit the data leading their authors to recommend that the scale be considered as a general measure of career decision self-efficacy.

The only study examining the dimensionality of the Portuguese version of the CDSE-SF (Silva et al. 2009), with high school students using PCA, did not replicate the five-factor model proposed for the scale. As the majority of the items saturated on the first component, the authors concluded that the Portuguese version of the scale predominantly measures generalized self-efficacy expectations concerned with the process of career decisions.

Except for the study by Miller et al. (2009), the theoretical model of measurement proposed for the CDSE is not replicable, regardless of the type of sample, language version or analytical procedures used. The literature reviewed equally emphasizes the presence of a latent unidimensional structure that would be sufficient to explain the pattern of responses in the CDSE (e.g., Creed et al. 2002). Specifically, it shows that the values of variance explained by this component/factor vary according to the study and range from 16% to 40% (M = 28%; *SD* = 9). Moreover, the number of items saturated in this first factor/component range from 8 to 21 (*M* = 12; *SD* = 4). Finally, Hampton (2005, p. 103) presents *eigenvalues* of 8.31, 1.25 and 1.02 for factors 1, 2 and 3, respectively. According to the criteria proposed by Reckase (1979), which have been used to assess the multidimensionality of the measures, the first factor explains more than 20% of the variability, and the first *eigenvalue* is several times greater than the second; these findings support the existence of a dominant first factor (Hambleton, Robin & Xing, 2000, p. 568).

Taylor and Betz (1983) had previously accepted this hypothesis when they concluded that the existence of a general factor was likely, provided the high values

of internal consistency for the full scale and the high values of correlation obtained between the subscales and factor structures.

The methods of factor analysis (both exploratory and confirmatory) that have been predominantly used in the literature present some weaknesses in assessing the dimensionality of the measures. The primary weakness is that these methods presume the existence of linear relationships between the variables and factors (Hambleton et al. 2000), which is questionable for the majority of scales used in psychosocial investigations. In addition, it is not appropriate to apply a factor analytic approach, which has been widely used for exploring or confirming the factor structure of measurement scales, directly to non-interval raw data (Wright & Linacre, 1989). In fact, these non-interval raw data must be constructed into sample-distribution free and item-distribution free measures before they can be analyzed using statistics requiring linear, interval data (Wright, 1997).

It is therefore important to explore alternative methods of psychometric analysis whose potential allows for the compensation of limitations inherent to the procedures of the Classical Test Theory (CTT). Betz and Turner (2011) propose the use of the Item Response Theory (IRT) and justify it based on the significant improvement achieved in terms of efficacy and precision of the psychological measures. The IRT was also suggested by Silva et al. (2009) for the calibration of the Portuguese version of the CDSE-SF.

The only study referenced in the literature on the application of the Rasch model to psychometric evaluation of the CDSE-SF was conducted with a sample of South Korean college students (Nam, Yang, Lee, Lee, & Seol, 2011). The results confirmed the unidimensionality of the CDSE-SF with the exception of three items, for which the high positive values of the respective structural coefficients suggested dependency, thus suggesting the possible existence of a second dimension.

The present study aimed to make a psychometric evaluation of the CDSE-SF in a sample of Portuguese high school students, within the framework of Rasch measurement. This study was performed in response to the growing need of new versions of the CDSE-SF for populations for whom English is not the official language, allowing for future cross-cultural studies on the equivalence of the construct of career decision self-efficacy with a well-established instrument. However, provided the methodological limitations identified in previous studies using CTT, the authors

chose to follow the suggestion of Betz and Turner (2011) and resorted to IRT using the Rasch model in an attempt to gather evidence for the measurement properties of the Portuguese version of the CDSE-SF.

2. The Study

2.1. Methods

2.1.1. Participants

The sample included 265 secondary education students who attended the 12th year of education in several public schools of the central and south regions of Portugal. Among these subjects, 130 (49.1%) were male, and 135 (50.1%) were female. Their ages ranged from 17 to 22 years old, with an average age of 17.60 (*SD* = .80). Only 52 students (19.6%) reported being unsuccessful in their academic paths, and 12 (4.5%) failed more than once. Information regarding ethnicity was not requested because these data are not typically obtained in studies conducted in Portugal (however, most respondents were Caucasian Europeans).

2.1.2. Procedures and tools

The investigation was monitored and approved by the Portuguese Ministry of Education; the school principals were then contacted, so the study could be conducted at their schools. The purpose of the study was explained, and informed consent was obtained from the students and their parents. Participation was voluntary, anonymous and confidential; the institutional contact information of the main researcher was also provided, so the results could be provided to interested parties. The surveys were administered by the first author in the context of the classroom in the presence of the class teacher. The students were provided no incentives for their participation. All participants completed the surveys during the class interval provided for the survey. The students who chose not to participate in the study were allowed to leave the classroom prior to the survey distribution.

The questionnaire included an initial sheet to obtain the student's demographic data (sex, age, school grade, academic success/failure and parents' academic qualifications) in addition to certain measures of their vocational behaviors.

The CDSE-SF (Betz & Taylor, 2001) was used to assess career decision self-efficacy. The scale included 25 items that were designed to measure five domains of

career decision self-efficacy; namely, accurate self-appraisal, gathering occupational information, goal selection, making plans for the future and problem solving. The answers were obtained using a scale with five alternatives, ranging from 1 = not at all confident to 5 = totally confident. The total score on the scale was calculated by adding the responses to the 25 items; higher scores indicated higher levels of career decision self-efficacy. The translation and retroversion of the CDSE-SF (Silva et al., 2009) were performed by bilingual doctoral fellows in educational and vocational guidance following the suggestion of Hambleton, Merenda and Spielberger (2005). A native English speaker compared the original scale and the respective retroversion for clarification and rewording.

2.1.3. Analyses

The Rasch analyses were performed using the computer program Winsteps (Linacre, 2011). Specifically, given the invariant polytomous format of all the items on the scale, the parameter estimates related to the subjects, items and response categories were calculated based on the Rating Scale Model (RSM, Wright & Masters, 1982). According to Linacre (2002), the RSM is an extension of the Rasch model for polytomous items and is provided by:

$$log\,[P_{nik} / P_{ni(k-1)}] = B_n - D_i - F_k$$

where P_{nik} is the probability that person n would respond in category k of item i, $P_{ni(k-1)}$ is the probability that person n would respond in category k-1 of item i, B_n is the ability of person n in the evaluated trait, D_i is the difficulty of item i and F_k is the difficulty of the step in category k-1 relative to category k (i.e., step calibration). This step calibration (F_k) is a threshold of the classification scale defined as the location corresponding to the equiprobability of observing the adjacent categories k-1 and k.

In psychometric terms, the choice for the RSM is justified for the transformation of ordinal data relative to the subjects' responses on an interval scale (Wright & Mok, 2004) and also for its ideal metric properties as a Rasch model, namely, sufficient statistics and specific and statistic objectivity for person and item fit. In practical terms, the RSM has the advantage of not requiring large samples for properly estimating parameters and of allowing the empirical determination of the quality of the response categories in Likert scales (Bond & Fox, 2007).

2.2. Results

The Rasch analyses produced indicators that allow for the quantification of model fit, estimations of items and person parameters and the diagnosis of the functioning of response categories to items (Fox & Jones, 1998). Table 1 shows the statistics of the items for fit (*infit* and *outfit*), location (D_i) and standard error (*SE*), which allowed for the assessment of content validity of the CDSE-SF and the coefficients for the analysis of structural validity of the scale (Wolfe & Smith, 2007).

Table 1. The CDSE-SF Item Psychometric Properties and Principal Component Analysis

	MNSQ					
Item	Infit	Outfit	D_i	*SE*	r_{pm}	SC
1	.71	.73	.05	.08	.60	.21
2	.84	.84	-.20	.08	.64	.57
3	.97	.97	.45	.08	.64	.19
4	.81	.81	.65	.08	.47	-.31
5	.89	.92	.03	.08	.47	-.10
6	.87	.87	-.21	.08	.66	.61
7	.56	.56	-.02	.08	.69	.06
8	1.12	1.15	-.20	.08	.47	-.13
9	1.25	1.25	-.34	.08	.62	.41
10	1.04	1.04	.72	.08	.58	-.04
11	1.04	1.02	-.44	.08	.67	.44
12	1.19	1.19	.06	.08	.53	-.43
13	1.42	1.44	.67	.08	.43	-.45
14	.65	.66	-.31	.08	.64	-.03
15	.93	.93	.16	.08	.58	-.24
16	1.48	1.51	1.10	.08	.45	.14
17	1.40	1.46	.43	.08	.41	-.33
18	.95	.95	-.36	.08	.54	.05
19	1.21	1.18	-.82	.09	.47	-.19
20	.87	.85	-.97	.09	.64	.51
21	.99	.98	.11	.08	.56	.10
22	.87	.85	-1.05	.09	.63	.37
23	.91	.91	-.03	.08	.56	-.23
24	1.02	1.02	.21	.08	.54	-.39
25	.90	.91	.32	.08	.59	-.41
Mean	1.00	1.00	.00	.08	---	---
SD	0.23	0.23	.51	.00	---	---

Note: D_i = item location; Infit = information-weighted mean square statistic; MNSQ = mean square fit indices; Outfit = outlier-sensitive mean square statistic; r_{pm} = point-measure correlation; SC = structure coefficients from the standardized residual PCA; and SE = standard error.

In terms of content validity, the analysis presented in Table 1 shows that the

mean values of *infit* and *outfit* equal the expected value of 1.0, which indicates a perfect fit of the items. All 25 items on the CDSE-SF showed individual values within the interval [.5 – 1.5], as stipulated by Wright and Linacre (1994) as being productive for the measure and indication of the absence of redundant items and the presence of homogeneity among the items. This initial indicator of unidimensionality of the data tends to be corroborated by other estimators that have been proposed in the context of the Rasch model with the purpose of analyzing the contribution of items in defining a central construct for the internal structure of the scale. The point-measure correlation (r_{pm}), which is similar to the item-total correlation of the CTT with values ranging from .41 to .69, suggests the absence of non-modeled noise or dependence among the data. This result allowed for the conclusion that each of the items contributed to define a common construct (Linacre, 2011). In addition, the values for the standard error of the items, which ranged from .08 to .09, also indicated that item reliability was high. The value of *item separation reliability*, which is a reliability estimator similar to Cronbach's alpha, was equal to .97. This result showed that the items of the CDSE-SF were measured with high precision.

To investigate the structural validity of the scale, a PCA of the standardized residuals was calculated after controlling the Rasch dimension in order to test for a basic assumption of Rasch modeling (e.g., unidimensionality) and determining to what extent the scale items corresponded to the defined construct (Smith, 2004). Given the absence of consensus on the criteria that are indicative of a secondary dimension (Chou & Wang, 2010), as an indicator of the unidimensionality of the CDSE-SF, the authors considered an *eigenvalue* below 3.0 (Linacre, 2011), and up to 10% of the variance explained by the first residual component. The last column in Table 1, which presents the structural coefficients of the 25 items of the CDSE-SF, shows that only eight of the items had values above the cutoff point of ± .40, and item 6 was the only item to slightly exceed the threshold of .60. Furthermore, the first component had an *eigenvalue* = 2.70, which represents a residual variance of 10.6% and was only slightly higher than the value proposed by Linacre (2011). This result suggests that the standardized residuals have no additional systematic information. Because the variance explained by the measure (37.9%) is classified as being moderately strong (Reckase, 1979), these results suggest the unidimensionality of the CDSE-SF, which is an assumption required by the Rasch model.

Substantive validity refers to the diagnosis of the empirical functioning of categories in the response scale, with the purpose of determining whether they function in accordance with what was expected by the author of the instrument when developing the respective items (Wolfe & Smith, 2007). In the context of the RSM, Linacre (2002) proposed a set of criteria for determining the efficiency of response categories: (a) the uniform distribution of the response frequencies throughout the different categories with a minimum of 10 observations in each; (b) the monotonic progression of the mean measure observed (B_n) and step calibration (F_k) throughout the response categories; and (c) an *outfit* mean square (MNSQ), which is more sensitive than the *infit* for unexpected responses, of the response categories of less than 2.0. Table 2 summarizes the statistics required to assess the extent to which the five response categories of the CDSE-SF fulfill these criteria. The analysis concludes that the structure of the response scale fulfills the criteria established by Linacre (2002), provided that each category has an observed frequency of response above 10 and an *outfit* value below 2.0. In addition, the mean measures observed and step calibration increased monotonically throughout the five response categories.

Table 2. The CDSE-SF Response Category Statistics

	Observed			MNSQ		
Category	Count	%	Average B_n	Infit	Outfit	F_k
1	174	3	-.71	1.32	1.37	---
2	798	12	-.24	.98	.98	-2.11
3	2443	37	.39	.93	.93	-1.00
4	2294	35	1.06	.94	.93	.80
5	916	14	1.78	.99	.99	2.31

Note: B_n = person trait; F_k = step calibration; Infit = information-weighted mean square statistic; MNSQ = mean square fit indices; and Outfit = outlier-sensitive mean square statistic.

The adequacy of the response scale, based on the efficient functioning of its five categories, is supported by the result of the analysis performed on person fit. The mean and standard deviation for the fit statistics were 1.01 and .62 (*infit*), respectively, and 1.00 and .60 (*outfit*), respectively. The model fit can be considered good because the percentage of people with *infit* and/or *outfit* above 1.5 is low (18.50%), which is the value of the upper threshold of psychometric acceptability for Rasch

modeling (Wright & Linacre, 1994). The parameters were estimated with high precision for most people and confirmed by the corresponding value of *person separation reliability* (equal to .89), which is a reliability estimator that measures the proportion of people variance that is not explained by measurement error.

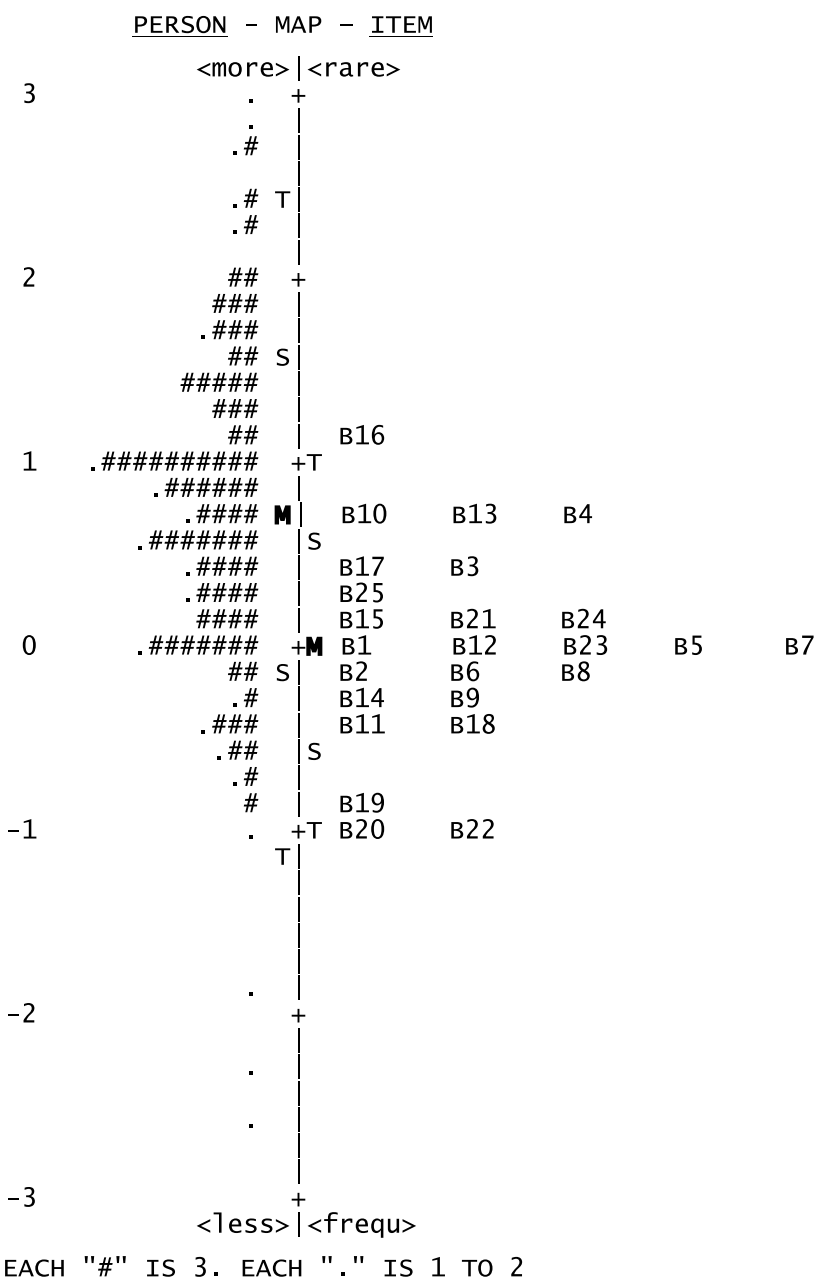

Figure 1. Joint Person and Item Representations along the Career Decision Self-Efficacy Variables.

Figure 1 shows the person-item joint representation in the identical common metric scale (i.e., logits). The 25 items of the CDSE-SF were organized in descending order of difficulty and operated by the amount of attributes measured in the person (i.e., career decision self-efficacy). Item 16 was the most difficult, whereas items 20 and 22 were the easiest to subscribe, i.e., people with more confidence necessary for making good career decisions were more likely to mark category 4 = very confident or 5 = totally confident on item 16 in comparison to people with low levels of confidence.

In this sense, the estimates related to the person's attributes and difficulty of the items were expected to overlap substantially, so all the items can be considered appropriately directed to the subject sample. Whenever the difference between the estimate means was less than 1 logit (Bond & Fox, 2007), as was observed in the present study (.71), the information contained in the items allowed for an accurate discrimination of people, in this case, an accurate discrimination at different levels of career decision self-efficacy.

Additionally, the differential item functioning (DIF) was analyzed with the purpose of assessing the validity of the results of the CDSE-SF in relation to sex. The standardized difference between the locations of male and female parameters were calculated for this purpose after matching for possible sex-related differences in the distribution of career decision self-efficacy. Therefore, the Bonferroni test was used, which corrected the level of significance according to the number of comparisons (.05/25) (Linacre, 2011). Following this conservative criterion, none of the 25 items of the CDSE-SF showed a location above the cutoff of .50 for either male or female individuals, which was the value stipulated by Wright and Douglas (1975) for establishing DIF.

3. Discussion

Rasch analyses were performed with the purpose of examining the psychometric properties of the Portuguese version of the CDSE-SF. Data were analyzed using the RSM (Wright & Masters, 1982), which, as a Rasch model, has ideal metric properties for ranking item difficulty and the person's ability (e.g., the level of the attribute measured) on a common scale. In addition to the joint measurement of people and items, and provided that the data fitted to the model's requirements, Rasch

modeling also allows for the comparison of people (items) regardless of the items (people) used in the measurement, which is a property designated as *specific objectivity* by Rasch (Andrich, 1988).

Once good data fit to the Rasch model was confirmed, therefore allowing for the estimation of item and people parameters with high precision, the psychometric properties of the Portuguese version of the CDSE-SF were analyzed in terms of its dimensionality. The main purpose was to test the basic assumption of unidimensionality required by the Rasch model. The results of the present study show that the 25 items on the scale are not distributed throughout the areas of competency of career decision-making initially defined by the authors (Taylor & Betz, 1983; Betz et al., 1996). Alternatively, the data show that the items predominantly measure a single latent dimension, which represents a generalized expectation of self-efficacy in relation to the process (tasks and behaviors) of career decisions. The Rasch analysis used in the present study thereby provides psychometric evidence that supports the hypothesis of unidimensionality of the CDSE-SF.

The obtained results indicate that the five categories of the response scale used in the CDSE-SF also function properly, which psychometrically validates the option provided by Betz et al. (2005) based on the CTT, a result previously confirmed by Nam et al. (2011) based on the IRT, although not for secondary education students. The Rasch model conferred substantive validity for the use of this short version of the response scale, which allowed for the accurate detailing of its functioning by calibrating each of the five categories.

Regarding the items of the CDSE-SF, their distribution reflects a wide range of individual differences in career decision self-efficacy, with the average level of this trait measured in the students being higher than the average difficulty level of the items used for their assessment. The difficulty level of the 25 items on the scale reflected an appropriate range of levels of career decision self-efficacy among students, thus allowing for the discrimination of different levels of this trait in the sample. The high internal consistency of people and items enabled an estimation of the respective parameters with high precision, allowing most of the response patterns to be fitted to the model (Prieto & Delgado, 2003).

Following a conservative criterion, the DIF analysis performed according to sex showed that the functioning of the 25 items on the CDSE-SF was consistent in

both sexes, and considered equally difficult for these two subgroups. The items were sufficiently robust to allow for the assessment of career decision self-efficacy regardless of the respondent's sex. Thus, the answers only quantified the person's level of construct, which was measured according to the difficulty of the items and not because of other competencies explained by the respondent's sex.

3.1. Conclusions

Overall, the present study confirms the Portuguese version of the CDSE-SF as a valid tool for assessing the expectations of career decision self-efficacy in a population of Portuguese high school students. Rasch modeling established the unidimensionality of the scale, allowing people's levels of career decision self-efficacy and the item difficulties of the CDSE-SF used to measure this construct to be ranked on a single continuous scale, a ranking achieved with high precision based on the adequacy of the response scale with five alternatives and no sex bias.

In terms of clinical practice in career counseling, it is possible to consider the development of a computerized version of the test (CAT) that applies the CDSE-SF and ranks the ability to recommend an item based on the level of career self-efficacy revealed by the consultants and the intended measurement error criterion, an approach that can help to streamlining the duration of intervention and minimizing the costs of the process by adapting the items to the consultant's level of career self-efficacy.

The present study has some limitations that should be overcome in future investigations, particularly with respect to the sample. Although the sample reproduces the demographic characteristics of the student population sought, it could not be considered representative of the diversity of Portugal. Future studies should utilize random sampling methods, so the results can be generalized.

Finally, confirmed the psychometric robustness of the CDSE-SF with the Rasch model, authors want to proceed with further research aiming to compare the experiences of career decision-making between individuals in Portugal and the United States, in order to map the construct of career decision self-efficacy in different cultures and nations.

References

Andrich, D. (1988). *Rasch models for measurement*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-068. Newbury Park, CA: SAGE.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavior change. *Psychological Review*, *84*, 191-215.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Betz, N. E. (2007). Career self-efficacy: Exemplary recent research and emerging directions. *Journal of Career Assessment*, *15*, 403-422.

Betz, N. E., Hammond, M. S., & Multon, K. D. (2005). Reliability and validity of five-level response continua for the Career Decision Self-Efficacy Scale. *Journal of Career Assessment*, *13*, 131-149.

Betz, N. E., Klein, K. L., & Taylor, K. M. (1996). Evaluation of a short-form of the Career Decision-Making Self-Efficacy Scale. *Journal of Career Assessment*, *4*, 47-57.

Betz, N. E., & Luzzo, D. A. (1996). Career assessment and the Career Decision-Making Self-Efficacy Scale. *Journal of Career Assessment*, *4*, 413-428.

Betz, N. E., & Taylor, K. M. (2012). *Career Decision Self-Efficacy Scale Manual*. Redwood City, CA: Mindgarden Inc.

Betz, N. E., & Turner, B. M. (2011). Using item response theory and adaptive testing in online career assessment. *Journal of Career Assessment*, *19*, 274-286.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Chaney, D., Hammond, M. S., Betz, N. E., & Multon, K. D. (2007). The reliability and factor structure of the Career Decision Self-Efficacy Scale-SF with African Americans. *Journal of Career Assessment*, *15*, 194-205.

Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residual. *Educational and Psychological Measurement*, *70*, 717-731.

Creed, P. A., Patton, W., & Watson, M. B. (2002). Cross-cultural equivalence of the Career Decision-Making Self-Efficacy Scale-Short Form: An Australian and South African comparison. *Journal of Career Assessment*, *10*, 327-342.

Crites, J. O. (1978). *Career Maturity Inventory*. Monterey, CA: CTB/McGraw-Hill.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology*, *45*, 30-45.

Gaudron, J.-P. (2011). A psychometric evaluation of the Career Decision Self-Efficacy Scale-Short Form among French university students. *Journal of Career Assessment*, *19*, 420-430.

Hackett, G., & Betz, N. E. (1981). A self-efficacy approach to the career development of women. *Journal of Vocational Behavior*, *18*, 326-339.

Hackett, G., & Betz, N. E. (1995). Self-efficacy and career choice and development. In J. E. Maddux (Ed.), *Self-efficacy, adaptation, and adjustment: Theory, research, and application* (pp. 249-280). New York: Plennum Press.

Hambleton, R. K., Robin, F., & Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In H. E: Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 553-581). San Diego, CA: Academic Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hampton, N. Z. (2005). Testing for the structure of the Career Decision Self-Efficacy Scale-Short Form among Chinese college students. *Journal of Career Assessment*, *13*, 98-113.

Hampton, N. Z. (2006). A psychometric evaluation of the Career Decision Self-Efficacy Scale-Short Form in Chinese high school students. *Journal of Career Development*, *33*, 142-145.

Kumar, M. E., Silva, J. T., & Paixão, M. P. (2007). Life projects of higher education students: Relationship with optimism and career self-efficacy. *Psychologica*, *44*, 45-62.

Lent, R. W., & Hackett, G. (1987). Career self-efficacy: Empirical status and future directions. *Journal of Vocational Behavior*, *30*, 347-382.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85-106.

Linacre, J. M. (2011). *Winsteps Rasch measurement computer program, version 3.73.0*. [Computer program.] Chicago, Il: Winsteps.com.

Miller, M. J., Roy, K. S., Brown, S. D., Thomas, J., & McDaniel, C. (2009). A confirmatory test of the factor structure of the short form of the Career Decision Self-Efficacy Scale. *Journal of Career Assessment*, *14*, 507-519.

Nam, S. K., Yang, E., Lee, S. M., Lee, S. H., & Seol, H.(2011). A psychometric evaluation of the Career Decision Self-Efficacy Scale with Korean students: a Rasch model approach. *Journal of Career Development*, *38*, 147-166.

Nilsson, J. E., Schmidt, C. K., & Meek, W. D. (2002). Reliability generalization: An examination of the Career Decision-Making Self-Efficacy Scale. *Educational and Psychological Measurement*, *62*, 647-658.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-144. Thousand Oaks, CA: SAGE.

Paixão, M. P., Leitão, L. M., Miguel, J. P., & Borges, G. (2004). Goal characteristics and vocational decision-making beliefs during the transition from under graduation in psychology to work – the role of individual agency indicators. Communication presented to the *9th International Conference on Motivation*, Lisbon, September, 30th – October, 2nd.

Peterson, S. L., & delMas, R. C. (1998). The component structure of career decision-making self-efficacy for underprepared college students. *Journal of Career Development*, *24*, 209-225.

Prieto, G., & Delgado, A. R. (2003). Análisis de un test mediante el modelo de Rasch [Analysis of a test using the Rasch model]. *Psicothema*, *15*, 94-100.

Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics*, *4*, 207-230.

Silva, J. T., & Paixão, M. P. (2005). Preliminary psychometric studies of the CDSE-SF. Communication presented to the *AIOSP International Conference 2005*, Lisbon, September, 14th – 16th.

Silva, J. T., Paixão, M. P., & Albuquerque (2009). Psychometric characteristics of the Portuguese version of the Career Decision Self-Efficacy Scale-Short Form (CDSE-SF). *Psychologica*, *51*, 27-46.

Smith Jr, E. V. (2004). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In: E. V. Smith Jr & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 575-600). Maple Grove, MN: JAM Press.

Taylor, K. M., & Betz, N. E. (1983). Applications of self-efficacy theory to the understanding and treatment of career indecision. *Journal of Vocational Behavior*, *22*, 63-81.

Watson, M. B., Brand, H. J., Stead, G. B., & Ellis, R. R. (2001). Confirmatory factor analysis of the Career Decision-Making Self-Efficacy Scale among South African university students. *Journal of Industrial Psychology*, *27*, 43-46.

Wolfe, E. W., & Smith Jr, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II – Validation activities. In E. V. Smith Jr & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 243-290). Maple Grove, MN: JAM Press.

Wright, B. D. (1997). S. S. Stevens revisited. *Rasch Measurement Transactions*, *11*, 552-553.

Wright, B. D., & Douglas, G. A. (1975). A better procedure for sample-free item analysis. *Research Memorandum*. Statistical Laboratory. Department of Education. University of Chicago.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal; Measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*, 857-860.

Wright, B D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, p. 370.

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

Wright, B. D., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. In E. V. Smith Jr & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 1-24). Maple Grove, MN: JAM Press.

Positive and Negative Affect Schedule—Portuguese European Version (PANAS-P): A Rasch Analysis

Abstract

The present study analyzes the psychometric properties of a Portuguese version of the PANAS (PANAS-P), directly developed from the original instrument, on a sample of secondary education students (N = 519) using the Rasch model. Overall, the results of the analysis were psychometrically sound; although one item in each subscale of the PANAS-P demonstrated lack of fit, the measurement system is largely unaffected because each subscale is adequately fit to a latent one-dimensional structure, as requested by Rasch modeling. The response scale, comprising 5 categories, demonstrated proper functioning; therefore, the people and item parameters could be estimated with high precision (.80 and .99, respectively). Differential item functioning (DIF) analyses confirmed that there were no differences in the results of the PANAS-P with respect to gender. Finally, psychometric implications derived from the results of the present study are discussed, and suggestions are provided for future investigations.

Keywords: positive affect, negative affect, dimensionality, Rasch analysis, rating scale model

Submitted

Miguel, J. P., Silva, J. T. & Prieto, G. (submitted). Positive and Negative Affect Schedule—Portuguese European Version (PANAS-P): A Rasch Analysis. *International Journal of Testing.*

Positive and Negative Affect Schedule—Portuguese Version (PANAS-P): A Rasch Analysis

The study of affective states has been marked by two important traditions. The categorical approach suggests that affective states are specific and conceptually distinct, constituting a small number of discrete emotions that must be measured independently (Ekman, 1993). In contrast, the dimensional approach argues that affective states are general and conceptually similar, preferring to focus on dimensions that produce important associations among fundamental emotions and, therefore, proposes that the best way to measure the states is by relating them with one another (Plutchik, 1997).

Although both approaches have been successfully used in the empirical exploration of affective states, in the present study we focus on the dimensional approach, while recognizing that the debate on this issue continues (Widiger & Frances, 1994). In fact, the dimensional models, as suggested by Cloninger, Bayon, and Svrakic (1998), retain more information compared to categorical models when characterizing the affective states as more or less prominent and adaptive.

The Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) is the measure most often used in the study of affective states by researchers adopting the dimensional approach. Comprising two subscales with 10 items each that assess positive affect (PA) and negative affect (NA), the PANAS can be administered to elicit more dispositional (trait) or situational (state) aspects, depending on the specified period of time in which the subject is instructed to reach different levels of retrospection (e.g., "right now", "today", "during the past few days," "during the past few weeks," "during the past year", and "in general").

According to the authors, the development of the PANAS lies on theoretical and empirical foundations. Conceptually, the PA and NA are conceived as the dispositional activation of positively and negatively valued affects. PA is a dimension of enthusiasm, activation and alertness, subsuming a positive state mood (e.g., joy, interest, enthusiasm and alertness); high PA is a state of high energy, full concentration and pleasurable engagement, whereas low PA is characterized by sadness and lethargy. NA is a general dimension of substantive distress and unpleasurable engagement that subsumes a variety of aversive mood states (e.g., anger, contempt, disgust, fear and nervousness); low NA is a state of calmness and serenity. Empirically, Watson et al. (1988) operationalized the PANAS items based on the mood categories proposed by Zevon and Tellegen (1982).

The adequacy of the psychometric properties of the PANAS, in terms of reliability and validity (Watson & Clark, 1994; Watson et al., 1988), along with its formal characteristics, help explain the widespread use of this instrument in clinical and non-clinical settings and with different subjects. The progressive interest of transcultural research on affective domain has promoted studies on PANAS adaptation and its translation into languages other than English, necessitating several international versions of this instrument (Allik & Realo, 1997; Balatsky & Diener, 1993; Engelen, De Peuter, Victoir, Van Diest, & Van Den Bergh, 2006; Gaudreau, Sanchez, & Blondin, 2006; Hilleras, Jorm, Herlitz, & Winblad, 1998; Joiner, Sandin, Chorot, Lostao, & Marquina, 1997; Krohne, Egloff, Kohlmann, & Tausch, 1996; Lim, Yu, Kim, & Kim, 2010; Pandey & Srivastava, 2008; Terraciano, McCrae, & Costa, 2003; Yamasaki, Katsuma, & Sakai, 2006).

Despite the widespread acceptance of the PANAS, the results of research conducted with the scale have not demonstrated consensus, particularly with respect to dimensionality. Different explanations have been advanced (Tuccito, Giaccobbi, & Leite, 2010). One reason, methodological in nature, suggests that the internal PANAS structure may be sensitive to sample type and/or time period specified for the retrospection. Another one, there are also statistical in basis, is supported on the criteria that have been used in the estimation of factorial parameters.

From a statistical standpoint, the research has mainly relied on classical test theory (CTT) to examine construct validity, specifically on exploratory factor analysis (EFA) and Cronbach's alpha. However, the use of these two methods has been

criticized as they both are considered insufficient for determining construct validity of psychological assessment instruments. In fact, Tabchnick & Fidell (2001) argued that the use of EFA alone does not compare data sets to any another criteria which, through multiple rotations, data can be easily manipulated. Moreover, correlation between scores on items does not mean that they are conceptually related, even in the presence of a strong loading on a single factor (Waugh & Chapman, 2005). According to Kline (2000), personality researchers tend to paraphrase items in order to form a factor based on correlational analysis and such factors (e.g., "tautologous factors"), fail the test of construct validity and cannot be generalized across samples.

Concerning the use of Cronbach alpha reliability estimates to validate results, the major drawback is the tendency to assume that they are related with construct validity when they are not (Cortina, 1993; Schmitt, 1996); the size of an internal consistency index is irrelevant to dimensionality (Embretson & Reise, 2000). Because of its dependency on the number of items in the construct (Furr, 2011), it is incorrect to use Cronbach's alpha as a measure of the internal structure of the constructs to be measured (Sijtsma, 2009).

Recently, a Portuguese version of the PANAS has been created (Galinha & Ribeiro, 2005) from which some psychometric properties are known. The authors chose to replicate the construction process of Watson et al. (1988) from the 60 original items of Zevon and Tellegen (1982) in an attempt to reach the same affective descriptors instead of translating the 20 items from the original scale. When the best factorial solution for the intercorrelation matrix observed in their study was sought, the authors observed that, although the factorial analysis (principal components method) isolated the two affective dimensions, this version only corresponds with the original scale on six items, in the case of the PA, and three items, in the case of the NA.

Given the limited comparability between the original scale and the version developed by Galinha and Ribeiro (2005), Miguel and Silva (2012) created a literal equivalent of the original PANAS (Watson et al., 1988) in European Portuguese (PANAS-P) with a sample of secondary school students. The psychometric study of the PANAS-P (Miguel and Silva, in press) replicated the factorial structure of the original scale, with the two subscales permitting measurements of PA and NA

with adequate reliability, despite the fact that each contains an item with problematic behavior.

Efforts made by these authors were an attempt to develop a Portuguese version able to be used in cross-cultural research, as the main problem found when an international version of the PANAS was tested (Thompson, 2007) was the difficulty to find good translations of the scales that could fully match the meanings of the original instrument. It is important to have sound information on the psychometric characteristics of this new literal equivalent of the scale before to advance in the studies of affect in Portugal. Only after such compelling information is available it is possible to undertake comparative studies in a cross-cultural perspective, comparing the Portuguese population with the original one for which the PANAS scale was devised.

The present study undertook a psychometric evaluation of the PANAS-P in a sample of Portuguese high school students within the framework of the item response theory (IRT) in order to provide an important approach to item calibration. Being a subject-independent statistical approach, it allows understanding the idiosyncrasies of the Portuguese samples. IRT modeling enables to understand individual response patterns and difficulty parameters obtained for each item, providing information on which are the easiest and the hardest. Moreover, IRT modeling has important clinical advantages as it enables professionals to understand the patient's behavior regarding a difficult or easy item, which is helpful for intervention purposes as well as for normative data (Embretson & Reise, 2000).

This study, complementing a former study carried out by the authors in response to the growing need for new versions of the PANAS for populations in which English is not the official language, aimed to probe the psychometric quality of the PANAS-P. However, given the methodological limitations previously identified in CTT, the authors chose to resort to IRT using the Rasch model in an attempt to gather further evidence for the measurement properties of the Portuguese version of the PANAS. In this paper, authors propose to use Rasch model measurement analysis to confirm construct dimensionality for the PANAS-P. Convergent and divergent validity with a measure of self-esteem, the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965), will be also explored with the Portuguese adolescent sample, as an association between high PA and low NA and self-esteem is expected to occur (Brown & Marshall, 2001).

Study

Method

Subjects

Only the surveys in which the student responded to all of the scale items were included in the analyses. The sample included 519 high school students in the 10th, 11th and 12th grade at several public schools in the central and southern regions of Portugal. Of these subjects, 256 (49.3%) were male and 263 (50.7%) were female. Their ages ranged from 14 to 21 years, corresponding to an average age of 16.26 years (SD = 1.19); three subjects did not report their age. Information on ethnicity was not requested, as these data are not routinely obtained in studies performed in Portugal; nevertheless, most respondents were Caucasian Europeans.

Measurements

For the present study, a questionnaire was developed, which included an initial page to collect demographic information about the students (gender, age, grade, school success/failure and academic qualifications of the parents) and some measures of their vocational behavior (these variables are not used in the present study). In addition to these parameters, measures of the type of emotional (affect) response and the level of self-esteem of the students were collected.

The type of affect was assessed using the Portuguese version of the Positive and Negative Affect Schedule (PANAS, Watson et al., 1988), translated from the original instrument by the first author. The scale consists of two subscales, each with 10 items, which provides a brief measure of positive affect (PA) and negative affect (NA). The items consist of adjectives indicating mood states related to PA and NA. Respondents are asked to evaluate to what extent they experienced each of the 20 emotions in a specific time period using a Likert-type scale with five response alternatives: 1 = very slightly or not at all, 2 = a little, 3 = moderately, 4 = quite a bit and 5 = extremely. Although different time periods can be used with the PANAS, depending on the desired level of retrospection, in the present study, the subjects were instructed to answer regarding what they felt "during the past few weeks."

The students' levels of self-esteem were assessed using the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965). The scale contains 10 items consisting of statements related to self-respect and self-acceptance from which the subjects express

their agreement through a Likert-type response scale with four categories ranging from 1 = strongly disagree to 4 = totally agree. Half of the items on the RSES are negatively formulated, requiring, therefore, that the respective responses be reversed to compute the scale's total score. The reliability of the RSES, based on the person separation index that corresponds in IRT to the Cronbach's alpha from CTT (Linacre, 2011), was established at .87 for the present sample.

Procedure

The present study was part of a broader investigation into the effects of self-efficacy and of mathematics anxiety in making career decisions in secondary school students. In the present study, measurements of the respondent's affect were made to examine its incremental contribution to the prediction of vocation. The choice of PANAS for affect operationalization was based on the notoriety of the measure in the literature and stems, in large part, from the psychometric quality of the results. For the reasons cited above, we decided to create a parallel version of the original PANAS instead of using the European Portuguese version proposed by Galinha and Ribeiro (2005).

Translation of the instrument was conducted by the first author from the original PANAS and the retroversion was made independently by two bilingual colleagues with doctorate degrees in psychology following the psychometric standards proposed for the development and transcultural adaptation of transliteration equivalent of psychological scales (Hambleton, Merenda, & Spielberger, 2005). A native English speaker compared the original scale and its retroversion for the clarification and rewording of the meaning of some words.

The survey was approved by the Portuguese Ministry of Education. The principals of the selected secondary schools were contacted to conduct the study in their schools. Additionally, the informed consent from students and their parents was obtained, and the purpose of the study explained. Participation was voluntary, anonymous and confidential; the institutional contact of the principal investigator was available for subsequent return of a summary of results to interested parties. The surveys were conducted in a classroom by the first author of the study in the presence of the teacher, and no incentive was given for the student's participation. All participants completed the surveys during the class time given for their application.

Students who declined to participate were allowed to leave the classroom before distribution of the surveys.

Analysis

The Rasch analyses were performed using the computer program Winsteps (Linacre, 2011). Specifically, given the invariant polytomous format of all of the items on the scale, the parameter estimates related to the subjects, items and response categories were calculated based on the Rating Scale Model (RSM; Wright & Masters, 1982). According to Linacre (2002), the RSM is an extension of the Rasch model for polytomous items and is provided by the following:

$$log\,[P_{nik} / P_{ni(k-1)}] = B_n - D_i - F_k$$

where P_{nik} is the probability that person n would respond in category k of item i, $P_{ni(k-1)}$ is the probability that person n would respond in category k-1 of item i, B_n is the ability of person n in the evaluated trait, D_i is the difficulty of item i and F_k is the difficulty of the step in category k-1 relative to category k (i.e., step calibration). This step calibration (F_k) is a threshold of the classification scale defined as the location corresponding to the equiprobability of observing the adjacent categories k-1 and k.

In psychometric terms, the choice of the RSM is justified by the transformation of the ordinal data relative to the subjects' responses on an interval scale (Wright & Mok, 2004) and for its ideal metric properties as a Rasch model, namely, sufficient statistics and specific and statistic objectivity for person and item fit, that is, the number-correct score is a sufficient statistics for estimating ability when the Rasch model fits (Wright & Masters, 1982). Moreover, the Rasch model produces measurement properties that are essential, namely, items maintain their difficulty order throughout the ability range, and the model is simple, elegant, easily understood, and the estimation of item parameters and abilities are "separable" (Yen & Fitzpatrick, 2006). In practical terms, the RSM has the advantage of not requiring large samples for properly estimating parameters and of allowing the empirical determination of the quality of the response categories in Likert scales (Bond & Fox, 2007).

The evaluation of PANAS-P with the RSM focused on aspects of content validity, structural validity and substantive validity proposed by Wolfe and Smith (2007) and based on work by Messick (1995).

Results

Rasch analyses produce indicators that allow quantifying the fit of the model to estimate the item and individual parameters and to diagnosis the functioning of the item response categories (Fox & Jones, 1998). Tables 1 and 2 present the statistics for the item adjustment (infit and outfit), the location (Di) and the standard error (SE) that permits the assessment of content validity of the two subscales of the PANAS-P beyond coefficients that allow analysis of their structural validity (Wolfe & Smith, 2007).

Table 1. PANAS (PA) Item Psychometric Properties and Principal Component Analysis

	MNSQ					
Item	Infit	Outfit	D_i	*SE*	r_{pm}	SC
1. Interested	.73	.74	-.14	.06	.61	.57
4. Proud	1.05	1.05	.79	.06	.66	-.28
6. Determined	.74	.73	-.52	.06	.72	.26
7. Active	.92	.91	-.74	.06	.68	-.07
9. Strong	.83	.83	-.04	.06	.67	-.41
12. Inspired	1.01	1.01	.44	.06	.64	-.14
13. Attentive	1.09	1.10	.07	.06	.46	.72
15. Excited	1.17	1.15	-.86	.06	.65	-.65
17. Enthusiastic	.73	.73	-.03	.06	.71	-.37
18. Alert	1.65	1.77	1.05	.06	.35	.31
Mean	.99	1.00	.00	.06	---	---
SD	.27	.30	.59	.00	---	---

Note. D_i = item location; Infit = information-weighted mean square statistic; MNSQ = mean square fit indices; Outfit = outlier-sensitive mean square statistic; r_{pm} = point-measure correlation; SC = structure coefficients from the standardized residual PCA; *SE* = Standard error.

Analysis of Tables 1 and 2 demonstrate that, in terms of content validity, the mean infit and outfit values tend to equalize the expected value of 1.0, which indicates a perfect fit of the items. With the exception of one item in each of the subscales, all items exhibit individual values within the interval [.5 - 1.5] that Wright and Linacre (1994) stipulate as productive for measurement, which indicates that the lack of redundant items and the existence of homogeneity among the PANAS-P items. In the specific cases of item 18 (PA) and item 5 (NA), the residual values indicate lack of homogeneity compared to the other items of the respective subscales;

however, as these values are less than 2.0, they neither contribute to the measuring process of the respective latent dimensions nor take away from them (Prieto & Delgado, 2007). This first indicator of the existence of data unidimensionality tends to be corroborated by other estimators proposed in the context of the Rasch model to analyze the contribution of items in defining a central construct for the internal structure of each of the subscales. In fact, the point-measure correlations (rpm), similar to the item-total correlations of the CTT, with values varying from .35 to .72 (PA) or from .38 to .71 (NA), suggest the absence of non-modeled noise or dependence on the data, permitting the conclusion that each one of the items contributes to the definition of their common construct (Linacre, 2011). In addition, the standard error values of the items, ranging from .05 to .06, suggests that the fit of the items is high in both scales. Indeed, the item separation reliability value equals .99, indicating that the items in the PANAS-P were measured with high reliability.

Table 2. PANAS (NA) Item Psychometric Properties and Principal Component Analysis

	MNSQ					
Item	Infit	Outfit	D_i	*SE*	r_{pm}	SC
2. Upset	.72	.73	-1.00	.05	.65	.11
3. Scared	.92	.92	.55	.05	.65	-.10
5. Ashamed	1.60	1.65	.70	.05	.38	-.54
8. Distressed	.86	.85	-.44	.05	.71	.69
10. Hostile	1.24	1.47	.50	.05	.41	-.52
11. Irritable	.97	.95	.01	.05	.63	-.07
14. Afraid	.78	.80	-.38	.05	.68	-.03
16. Guilty	1.29	1.14	.94	.06	.57	-.33
19. Nervous	.74	.73	-.45	.05	.70	.71
20. Jittery	1.02	1.04	-.45	.05	.62	.38
Mean	1.01	1.03	.00	.05	---	---
SD	.27	.30	.61	.00	---	---

Note. D_i = item location; Infit = information-weighted mean square statistic; MNSQ = mean square fit indices; Outfit = outlier-sensitive mean square statistic; r_{pm} = point-measure correlation; SC = structure coefficients from the standardized residual PCA; *SE* = Standard error.

To investigate the structural validity of the scale, the PCA of the standardized residuals was calculated after controlling for Rasch to determine the extent to which the subscale items correspond to the defined constructs (Smith, 2004). Given the lack of consensus on criteria indicative of a secondary dimension (Chou & Wang,

2010), the authors opted to consider an eigenvalue less than 3.0 (Linacre, 2011) and a variance explained by the first component of the residue, not exceeding 10% as indicators of the unidimensionality of the PANAS-P subscales. The last column on Tables 1 and 2, present the structural coefficients for the PA and NA items, indicates that only four of the items exhibit values above the cutoff value (± .40), with only two exceeding the threshold (.60) in each of the subscales. In addition, the first component exhibits an eigenvalue of 1.80 and, although it represents a residual variance of 18.3% (PA) or 18.4% (NA) greater than the value proposed by Linacre (2011), suggests that the standardized residues do not exhibit additional systematic information. As the variance explained by PA (48.1%) and by NA (45.4%) classifies one of these dimensions as strong (Reckase, 1979), these results may be interpreted as indicators of the unidimensionality of each of the subscales, which is an assumption required by the Rasch model.

Table 3. PANAS (PA) Response Category Statistics

	Observed			MNSQ		
Category	Count	%	Average B_n	Infit	Outfit	F_k
1	246	5	-1.30	1.25	1.25	---
2	747	12	-.69	.90	.91	-2.15
3	1989	39	.27	.90	.90	-1.11
4	1531	30	1.07	.96	.99	.92
5	597	12	1.75	1.05	1.05	2.33

Note. B_n = person trait; F_k = step calibration; Infit = information-weighted mean square statistic; MNSQ = mean square fit indices; Outfit = outlier-sensitive mean square statistic.

The substantive validity refers to the diagnosis of the empirical operation of the response scale categories, to determine whether they function in accordance with what was expected by the authors of the instrument, upon development of the respective items (Wolfe & Smith, 2007). In the context of RSM, Linacre (2002) proposed a set of criteria to determine the efficiency of the response categories: (a) uniform distribution of the frequency of responses for different categories with at least 10 observations in each category; (b) monotonic progression of the observed mean measurement - Bn - and steps calibration - Fk - over the response categories; and (c) MNSQ outfit, more sensitive than the infit to unexpected responses, the response categories less than 2.0. Tables 3 and 4 summarize the statistics needed to as-

sess to what extent the five response categories of PANAS-P meet these criteria. The analysis leads to the conclusion that the structure of each response subscales meets the criteria established by Linacre (2002), as each category has an observed frequency response greater than 10 and an outfit value less than 2.0; furthermore, the observed mean measurements and step calibration increase monotonically along the 5 response categories.

Table 4. PANAS (NA) Response Category Statistics

	Observed			MNSQ		
Category	Count	%	Average B_n	Infit	Outfit	F_k
1	1181	23	-1.55	.99	1.00	---
2	1287	25	-.85	.93	.91	-1.30
3	1406	27	-.30	1.00	1.08	-.67
4	918	18	.31	.97	1.05	.43
5	368	7	.96	1.10	1.13	1.53

Note. Bn = person trait; Fk = step calibration; Infit = information-weighted mean square statistic; MNSQ = mean square fit indices; Outfit = outlier-sensitive mean square statistic.

The fitness of the response scale, by the efficient functioning of the 5 categories that comprise it, is supported by the results of the analysis performed to the adjustment of individuals. The mean and the standard deviation for the fit statistics for the PA subscale were 1.00 and .64 (infit), respectively, and 1.00 and .66 (outfit), respectively; in the case of NA subscale, the fit statistics exhibited mean and standard deviation values equal to 1.01 and .59 (infit), respectively, and 1.03 and .61 (outfit), respectively. Because the percentage of people with infit and/or outfit greater than 1.5, the value of the upper psychometric acceptability threshold for Rasch modeling (Wright & Linacre, 1994), is low in both subscales (18.0% and 18.6%, respectively PA and NA), the fit of the model is confirmed to be good. In fact, for most individuals, the parameters were estimated with adequate reliability; this is confirmed by the corresponding person separation reliability value, a fidelity estimator that measures the person's variance proportion that is not explained by measurement error (.80 in the two subscales).

Figures 1 and 2 demonstrate the individual-item representation in the same common metric scale (i.e., logits). The 10 items from each of the PANAS-P subscales are arranged in descending order of difficulty, here operationalized by the amount of

the attribute measured in the individual (i.e., PA or NA). In the case of the PA subscale, item 18 represents the greatest difficulty in choice, while item 15 is the easiest to subscribe, i.e., individuals with higher positive affect are more likely to choose the category 4 = "quite a bit" or category 5 = "extremely" for item 18, compared with individuals who exhibit low levels of the attribute. In the case of the NA subscale, item 16 presents greater difficulty in choice, whereas item 2 will be more easily chosen, i.e., individuals with more negative affect are more likely to choose category 4 = "quite a bit" or category 5 = "extremely" for item 16.

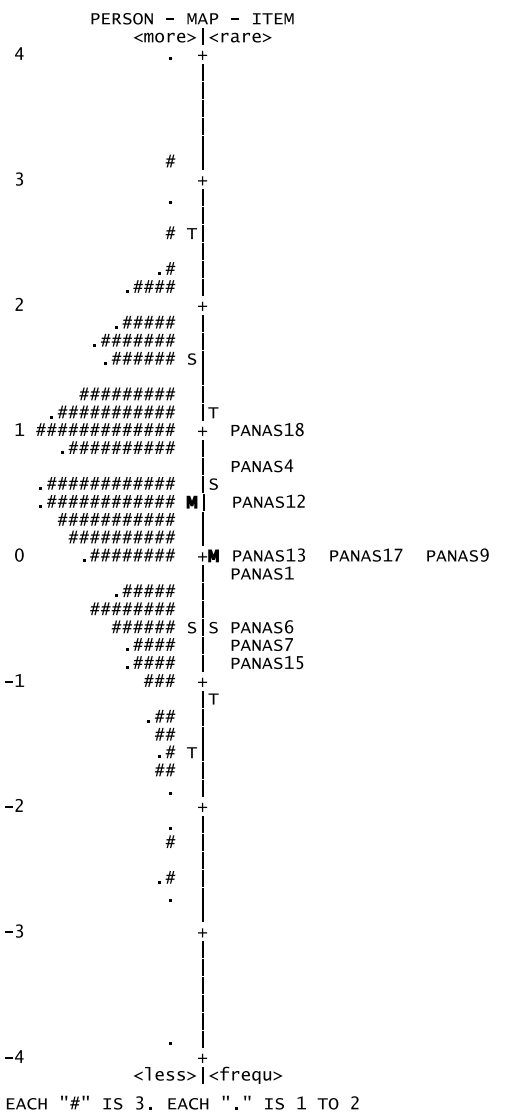

Figure 1. Joint Person and Item Representation along PANAS-P PA variable

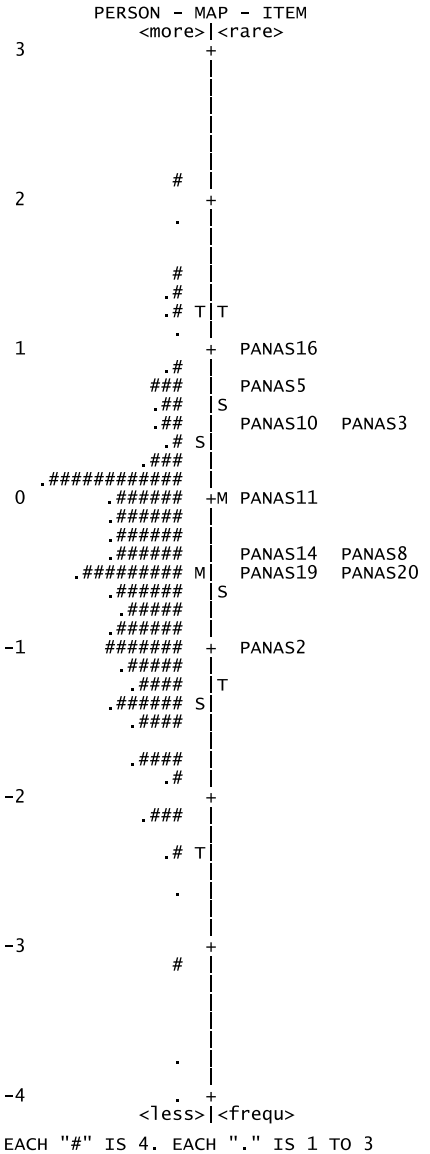

Figure 2. Joint Person and Item Representation along PANAS-P NA variable

It is therefore expected that the estimates of the individual's attribute and the item's difficulty overlap substantially so that the group of items can be considered appropriately directed to the sample of subjects. Whenever the difference between the mean of these estimates is less than 1 logit (Bond & Fox, 2007), as indeed happened in the PA subscale (.47) and in the NA subscale (- .54), the information contained in the items permits very reliable discrimination of individuals, in this case, at different levels of positive and negative affect.

In addition, analyses were performed on the differential item functioning (DIF) to assess the validity of the results of the PANAS-P on gender. For this purpose, the parameter of the standardized difference between the locations of males and females was calculated, followed by the adjustment of possible differences related to gender in the distribution of PA and NA. To this end, a Bonferroni procedure was used, which corrected the chosen significance level for the number of comparisons (.05/10) (Linacre, 2011). Following this conservative criterion, none of the 10 items of the PANAS-P subscales in males or females revealed location above .50, the value stipulated by Wright & Douglas (1975) as the cutoff point for the DIF contrast.

To establish the convergent and discriminant validity of the PANAS-P, we correlated the two dimensions of affect (PA and NA) with a measure of self-esteem (RSES; Rosenberg, 1965) in the logit scale. All correlations are statistically significant at a significance level of .001. The PA revealed a significant positive correlation with self-esteem (r = .48), whereas the NA exhibited a significant negative correlation (r = -.42). These results support the already established association between high PA and low NA and self-esteem (Brown & Marshall, 2001), following the direction of what was theoretically expected, and confirm the findings of previous studies with the PANAS-P with CTT (Miguel & Silva, 2012; in press).

Discussion

Rasch analyses were performed to examine the psychometric properties of PANAS-P. Data were analyzed using the RSM (Wright & Masters, 1982), which, as a Rasch model, has ideal metric properties for ranking item difficulty and person's ability (e.g., the level of the attribute measured) on a common scale. In addition to the joint measurement of people and items and provided that the data fit the model's requirements, Rasch modeling also allows for the comparison of people (items) regardless of the items (people) used in the measurement, which is property designated as *specific objectivity* by Rasch (Andrich, 1988).

Upon confirmation of the good fit of the data to the Rasch model, allowing the items and individual parameters to be estimated with high reliability, we proceeded to analyze the psychometric properties of the PANAS-P subscales in terms of their dimensionality. The main purpose was to test the basic assumption of uni-

dimensionality required by the Rasch model. The results of the present study reveal that 10 items from each of the subscales measure a single latent dimension. The Rasch analysis used in the present study provided, therefore, psychometric evidence to support the unidimensionality thesis of each of the two PANAS-P subscales.

The results suggest that the 5 response scale categories used in the PANAS-P function adequately with secondary school students, a conclusion that psychometrically supports the option based on CTT (Miguel & Silva, 2012; in press) and that is now confirmed by the Item Response Theory (IRT). The Rasch model provided substantive validity to the use of this response scale, allowing detailing of its function through the reliable manner in which the calibration of the five categories was performed.

With respect to the PANAS-P items, their distribution reflects a wide range of individual differences at the affect level. The mean PA level measured in the students is higher than the mean difficulty level of the items used to assess, verifying the reverse with NA. Nonetheless, the difficulty level of the 10 items for each subscale reflects an appropriate range of PA/NA levels in the students, allowing discrimination of different levels of this trait in the sample. The high internal consistency of people and items enabled an estimate of the respective parameters with high reliability, adjusting most of the response patterns to the model (Prieto & Delgado, 2003).

Following a conservative criterion, the DIF analysis performed as a function of gender revealed that the functioning of the PANAS-P items, both in the PA subscale or in the NA subscale, was constant in both genders, assumed difficult for both subgroups. The items are sufficiently robust to allow and assessment of the PA/NA independent of the gender of the individual who answered the scale. Therefore, it can be concluded that the answers only reflect the construct level of the individual, which is measured as a function of the difficulty of the items and not by other competencies explained by the gender to which the individual belongs.

Conclusion

Overall, this study confirms that the PANAS-P is a valid instrument for assessing the PA/NA in the Portuguese secondary school student population. Rasch modeling established the unidimensionality of both subscales, allowing people's levels of PA/NA and the item difficulties of the PANAS-P used to measure each con-

struct to be ranked on a single continuous scale, a ranking achieved with high precision based on the adequacy of the response scale with five alternatives and no sex bias.

In terms of clinical practice in counseling psychology, it is possible to consider the development of a computerized adaptive version of the test (CAT) that applies each of the subscales of the PANAS-P and ranks the ability to recommend an item based on the level of PA/NA revealed by the consultants and the intended measurement error criterion, an approach that can help streamlining the duration of intervention and minimize the costs of the process by adapting the items to the consultant's level of PA/NA.

The present study exhibits some limitations that should be overcome in future investigations, particularly with respect to the sample. Although the sample reproduces the demographic characteristics of the student population sought, it could not be considered representative of the diversity of Portugal. Future studies should utilize random sampling methods so that the results can be generalized and must assess the test-retest reliability of the PANAS-P.

Finally, having confirmed the psychometric robustness of the PANAS-P with the Rasch model among adolescents, the authors wish to proceed with further research, namely cross-cultural studies aimed at mapping the constructs of PA/NA in different cultures and nations, using other groups (adults), fields of application (work and health settings) and time frame retrospection. It will also be interesting to conduct more research with respect to the construct that is measured, studying its level of overlap with social desirability, for example.

References

Allik, J., & Realo, A. (1997). Emotional experience and its relation to the five-factor model in Estonian. *Journal of Personality, 65*, 625-647.

Andrich, D. (1988). *Rasch models for measurement*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-068. Newbury Park, CA: SAGE.

Balatsky, G., & Diener, E. (1993). Subjective well-being among Russian students. *Social Indicators Research, 28*, 225-243.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, J. D., & Marshall, M. A. (2001). Self-esteem and emotion: Some thoughts about feelings. *Personality and Social Psychology Bulletin, 27*, 575-584.

Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residual. *Educational and Psychological Measurement, 70*, 717-731.

Cloninger, C. R., Bayon, C., & Svrakic, D. M. (1998). Measurement of temperamental and character in mood disorders: A model of fundamental states as personality types. *Journal of Affective Disorders, 51*, 21-32.

Cortina, (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *29*, 32-62.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*, 384-392.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Engelen, U., De Peuter, S., Victoir, A., Van Diest, I., & Van Den Bergh, O. (2006). Verdere validering van de Positive and Negative Affect Schedule (PANAS) en vergelijking van tweenederlandstalige versies [Subsequent validation of the Positive and Negative Affect Schedule (PANAS) and comparison of two Dutch versions]. *Gedrag & Gezondheid: Tijdsschrift Voor Psychologie En Gezondheid*, *34*, 89-102.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology, 45*, 30-45.

Furr, R. M. (2011). *Scale construction and psychometrics*. London: SAGE.

Galinha, I., & Ribeiro, J. (2005). Contribution to the study of the Portuguese version of the Positive and Negative Affect Schedule (PANAS): II—Psychometric Study (Contribuição para o estudo da versão Portuguesa da Positive and Negative Affect Schedule (PANAS): II—Estudo psicométrico). *Análise Psicológica, 23*, 219-227.

Gaudreau, P., Sanchez, X., & Blondin, J.-P. (2006). Positive and negative affect states in a performance-related setting: Testing the factorial structure of the PANAS across two samples of French-Canadian participants. *European Journal of Psychological Assessment, 22*, 240-249.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hilleras, P. K., Jorm, A. F., Herlitz, A., & Winblad, B. (1998). Negative and positive affect among the very old: A survey on a sample age 90 years or older. *Research on Aging, 20*, 593-610.

Joiner, T. E., Sandin, B., Chorot, P., Lostao, L., & Marquina, G. (1997). Development and factor analytic validation of the SPANAS among women in Spain: (More) cross-cultural convergence in the structure of mood. *Journal of Personality Assessment, 68*, 600-615.

Kline, P. (2000). The future of personality measurement. In J. Mohan (Ed.), *Personality across cultures: Recent developments and debates* (pp. 336-351). New Delhi, India: Oxford University Press.

Krohne, H. W., Egloff, B., Kohlmann, C.-W., & Tausch, A. (1996). Investigations with a German version of the Positive and Negative Affect Schedule (PANAS). *Diagnostica, 42*, 139-156.

Lim, Y.-J., Yu, B.-H., Kim, D.-K., & Kim, J.-H. (2010). The Positive and Negative Affect Schedule: Psychometric properties of the Korean version. *Psychiatry Investigation, 7*, 163-169.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85-106.

Linacre, J. M. (2011). *Winsteps Rasch measurement computer program, version 3.73.0*. [Computer program.] Chicago, IL: Winsteps.com.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Miguel, J. P., & Silva, J. T. (2012). *PANAS—Portuguese version: A psychometric study*. Paper presented at the 8th Conference of the International Test Commission, Amsterdam.

Miguel, J. P., & Silva, J. T. (in press). Positive and Negative Affect Schedule—Portuguese European version: A psychometric study. *Psychological Reports*.

Pandey, R., & Srivastava, N. (2008). Psychometric evaluation of Positive and Negative Affect Schedule. *Industrial Psychiatry Journal, 17*, 49-54.

Plutchik, R. (1997). The circumplex as a general model of the structure of emotions and personality. In R. Plutchik & H. R. Conte (Eds.), *Circumplex models of personality and emotions* (pp. 17-45). Washington, DC: American Psychological Association.

Prieto, G., & Delgado, A. R. (2003). Análisis de un test mediante el modelo de Rasch [Analysis of a test using the Rasch model]. *Psicothema, 15*, 94-100.

Prieto, G., & Delgado, A. R. (2007). Measuring math anxiety (in Spanish) with the Rasch rating scale model. *Journal of Applied Measurement, 8*, 149-160.

Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350 -353.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120.

Smith E. V., Jr. (2004). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In: E. V. Smith Jr & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 575-600). Maple Grove, MN: JAM Press.

Tabchnick, B., & Fidell, L. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Pearson Education.

Terraciano, A., McCrae, R. R., & Costa, P. T., Jr. (2003). Factorial and construct validity of the Italian Positive and Negative Affect Schedule (PANAS). *European Journal of Psychological Assessment, 19*, 131-141.

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (PANAS). *Journal of Cross-Cultural Psychology*, *38*, 227-242.

Tuccitto, D. E., Giaccobbi, P. R., & Leite, W. L. (2010). The internal structure of positive and negative affect: A confirmatory factor analysis of the PANAS. *Educational and Psychological Measurement, 70*, 125-141.

Watson, D., & Clark, L. A. (1994). *The PANAS-X: manual for the Positive and Negative Affect Schedule—Expanded form*. Iowa City, IA: University of Iowa.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063-1070.

Waugh R. F., & Vhapman, E. S. (2005). An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: What is the difference? *Journal of Applied Measurement*, *6*, 80-99.

Widiger, T. A., & Frances, A. J. (1994). Toward a dimensional model for personality disorders. In P. T. Costa Jr., & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (pp. 19-39). Washington, DC: American Psychological Association.

Wolfe, E. W., & Smith, E. V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. In E. V. Smith Jr & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 243-290). Maple Grove, MN: JAM Press.

Wright, B. D., & Douglas, G. A. (1975). A better procedure for sample-free item analysis. *Research Memorandum*. Statistical Laboratory. Department of Education. Chicago, IL: University of Chicago.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

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

Wright, B. D., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. In E. V. Smith Jr & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 1-24). Maple Grove, MN: JAM Press.

Yamasaki, K., Katsuma, R., & Sakai, A. (2006). Development of a Japanese version of the Positive and Negative Affect Schedule for children. *Psychological Reports, 99*, 535-546.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (pp. 111-154). Westport, CT: ACE/Praeger.

Zevon, M. A., & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology, 43*, 111-122.

Positive and Negative Affect Schedule, European Portuguese Version (Panas-P): A Psychometric Study

Abstract

This study examines the psychometric properties of a Portuguese version of the PANAS (PANAS-P) that was directly developed from the original instrument using the international psychometric standards on a sample of high school students (*N* = 528). Item analyses performed for all theoretical subscales generally demonstrated adequate internal consistency, although one item in each subscale showed poor performance. After dividing the sample, exploratory and confirmatory factor analyses were successively conducted, and the original theoretical structures of the two affect dimensions were obtained for both samples. These findings suggest that the PANAS-P is a reliable and valid assessment of the affect dimensions of Portuguese adolescents in a school context. Future steps to validate the PANAS-P are discussed.

Keywords: Positive affect; negative affect; reliability; validity; factor analysis.

Submitted

Miguel, J. P., Silva, J. T. & Prieto, G. (submitted). Positive and Negative Affect Schedule, European Portuguese Version (Panas-P): A Psychometric Study. *Psychological Reports*.

Positive and Negative Affect Schedule, European Portuguese Version (Panas-P): A Psychometric Study

Two important traditions have influenced the study of affective states. On one hand, the categorical approach asserts that affective states are specific and conceptually different and that they comprise a small number of discrete emotions that must be measured independently (Ekman, 1993). On the other hand, the dimensional approach asserts that affective states are general and conceptually similar. Furthermore, this approach focuses on the dimensions that induce important associations among the fundamental emotions; thus, it suggests that affective states are best measured by correlating emotions with one another (Plutchik, 1997).

The *Positive and Negative Affect Schedule* (PANAS; Watson, Clark & Tellegen, 1988) is the affective state measure that is most widely used by the investigators who follow the dimensional approach. It is composed of two 10-item subscales that assess positive (PA) and negative (NA) affect, respectively, and may be administered to elicit either dispositional (trait) or situational (state) features, depending on the indicated length of retrospection (e.g., "right now", "today", "during the past few days", "during the past few weeks", "during the past year", or "in general").

According to its authors, the PANAS has theoretical and empirical foundations. Conceptually, PA and NA are conceived of as the dispositional activation of positive and negative affects, respectively. PA is a dimension including the enthusiasm, activation, and alertness that comprise a positive mood (e.g., joy, interest, enthusiasm, and alertness); high PA denotes a state of high energy, full concentration, and pleasurable engagement, whereas low PA is characterized by sadness and lethargy. NA is a general dimension including substantive distress and unhappy engagement

that comprises a variety of aversive moods (e.g., anger, contempt, disgust, fear, and nervousness); low NA is a state of calmness and serenity. Empirically, Watson et al. (1988) created the items that comprise the PANAS using the mood categories proposed by Zevon and Tellegen (1982).

The appropriateness of the reliability and validity of the PANAS (Watson & Clark, 1994; Watson et al., 1988), together with its formal characteristics, account for its wide use in clinical and non-clinical settings as well as in different samples of individuals. The increasing amount of cross-cultural affective research has promoted the development of studies that focus on the adaption and translation of the PANAS to languages other than English, of which several international versions are available (Allik & Realo, 1997; Balatsky & Diener, 1993; Engelen, De Peuter, Victoir, Van Diest, & Van Den Bergh, 2006; Gaudreau, Sanchez, & Blondin, 2006; Hilleras, Jorm, Herlitz, & Winblad, 1998; Joiner, Sandin, Chorot, Lostao, & Marquina, 1997; Krohne, Egloff, Kohlmann, & Tausch, 1996; Lim, Yu, Kim, & Kim, 2010; Pandey & Srivastava, 2008; Terraciano, McCrae, & Costa, 2003; Yamasaki, Katsuma, & Sakai, 2006).

Despite the widespread acceptance of the PANAS, its results are not unanimous, particularly with regard to its dimensionality. In fact, although some studies have replicated the two-factor structure suggested by the original PANAS authors (Crawford & Henry, 2004; Crocker, 1997; Joiner et al. 1997; Melvin & Molloy, 2000; Molloy, Pallant & Kantas, 2001; Terraciano et al., 2003), others suggest that three factors are involved (Gaudreau et al., 2006; Killgore, 2000; Leue & Beauducel, 2011; Mehrabian, 1997). In the three-factor solution, NA is divided into two conceptually significant factors: discomfort and fear. The relationship between PA and NA is also currently debated; some studies have replicated the orthogonal relationship asserted by the PANAS authors (Leue & Beauducel, 2011; Melvin & Molloy, 2000), whereas others have concluded that these factors are correlated (Crawford & Henry, 2004; Crocker, 1997; Gaudreau et al., 2006; Joiner et al., 1997; Killgore, 2000; Mehrabian, 1997; Molloy, et al., 2001; Terraciano et al., 2003).

The ambiguity of these results has different explanations (e.g., Tuccitto, Giacobbi & Leite, 2010). Some reasons are conceptual given that, although not included in the structure of the PANAS, Zevon and Tellegen's (1982) mood categories are first-order factors, whereas PA and NA were conceived of as second-order affect dimensions. Other reasons are methodological and suggest that the internal structure

of the PANAS might be sensitive to the types of samples, the period of time indicated for retrospection, or both. Still other reasons are statistical and based on the wide variation among the criteria applied to assess the factorial model parameters.

The Portuguese version of the PANAS was recently formulated (Galinha & Ribeiro, 2005), and some of its psychometric properties are known. Its authors aimed to attain the same affect descriptors by choosing to replicate the formulation process undertaken by Watson et al. (1988) (i.e., using Zevon and Tellegen's [1982] original 60 items, rather than translating the 20 items that comprise the PANAS). Upon attempting to establish the best factorial solution for the intercorrelation matrix observed in their study, these authors found that despite the results of their principal components analysis (PCA) that isolated two dimensions of affect, correspondence between the Portuguese and the original scales existed only for six PA items and three NA items.

Considering the limited comparability between the original scale and the version developed by Galinha and Ribeiro (2005), who used a sample of university students, we cannot know whether the psychometric divergence reported by these authors is real. Thus, the present study developed a literal equivalent of the original PANAS (Watson et al., 1988) using European Portuguese to attempt to control for a potentially confounding variable.

The present study aimed to collect empirical evidence on the psychometric properties of the European Portuguese translation of the original PANAS using a sample of adolescents attending high school. After estimating the internal consistency of the two subscales, the internal structure of the PANAS was subjected to exploratory (EFA) and confirmatory factor analyses (CFA) to elucidate its dimensionality and the type of relationship between PA and NA. In addition, we sought to assess which interrelation structure between the subscales exhibited a better fit to the data: orthogonal or oblique. Finally, the convergent/divergent validity of the subscales was investigated with regard to self-esteem.

Methods

Participants

The sample was composed of 528 high school students who attended the 10th, 11th, and 12th grades across several public schools in the central and southern regions

of Portugal. A total of 227 (43.1%) students were male, and 295 (55.9%) were female. Six participants did not report their gender. Their age ranged from 14 to 22 years old, corresponding to an average of 16.26 years (SD = 1.19). Three participants did not report their age. Data concerning ethnicity were not collected because this information is not usually collected in Portugal. Nevertheless, most respondents were Caucasian Europeans.

Measures

The first page of the questionnaire collected the demographic data of the students (gender, age, school grade, school success/failure, and education level of parents) and some measures regarding their vocational behavior (these variables were not used in the present study). In addition, data on the types of affect and the self-esteem level of the students were collected.

The affect types were assessed using the Portuguese version of the PANAS (Watson, Clark & Tellegen, 1988) translated from the original instrument by the first author. This survey is composed of two 10-item subscales. These items briefly measure PA and NA, respectively. The items consist of adjectives that indicate mood states related to PA and NA. Respondents assess the degree to which they experienced each of the 20 listed emotions over a specified time period on a Likert scale with five answer options: 1 = slightly or not at all, 2 = a little, 3 = moderately, 4 = quite a bit, and 5 = extremely. Although several periods of time may be used depending on the intended level of retrospection, the present study instructed participants to respond how they felt "during the past few weeks". For the purpose of description, the results of the PA and NA subscales were calculated by adding the 10 responses given by each participant to the corresponding items.

Self-esteem level was assessed using the *Rosenberg Self-Esteem Scale* (RSES; Rosenberg, 1965). That scale is composed of 10 items that consist of statements concerning self-worth and self-acceptance to which participants respond using a four-point Likert scale varying from 1 = strongly disagree to 4 = strongly agree. Because half of the items in the RSES are worded negatively, their results must be inverted when calculating the total score.

Procedure

The present study is a portion of a larger research project on the effects of self-efficacy and mathematical anxiety on the career decision making of high school students. In the present study, measurements of respondent affect were performed to investigate its incremental contribution to the prediction of the vocational variable. The PANAS was chosen to measure affect due to its popularity in the emotion literature, which largely derives from its satisfactory psychometric qualities. Due to the aforementioned reasons, we decided to create a parallel version of the original PANAS rather than using the European Portuguese version developed by Galinha and Ribeiro (2005).

The first author translated the instrument using the original PANAS, and the back-translation was performed by two independent bilingual colleagues who held PhDs in psychology according to the psychometric recommendations for the formulation and cross-cultural adaptation of trans-literal psychological scale equivalents (Hambleton, Merenda & Spielberger, 2005). A native English speaker compared the original and back-translated versions to clarify and reformulate the meaning of certain words.

The Education Ministry of Portugal approved the study. Authorization to perform the study was requested from the principals of the selected schools. Informed consent was requested from the students and their legal guardians after explaining the study aims. Participation was voluntary, anonymous, and confidential; the principal investigator supplied his professional contact details to provide a synthesis of the results to interested parties. The first author administered the questionnaires in the classroom in the presence of the participants' teachers; students were not offered any reward for their participation. All participants completed the questionnaires during the selected class period, and the students who refused to participate were allowed to leave the classroom before distributing the questionnaires.

Analyses

The analyses were performed without considering the missing results of the investigated variables.

To investigate the psychometric properties of the Portuguese version of the

PANAS (PANAS-P) that resulted from the direct translation of the original instrument, descriptive statistics (means and standard deviations as well as measures of symmetry and kurtosis) were calculated for each of the 20 scale items. Several tests were performed to identify univariate or multivariate outliers. Finally, the reliabilities of the PA and NA subscales were assessed using Cronbach's alpha (i.e., internal consistency with a cutoff score of 0.80; Nunnaly & Bernstein, 1994). SPSS 20.0 was used for all analyses.

The sample was randomly divided in two equivalent groups of participants using the randomization strategies supplied by SPSS. One subsample (N = 255) was used to perform EFA to establish the number and nature of the latent factors in the responses to the PANAS-P items using FACTOR 8.02 software (Lorenzo-Seva & Ferrando, 2006; 2011). PCA was applied to a matrix of polychoric correlations for ordinal data (Drasdow, 1988) because the PANAS-P items must be treated as ordinal variables as a function of the five categories of response. To establish the number of dimensions present in the latent space, an improvement of the *Parallel Analysis* (Horn, 1965) technique formulated by Timmerman and Lorenzo-Seva (2011) was used. Following Tabachnick and Fidell (2007), a value of .32 was established as the cutoff point to define the minimum factor saturation value to attribute items to factors.

The other subsample (N = 273) was used to test the factorial validity of the PANAS-P. CFA was performed using AMOS 20 (Arbuckle, 2011) to investigate the items' fit to the two-factor (PA and NA) theoretical structure suggested by Watson et al. (1988). The maximum likelihood (ML) method was used because it is robust with regard to non-normal data and independent observations (Tabachnick & Fidell, 2007). The confirmatory approach was chosen because the results of a previous study suggested that this same two-factor structure might be latent in the PANAS-P (Miguel & Silva, 2012). To investigate the models' goodness of fit, several absolute-fit, relative-fit, and population-based-fit indices usually recommended by the specialized literature were used in addition to the chi-square test. Specifically, the *root mean square residual* (RMR; Hu & Bentler, 1999), the *goodness-of-fit index* (GFI; Bentler, 1990), the *comparative fit index* (CFI; Bentler, 1990) and the *root mean square error of approximation* (RMSEA; Steiger, 1990) were used. A model is considered to fit the data when the criteria for two such indices are met (Blunch, 2008). The factor loadings and inter-factor correlations were tested at a significance level of .05.

Two models were tested. Model 1 tested the hypothesis stating that the PANAS-P items follow a two-dimensional model in which PA and NA are not correlated (Watson et al., 1988). To compare with the results of previous studies that describe PA and NA as non-orthogonal dimensions (e.g., Crawford & Henry, 2004), the two-factor Model 2 posited that PA and NA are correlated.

Results

Preliminary analyses showed that the items in both PANAS-P subscales exhibited approximately normal distributions. The symmetry values of PA and NA varied from −.422 to .171 and from −.343 to 1.085, respectively, whereas the kurtosis varied from −.564 to .428 and from −1.118 to .322, respectively. Mahalanobis' critical distance value was used for all 20 variables ($\chi^2_{(20)} = 45.31, p < .001$), and only one outlier was identified. In addition, a univariate analysis of the PANAS-P items using Tabachnick and Fidell's (2007) *Z*-score < 3.00 criterion did not reveal anomalies. Based on these preliminary analyses, the final sample included all 255 students.

Table 1 describes the reliability values and descriptive statistics that correspond to the PA and NA subscales in the total sample, distributed by gender. The average PA values tend to be higher than the NA values in the full sample and when divided by gender; significant differences were found between boys and girls. The boys scored significantly higher on the PA subscale ($t_{(252)} = 2.21, p = .014$; Cohen's d = .29) than girls, whereas the opposite results were found for the NA subscale ($t_{(252)} = - 3.23, p < .001$; Cohen's d = .41). Although these gender differences were not observed in the North American normative sample (Watson et al., 1988), they match those identified by Crawford and Henry (2004) in a sample of English adults and by McCrae and Costa (2003) in a sample of Italian university students and adults.

Table 1. Descriptive statistics and reliabilities of the PANAS-P (N = 255)

PANAS-P	Total		Women		Men		Cronbach's
scales	*M*	*SD*	*M*	*SD*	*M*	*SD*	α
PA	31.99	6.11	31.25	6.45	32.95	5.43	.83 (.88)
NA	25.68	7.59	26.96	7.59	23.89	7.28	.85 (.87)

Note. The American normative values (Watson et al., 1988) are provided in parentheses. PA = Positive Affect; NA = Negative Affect.

Table 2. PANAS-P reliability and item-level analysis (N = 255)

Items				Corrected item-total correlation	Squared multiple correlation	Alpha if item deleted
PA	NA	*M*	*SP*			
1. Interested		3.38	0.83	.495	.346	.812
	2. Upset	3.45	1.09	.653	.486	.832
	3. Scared	2.14	1.14	.663	.513	.831
4. Proud		2.74	1.01	.522	.329	.809
	5. Ashamed	1.99	1.10	.382	.164	.855
6. Determined		3.43	0.95	.681	.522	.792
7. Active		3.60	0.98	.605	.433	.800
	8. Distressed	2.99	1.31	.692	.556	.827
9. Strong		3.18	1.03	.500	.335	.811
	10. Hostile	2.16	1.04	.213	.129	.867
	11. Irritable	2.43	1.13	.577	.419	.839
12. Inspired		2.98	1.01	.624	.469	.798
13. Attentive		3.17	0.92	.445	.334	.816
	14. Afraid	2.84	1.19	.672	.598	.830
15. Excited		3.73	1.01	.550	.461	.806
	16. Guilty	1.84	1.04	.478	.283	.847
17. Enthusiastic		3.22	0.96	.642	.521	.796
18. Alert		2.55	1.06	.115	.030	.851
	19. Nervous	2.93	1.26	.779	.664	.819
	20. Jittery	2.92	1.19	.456	.260	.849

Note. PA = Positive Affect; NA = Negative Affect.

Table 1 also describes the internal consistency of the PANAS subscales, which are satisfactory according to the aforementioned criteria. The Cronbach's α coefficients were similar to those found in the original studies (Watson et al., 1988), although the present study of Portuguese adolescents exhibited an inversion in which the internal consistency of the NA subscale was higher than that of the PA subscale. This result matches the findings of Terraciano et al. (2003). Although separate item analyses for each subscale were homogeneous (Table 2), two items were problematic from a psychometric perspective because the average inter-item correlations were each over .30. The values of PA Item 18 ("alert") and NA Item 10 ("hostile") were lower than the minimum psychometric acceptance threshold (.30); thus, their ability to discriminate PA from NA is poor. Therefore, the predicted percentage of variability in the responses to those items based on the results of the remainder of the

items in the corresponding subscale is lower than the minimum required value (.20). Those two items represent the most serious psychometric flaws because their exclusion from the corresponding scale increases the α coefficient compared with the one calculated that includes them.

Table 3. Item statistics, PCA loadings, and communalities of the PANAS-P (N = 255)

Items	PA	NA	h^2
6. Determined	**.710**	−.055	.507
17. Enthusiastic	**.665**	−.105	.453
7. Active	**.657**	−.184	.465
12. Inspired	**.646**	−.062	.421
1. Interested	**.581**	.093	.346
15. Excited	**.571**	−.311	.423
4. Proud	**.569**	−.066	.328
13. Attentive	**.538**	.098	.299
9. Strong	**.527**	−.208	.321
18. Alert	.198	**.431**	.225
19. Nervous	−.182	**.830**	.722
2. Upset	.034	**.725**	.527
14. Afraid	−.117	**.724**	.538
8. Distressed	−.210	**.712**	.552
3. Scared	−.149	**.706**	.521
11. Irritable	−.245	**.557**	.370
16. Guilty	−.340	**.521**	.387
5. Ashamed	−.008	**.504**	.254
20. Jittery	−.016	**.492**	.242
10. Hostile	−.015	**.301**	.091
Eigenvalue	5.24	2.76	—
Explained variance (%)	21.60	18.40	—
Reliability[a]	.89	.85	—

Note. PA = Positive Affect; NA = Negative Affect. h^2 = Communalities. Loadings over .30 are in boldface. The percentage of variance accounted for by each factor is taken after the rotation.

[a]Reliability estimates were computed using the computer program Factor (Lorenzo-Seva & Ferrando, 2006)

An exploratory PCA with varimax rotation was performed to complement the identification of the items with unsatisfactory performance. The specification of a two-component solution resulted in a structure with two factors that corre-

sponded to PA and NA (Table 3). Overall, the items exhibited strong primary loadings on the appropriate factor and acceptably low secondary loadings on the other. Only Items 15 ("excited") and 16 ("guilty") of the PA and NA subscales, respectively, exhibited cross loadings above the cutoff point (|.25|) employed by Watson et al. (1988) in constructing the PANAS. These results might reduce the independence of the scales. Contrary to our theoretical expectations, Item 18 ("alert") loaded onto the NA subscale, and its communality value approximately coincided with the psychometric acceptance threshold. This result confirms its weak discriminant ability found in the item analysis. Although Item 10 ("hostile") loaded onto the theoretically predicted dimension, it exhibited the lowest loading among all the adjectives in the scale, and it was the only one whose communality value was significantly lower than .20 (i.e., it did not meet the established criterion for psychometric acceptance). Overall, the factorial structure of the PANAS-P closely reproduced the dimensionality of the original scale. The correlation between the PA and NA subscales was negative and significant ($r = -.19$, $p < .001$). Thus, an association existed between high PA scores and low NA scores.

The second sample ($N = 273$) was used to test the factorial validity of the PANAS-P based on theoretical and statistical criteria. The analysis of the item distribution properties did not reveal any deviations from univariate normality within any subscale. The PA and NA symmetry values varied from −.455 to −.016 and from −.278 to .759, respectively, and the kurtosis values varied from −.948 to .369 and from −1.005 to −.362, respectively. These results agree with the fact that serious univariate outliers are not present. However, applying Mahalanobis' critical distance value for all 20 variables ($\chi^2_{(20)} = 45.31$, $p < .001$) revealed six multivariate outliers. Because these individuals were also univariate outliers (i.e., some of the items exhibited Z-scores ≥ 3.00; Tabachnick & Fidell, 2007), they were excluded from the sample. The final sample was composed of 267 students.

Table 4 describes the goodness-of-fit statistics of the tested models and shows that the two-dimensional model with two orthogonal factors (Model 1) was rejected using the various fit indices. The χ^2, RMR, and RMSEA values were high, whereas the GFI and CFI values were low. PA and NA covariation was specified to allow for their mutual correlation (Model 2). Although this change improved the goodness-of-fit indices, the model was rejected (Table 4).

Based on the results described above, the model that allowed PA and NA to correlate was kept because it exhibited a better fit. The two-dimensional model that included two non-orthogonal factors (Model 3) differed from Model 2 by the exclusion of NA Item 10 and PA Item 18 (i.e., the items that exhibited poor psychometric properties in the previous analysis). Although the values of the goodness-of-fit indices approached the psychometric acceptance threshold, Model 3 was also rejected.

Table 4. Goodness-of-fit estimations for the PANAS-P two-factor model (N = 267)

Model	χ^2	*df*	*p*	χ^2/df	RMR	GFI	CFI	RMSEA
Model 1 Two-factor uncorrelated	487.62	170	< .001	2.87	.111	.84	.80	.084
Model 2 Two-factor correlated (all 20 PANAS items)	472.47	169	< .001	2.80	.088	.85	.81	.082
Model 3 Two-factor correlated (Items 10 and 18 deleted)	365.98	134	< .001	2.73	.079	.86	.85	.081
Model 4 Two-factor correlated (Items 10 and 18 deleted; Error terms allowed to intercorrelate[a])	296.24	130	< .001	2.28	.075	.89	.89	.069

Note. RMR = *root mean square residual*; GFI = *goodness-of-fit index*; CFI = *comparative fit index*; RMSEA = *root mean square error of approximation*.

[a] Error terms from the items that were theoretically determined to share substantial content were allowed to intercorrelate with the assumption that the items of the correlated error variances were conceptually similar.

Given these results, we surveyed the modification indices supplied by the software in an attempt to possibly improve the goodness of fit. Therefore, Model 3 was re-specified to include some of the students' idiosyncrasies in the items that belonged to the same psychological factors. Thus, the errors of the items were allowed to correlate because they were conceptually similar on the content analysis (Mueller & Hancock, 2007). Thus, Model 4 allowed for the intercorrelation of PA Items 1 (interested) and 9 (strong), 1 (interested) and 13 (attentive), and 13 (attentive) and 15 (excited) as well as NA Items 3 (scared) and 16 (guilty). The data in Table 4 conclude that the fit of Model 4 was acceptable. Thus, contrary to Watson et al.'s (1988) assumptions, the dimensions of PA and NA were moderately interdependent ($r = -.28$, $p < .001$) in the present study, which agrees with the conclusions of Crawford and Henry (2004).

To establish the convergent and discriminant validity of the results obtained using the revised measure, a correlation between the factors relative to the 18 items in the PANAS-P and the RSES (Rosenberg, 1965) was assessed. For the sake of comparison, the results of the correlation between the 20-item PANAS-P and the RSES are also described. All the correlations were significant at the .001 level. The PA factor was significantly and positively correlated with self-esteem (r = .479, $r_{\text{PANAS-P-20}}$ = .449), whereas the NA factor was significantly and negatively correlated with this construct (r = −.545, $r_{\text{PANAS-P-20}}$ = −.529). These findings provide additional support to the previously identified association between high PA (or low NA) and self-esteem (Brown & Marshall, 2001). This result met our theoretical expectations.

Discussion

The primary goal of the present study was to assess the psychometric properties of the PANAS in a sample of Portuguese adolescents by analyzing the reliability of its factor structure and its hypothetical two-factor model fit to the data (Watson et al., 1988).

Although the internal consistency of the PANAS-P subscales was appropriate and attained values similar to those of the original scale, two items proved to be problematic. The exclusion of the PA item "alert" and the NA item "hostile" increased the reliability coefficient of the corresponding subscales; thus, these items are not satisfactory indicators of the dimension they are intended to assess.

An EFA confirmed these results by showing that these items required reformulation despite replicating the original factorial structure (Watson et al., 1988). The item "alert" did not perform well and loaded onto a factor that was not predicted by theory, which corroborates the findings of other studies (Gaudreau et al., 2006; Villodas, Villodas & Roesch, 2011). Perhaps this finding is due to a lack of understanding regarding the meaning of this term among participants. In fact, we suspect that some adolescents did not know the exact meaning of this term and responded to its meaning in general terms rather than to its specific affective content. Although the item "hostile" loaded onto the corresponding factor, its factor loading coincided with the psychometric acceptance threshold, whereas two other items, namely, "excited" (PA) and "guilty" (NA) were correlated with the non-corresponding factor. These results might reduce the independence of the subscales.

After testing the reliability of the factorial structure in the PANAS-P, the CFA revealed that the model formulated by Watson et al. (1988) did not fit the data (Model 1). In agreement with the conclusions of Crawford and Henry (2004), Gaudreau et al. (2006), and Terraciano et al. (2003), an alternative model that assumed a moderate correlation between the two factors (Model 2) was tested. Although its goodness-of-fit improved, the model fit remained mediocre. A reformulation of Model 2 excluded the two items that were rated as problematic based on previous item analysis. This model, Model 3, represented a slight improvement with regard to goodness-of-fit; however, the fit remained mediocre. An acceptable fit was achieved only after the model accounted for the intercorrelations among the error terms of the items that shared theoretical content, based on the assumption that these items are conceptually similar (Crawford & Henry, 2004). According to Tuccitto et al. (2011), these correlations represent latent factors that are not specified in the model (and not measurement error); thus, the most effective solution to this problem includes specifying Zevon and Tellegen's (1982) mood categories as first-order latent factors and PA and NA as the second-order factors in the PANAS CFA model.

Given the theories that underlie the PANAS and its widespread use as an affect self-assessment, the inability to attain goodness-of-fit indices that met the criteria formulated by Hu and Bentler (1999) might indicate that the orthogonal two-factor structure does not generalize across cultures, especially for studies that have used translated versions of the original scale. Therefore, the clinical use of the PANAS-P among adolescents must be treated with caution, and more studies are needed to compensate for the limitations identified in the present study before its clinical use can be recommended.

In addition, the results of the present study revealed that the two dimensions of the PANAS-P are correlated with self-esteem, as our theory predicted. The findings of the present study match those of other authors (Brown & Marshall, 2001) and confirm the associations between self-esteem and high PA as well as low NA. These results allow positive and negative affects to predict psychological results.

Conclusions

Overall, the present study confirms that the PANAS-P is a reliable assessment of the affective states of Portuguese high school students. Our study showed that the

validity of the PANAS-P exhibits certain limitations, especially in regard to construct validity. Although the modified two-factor model failed to meet the rigorous goodness-of-fit criteria formulated by Hu and Bentler (1999), it nonetheless provided evidence supporting the acceptability of the factorial structure of the PANAS-P. In agreement with the recent studies that applied CFA to the structure of the original instrument, the results of the present study confirm that the hierarchical conception of affect (Watson & Clark, 1994) might be met via two first-order latent factors. However, considering that the correlations among the errors of the items might represent latent factors that are not specified in the model, future research must test second-order CFA models specifying these mood categories (Zevon & Tellegen, 1982) as first-order factors as well as PA and NA as second-order factors. Furthermore, the present study must be replicated using larger samples of other populations (e.g., adults) because the results might differ from those obtained in the current adolescent sample.

In fact, sampling issues relative to the number and diversity of participants limit the present study; future research should overcome this limitation. The present study used a convenience sample whose size might not have been sufficient to attain a stable factorial solution (Guadagnoli & Velicer, 1988); thus, caution should be taken when generalizing the results. Importantly, however, the sample size of the present study (even when split into two subsamples) surpassed the ratio of respondents to items (10:1) that is usually recommended for factor analyses. The modifications that were introduced to the model to achieve a satisfactory data fit also compromise the generalization of the current results. Although these modifications were conceptually and statistically warranted, the interpretation of the results requires caution. Finally, future studies must also assess the test-retest reliability of the PANAS-P.

The validation of the PANAS-P among adolescents is relevant for research in Portugal and for cross-cultural studies because it allows the two-factor model to be compared across more languages and cultures.

References

Allik, J., & Realo, A. (1997) Emotional experience and its relation to the five-factor model in Estonian. *Journal of Personality*, 65, 625-647.

Arbuckle, J. L. (2011) Amos (Version 20) [Computer Program]. Chicago: SPSS.

Balatsky, G., & Diener, E. (1993) Subjective well-being among Russian students. *Social Indicators Research*, 28, 225-243.

Bentler, P. M. (1990) Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246.

Blunch, N. J. (2008) *Introduction to structural equation modeling using SPSS and AMOS*. Thousand Oaks, CA: SAGE.

Brown, J. D. & Marshall, M. A. (2001) Self-esteem and emotion: Some thoughts about feelings. *Personality and Social Psychology Bulletin*, 27, 575-584.

Crawford, J. R. & Henry, J. D. (2004) The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, 43, 245-265.

Crocker, P. R. (1997) A confirmatory factor analysis of the Positive and Negative Affect Schedule (PANAS) with a youth sport sample. *Journal of Sport & Exercise Psychology*, 19, 91-97.

Drasdow, F. (1988) Polychoric and polyserial correlations. In S. Kotz & N. L. Johnson (Eds.), Encyclopedia of statistical sciences (Vol. 7) (Pp. 69-74). New York: Wiley.

Ekman, P. (1993) Facial expression and emotion. *American Psychologist*, 48, 384-392.

Engelen, U., De Peuter, S., Victoir, A., Van Diest, I., & Van Den Bergh, O. (2006) Verdere validering van de positive and negative affect schedule (PANAS) en vergelijking van twee Nederlandstalige versies. *Gedrag & Gezondheid: Tijdsschrift Voor Psychologie En Gezondheid*, 34, 89-102.

Galinha, I., & Ribeiro, J. (2005) Contribuição para o estudo da versão Portuguesa da Positive and Negative Affect Schedule (PANAS): II – Estudo psicométrico [Contribution to the study of the Portuguese version of Positive and Negative Affect Schedule (PANAS): II – Psychometric study]. *Análise Psicológica*, 23, 219-227.

Gaudreau, P., Sanchez, X., & Blondin, J.-P. (2006) Positive and negative affect states in a performance-related setting. Testing the factorial structure of the PANAS across two samples of French-Canadian participants. *European Journal of Psychological Assessment*, 22, 240-249.

Guadagnoli, E., & Velicer, W. F. (1988) Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103, 265-275.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.) (2005) *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hilleras, P. K., Jorm, A. F., Herlitz, A., & Winblad, B. (1998) Negative and positive affect among the very old: A survey on a sample age 90 years or older. *Research on Aging*, 20, 593-610.

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

Hu, L.-T., & Bentler, P. M. (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria vs. new alternatives. *Structural Equation Modeling*, 6, 1-55.

Joiner, T. E., Sandin, B., Chorot, P., Lostao, L., & Marquina, G. (1997) Development and factor analytic validation of the SPANAS among women in Spain: (More) Cross-cultural convergence in the structure of mood. *Journal of Personality Assessment*, 68, 600-615.

Killgore, W. S. (2000) Evidence for a third factor on the Positive and Negative Affect Schedule in a college student sample. *Perceptual and Motor Skills*, 90, 147-152.

Krohne, H. W., Egloff, B., Kohlmann, C.-W., & Tausch, A. (1996) Investigations with a German version of the Positive and Negative Affect Schedule (PANAS). *Diagnostica*, 42, 139-156.

Leue, A., & Beauducel, A. (2011) The PANAS structure revisited: On the validity of a bifactor model in community and forensic samples. *Psychological Assessment*, 23, 215-225.

Lim, Y.-J., Yu, B.-H., Kim, D.-K., & Kim, J.-H. (2010) The Positive and Negative Affect Schedule: Psychometric properties of the Korean version. *Psychiatry Investigation*, 7, 163-169.

Lorenzo-Seva, U. & Ferrando, P. J. (2006) FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers*, 38, 88-91.

Lorenzo-Seva, U. & Ferrando, P. J. (2011) *Manual of the program FACTOR v.8.02*. Tarragona: Universitat Rovira I Virgili.

Mehrabian, A. (1997) Comparison of the PAD and the PANAS as models for describing emotions and for differentiating anxiety from depression. *Journal of Psychopathology and Behavioral Assessment*, 19, 331-357.

Melvin, G. A., & Molloy, G. N. (2000) Some psychometric properties of the Positive and Negative Affect Schedule among Australian youth. *Psychological Reports*, 86, 1209 -1212.

Miguel, J. P. & Silva, J. T (2012) *PANAS — Portuguese version: A psychometric study*. Paper presented at The 8th Conference of The International Test Commission, Amsterdam.

Molloy, G. N., Pallant, J. F., & Kantas, A. (2001) A psychometric comparison of the Positive and Negative Affect Schedule across age and sex. *Psychological Reports*, 88, 861-862.

Mueller, R. O. & Hancock, G. R. (2007) Best practices in structural equation modeling. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (Pp. 488-508). New York: SAGE.

Nunnally, J. C., & Bernstein, I. J. (1994) *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Pandey, R., & Srivastava, N. (2008) Psychometric evaluation of Positive and Negative Affect Schedule. *Industrial Psychiatry Journal*, 17, 49-54.

Plutchik, R. (1997) The circumplex as a general model of the structure of emotions and personality. In R. Plutchik & H. R. Conte (Eds.), *Circumplex models of personality and emotions* (Pp. 17-45). Washington, DC: APA.

Roesch, S. C. (1998) The factorial validity of trait positive affect scores: Confirmatory factor analyses of unidimensional and multidimensional models. *Educational and Psychological Measurement*, 58, 451-466.

Rosenberg, M. (1965) *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Steiger, J. H. (1990) Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25, 173-180.

Tabachnick, B. G., & Fidell, L. S. (2007) *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education / Allyn and Bacon.

Terraciano, A., McCRae, R. R., & Costa, P. T., Jr. (2003) Factorial and construct validity of the Italian Positive and Negative Affect Schedule (PANAS). *European Journal of Psychological Assessment*, 19, 131-141.

Timmerman, M. E. & Lorenzo-Seva, U. (2011) Dimensionality assessment of ordered polytomous items with Parallel Analysis. *Psychological Methods*, 16, 209-220.

Tuccitto, D. E., Giaccobbi, P. R., & Leite, W. L. (2010) The internal structure of positive and negative affect: A confirmatory factor analysis of the PANAS. *Educational and Psychological Measurement*, 70, 125-141.

Villodas, F., Villodas, M. T., & Roesch, S. (2011) Examining the factor structure of the Positive and Negative Affect Schedule (PANAS) in a multiethnic sample of adolescents. *Measurement and Evaluation in Counseling and Development*, 44, 193-203.

Watson, D., & Clark, L. A. (1994) *The PANAS-X: manual for the positive and negative affect schedule – Expanded form*. The University of Iowa.

Watson, D., Clark, L. A., & Tellegen, A. (1988) Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54, 1063-1070.

Yamasaki, K., Katsuma, R., & Sakai, A. (2006) Development of a Japanese version of the Positive and Negative Affect Schedule for children. *Psychological Reports*, 99, 535 -546.

Zevon, M. A. & Tellegen, A. (1982) The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology*, 43, 111-122.

Rosenberg Self-Esteem Scale—Portuguese European version (RSES-P): A Rasch analysis

Abstract

Originally developed and presented in English, the *Rosenberg Self-Esteem Scale* (RSES) is a widely used self-report instrument for assessing individual self-esteem. This study examines the psychometric properties of a Portuguese version of the RSES (RSES-P), directly developed from the original instrument using the psychometric standards for developing translated and cross-cultural equivalent adaptations of psychological instruments on a sample of high school students. The results indicate that the 10 items of the RSES are well fitted to a latent unidimensional structure, as required by Rasch modeling. The response scale (four categories) showed proper functioning; therefore, the people and item parameters could be estimated with high precision (.81 and .99, respectively). Differential item functioning (DIF) analysis confirmed that there were no differences in the results of the RSES-P concerning gender. Finally, psychometric implications derived from the results of the present study are discussed, and suggestions are provided for future investigations.

Key-words: Self-esteem; dimensionality; Rasch analysis; rating scale model; Winsteps

Submitted

Miguel, J. P., Silva, J. T. & Prieto, G. (submitted). Rosenberg Self-Esteem Scale—Portuguese European version (RSES-P): A Rasch analysis. *Journal of Applied Measurement*.

Rosenberg Self-Esteem Scale—Portuguese European version (RSES-P): A Rasch analysis

1. Introduction

From the conceptual point of view, self-esteem is an evaluative component of self-concept (Rosenberg, 1979) that corresponds to the evaluation individuals make of their self-image based on the information they receive during personal interactions associated with the performance of various social roles (Brown, Collins, & Schmidt, 1988). Such global self-assessment results from the integration of specific evaluations each individual performs according to the relative significance he or she attributes to these factors based on his or her individual aspirations and ideals (Marsh, 1993). As previously established, this global sense of worth might coexist with a variety of more specific senses of worth associated with specific accomplishments or areas of individual competence. Nevertheless, global and specific self-esteem may not be separately deduced, as they are neither equivalent nor interchangeable (Rosenberg, Schooler, Schoenbach, & Rosenberg, 1995). To adjust the discrepancy between global and specific self-esteem, both being components of self-concept, a hierarchical model was suggested (Shavelson, Hubner, & Stanton, 1976) and subjected to empirical validation (Byrne & Shavelson, 1996).

By representing self-concept as hierarchically organized and interrelated parts and components (Rosenberg, 1979), any adequate assessment strategy requires a measure capable of globally assessing this construct, thus allowing for a global view of the attitudes that are positive and negative relative to the *self*. Furthermore, the social scientists that study self-esteem tend to prioritize global self-esteem. Among the manifold measures formulated to assess global self-esteem, the *Rosenberg Self-Esteem*

Scale (RSES; Rosenberg, 1965) is the most widely used (Blascovich & Tomaka, 1991; Byrne, 1996). Conceived of as a unidimensional instrument, RSES was originally designed as a Guttman scale, although it is widely applied using a Likert scale to measure global self-esteem. The popularity of this self-report instrument is due to several acknowledged advantages, including quick application as a function of its small number of items (10), the accessibility of the language used (equivalent to fifth-grade level), and the corresponding face validity.

The instrument's relative simplicity and accessibility, in addition to the increasing interest of cross-cultural research on self-esteem (Schmitt & Allik, 2005), contributed to the dissemination of studies that were devoted to adapt RSES to languages other than English and that resulted in the formulation of several international versions of this scale (Cheng & Hamid, 1995; Franck, De Raedt, Barbez & Rosseel, 2008; Kamakura, Ando & Ono, 2001; Martín-Albo, Nuñez, Navarro & Grijalvo, 2007; Prezza, Trombaccia, & Armento, 1997; Pulmann & Allik, 2000; Roth, Decker, Herzberg & Brähler, 2008; Shapurian, Hojat & Nayerahmadi, 1987; Šmídová, Hátlová, & Stochl, 2008; Vallieres & Vallerand, 1990), including versions in Portuguese (Santos & Maia, 2003; Vasconcelos-Raposo, Fernandes, Teixeira & Berteli, 2012) resulting from independent efforts by various authors.

The study of RSES psychometric properties in different cultural settings confirmed the adequacy of its internal consistency (Gray-Little, Williams & Hancock, 1997; Schmitt & Allik, 2005) as measured by Cronbach's alpha. However, in the terms of factorial structure, the conclusions of several studies that applied exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are not consensual (Corwyn, 2000). While some authors were able to confirm the unidimensionality of RSES (Martín-Albo et al., 2007; O'Brien, 1985; Pullmann & Allik, 2000; Santos & Maia, 2003; Vasconcelos-Raposo et al., 2012), others found that the scale appears to reflect a two-dimensional construct (Shevlin, Bunting & Lewis, 1995; Vallieres & Vallerand, 1990). In addition, some authors believe that this two-dimensional character might actually represent a methodological artifact resulting from the items' formulation, as five of them are negative assertions to control for eventual bias resulting from acquiescence and/or the social desirability response bias (Greenberger, Chen, Dimitrieva & Farragia, 2003; Marsh, 1996; Tomás & Oliver, 1999).

Moreover, most of the psychometric studies of RSES are based on the assumptions underpinning the classical test theory (CTT) and do not integrate the innovations introduced by the new measurement models associated with the item response theory (IRT), particularly, Rasch modeling. Given their intrinsic limitations, factor analysis and reliability estimates based on Cronbach's alpha have been criticized as strategies to investigate the construct validity of instruments for psychological assessment (Embretson & Reise, 2000; Sijtsma, 2009; Tabachnick & Fidell, 2001; Waugh & Chapman, 2005).

To summarize, the attractive features of RSES notwithstanding, some of aspects still require close attention, as the psychometric studies investigating its internal consistency and factorial structure do not allow for thorough item analysis. For instance, knowledge of the items' degree of precision to discriminate among people with different levels of self-esteem (low, average, or high) is necessary. Thorough analysis of RSES psychometric properties at the item level is indeed possible—but through using IRT-based models only, as they associate an individual's response probabilities with his or her level relative to the latent construct (e.g., self-esteem) that the instrument for psychological assessment is intended to measure (Embretson & Reise, 2000).

From the various methods developed within the framework of IRT, Rasch modeling stands out as particularly appropriate (Rasch, 1960). By transforming ordinal data into an interval scale, the Rasch method affords a psychometrically sound alternative to the sum of item scores that is characteristic of Likert scales (Wright & Masters, 1982). Consequently, Rasch modeling enables testing the hypothesis that RSES items constitute a unidimensional variable, calibrating the magnitude of the differences among items on an interval scale and thus assessing each individual relative to the resulting newly created variable (Fox & Jones, 1998; Prieto & Delgado, 2007).

A single study in the literature applied Rasch modeling to the analysis of RSES psychometric properties, and that was conducted with a sample of university students from Portugal (Quintão, Prieto & Delgado, 2011). The results of that study confirmed the scale's unidimensionality and the adequacy of the number of categories included in the response scale. However, the authors concluded that two items in the scale exhibited differential item functioning (DIF), one favorable to the male gender and the other to the female gender.

The aim of the present study was to replicate the investigation by Quintão et al. (2011) with a sample of younger respondents. As a function of the necessity to perform studies assessing the equivalence of RSES among non-native English speaking populations (Sinclair, Blais, Gansler, Sandberg, Bistis, & LoCicero, 2010), the present study aimed to perform a psychometric evaluation of the RSES-P within the framework of IRT to provide an important approach to item calibration. This subject-independent statistical approach reveals the idiosyncrasies of Portuguese adolescent samples. IRT modeling elucidates the individual response patterns and the difficulty parameters obtained for each item, providing information on which are the easiest and the hardest. Moreover, IRT modeling has important clinical advantages, as it enables professionals to understand the patient's behavior regarding a difficult or easy item, which is helpful for intervention purposes and for normative data (Embretson & Reise, 2000).

This study aimed to probe the psychometric quality of the RSES-P. However, given the methodological limitations previously identified in CTT, the authors chose to resort to IRT using the Rasch model in an attempt to gather further evidence for measuring the properties of the Portuguese version of the RSES. In this paper, the authors propose to use Rasch model measurement analysis to confirm construct dimensionality for the RSES-P. Involving convergent and divergent validity with a measure of affect, the Positive and Negative Affect Schedule (PANAS; Wtason, Clark, & Telleguen, 1988) will also be explored with the Portuguese adolescent sample, as an association between high positive affect (PA) and low negative affect (NA) and self-esteem is expected to occur (Brown & Marshall, 2001).

2. Study

2.1. Methods

2.1.1. Participants

Only the students who responded to all the items in the scales were included for analysis. The sample comprised 508 students attending the tenth, eleventh, and twelfth grades at several public secondary schools in Portugal's northern, central, and southern regions. Of the 508 students, 234 (46.1%) participants were male, and 274 (53.9%) were female. The volunteers' ages varied from 14 to 20 years, which corresponded to an average of 16.23 (standard deviation – SD = 1.18) years old;

five participants did not report their age. Data related to ethnicity were not collected, as this information is not routinely gathered in studies conducted in Portugal. Nevertheless, most respondents were Caucasian Europeans.

2.1.2. Instruments

A questionnaire was developed for the present study, the front page of which was devoted to demographic data (gender, age, school year, school success/failure, and the parents' educational levels) and certain measurements of vocational behavior (which were not used in the present study). The questionnaire further included measurements of the participants' self-esteem level and emotional (affective) response.

The self-esteem level was assessed using the RSES (Rosenberg, 1965), which comprises 10 items corresponding to statements on self-respect and self-acceptance and in which the participants are requested to express their agreement with such statements on a four-point Likert scale varying from 1 = strongly disagree to 4 = strongly agree. Half of the items (#2, 5, 6, 8, and 9) are formulated as negative statements, and thus the corresponding scores should be reversed before calculating the total scale score. The total score is calculated by adding the individual scores attributed to all 10 items, and the higher the total score, the higher is the self-esteem level.

The affect types were assessed using the Portuguese version of the PANAS (Watson, Clark & Telleguen, 1988), which was translated from the original instrument by the first author. The PANAS includes two 10-item subscales that provide a brief measure of PA and NA. The items consist of adjectives representing mood states related with the PA and NA, and respondents are requested to indicate the extent to which they feel each of the listed 20 emotions within a specific time frame on a five-point Likert scale with the following alternatives: 1 = very slightly or not at all, 2 = a little, 3 = moderately, 4 = quite a bit, and 5 = extremely. The time frame may be set according to the desired level of introspection. In the present study, the participants were requested to report on their feelings "during the past few weeks". For descriptive purposes, the scores of the PA and NA subscales were calculated by totaling the scores attributed to the 10 corresponding items by each participant. The reliability of the PANAS, based on the participant's separation index that corresponds in the IRT to Cronbach's alpha from the CTT (Linacre, 2011), was established at .84 and .83 for the PA and NA, respectively, for the present sample.

2.1.3. Procedure

The present study is a part of a larger research project that includes a broader battery of instruments to assess the effects of math self-efficacy and anxiety on career decision-making by secondary school students. The present study analyzed several psychological measures, including self-esteem. The selection of the RSES to assess self-esteem was based on its popularity in the specialized literature, which is largely due to the high psychometric quality of its results. Although there are several Portuguese versions available, as we stated above, doubts regarding their fundamental psychometric properties remain. For this reason, and before discussing the variables that exhibit correlations with the RSES, some of these doubts should be elucidated, including the uncertainty related to the unidimensionality of the scale of the measures' (i.e., the items' and participants') invariance properties. For the above-stated reasons, we decided to formulate a novel version of the original RSES instead of using the one in European Portuguese elaborated by Santos and Maia (2003).

The instrument was translated by the first author from the original RSES, and the back-translation was independently performed by two bilingual colleagues who hold PhDs in psychology to comply with the psychometric standards formulated for developing and cross-culturally adapting transliterated equivalents of psychological scales (Hambleton, Merenda & Spielberger, 2005). A native English speaker compared the original scale and the corresponding back-translation to elucidate and reformulate the meaning of certain words.

The study was approved by the Ministry of Education of Portugal. The headmasters of the selected secondary schools were contacted for authorization to conduct the study at their respective schools. Informed consent was obtained from the students and their parents or guardians after a thorough explanation of the study aims. Participation was voluntary, anonymous, and confidential. The principal investigator provided his personal contact information to the school liaison for subsequent delivery of a summary of the study results to the individuals expressing interest in them; in these cases, only the confidentiality of the responses could be guaranteed. The questionnaires were applied in the classroom by the first author and in the presence of the teachers in charge of the respective classes. The students were offered no incentives to participate in the study. All the volunteers completed the questionnaire in the time allotted for its application. The students who refused participation were allowed to leave the classroom before distribution of the questionnaires.

2.1.4. Data analyses

The Rasch analyses were performed using the computer program Winsteps (Linacre, 2011). Specifically, given the invariant polytomous format of all the items on the scale, the parameter estimates related to the subjects, items, and response categories were calculated based on the Rating Scale Model (RSM, Wright & Masters, 1982). According to Linacre (2002), the RSM is an extension of the Rasch model for polytomous items and is provided by:

$$log\,[P_{nik} / P_{ni(k-1)}] = B_n - D_i - F_k$$

where P_{nik} is the probability that person *n* would respond in category *k* of item *i*, $P_{ni(k-1)}$ is the probability that person *n* would respond in category *k-1* of item *i*, B_n is the ability of person *n* in the evaluated trait, D_i is the difficulty of item *i*, and F_k is the difficulty of the step in category *k-1* relative to category *k* (i.e., step calibration). This step calibration (F_k) is a threshold of the classification scale defined as the location corresponding to the equiprobability of observing the adjacent categories *k-1* and *k*.

In psychometric terms, selection of the RSM is justified for transforming ordinal data relative to the subjects' responses on an interval scale (Wright & Mok, 2004) and for its ideal metric properties as a Rasch model, namely, sufficient statistics and specific and statistic objectivity for person and item fit. In practical terms, the RSM has the advantage of not requiring large samples for properly estimating parameters and of allowing the empirical determination of the quality of the response categories in the Likert scales (Bond & Fox, 2007).

The assessment of the RSES-P using the RSM focused on the features related to content validity, structural validity, and substantive validity formulated by Wolfe & Smith (2007), based on Messick's study (1995).

2.2. Results

Rasch analysis provides indicators that facilitate quantifying the model fit, estimate item and person parameters, and diagnose the functioning of the item response categories (Fox & Jones, 1998). Table 1 describes the item statistics relative to fit (*infit* and *outfit*), location (*Di*), and standard error (*SE*) that allow assessing the content validity of the RSES-P, in addition to coefficients that allow determining the scale's structural validity (Wolfe & Smith, 2007).

Table 1. RSES Item Psychometric Properties and Principal Component Analysis

	MNSQ					
Item	Infit	Outfit	D_i	*SE*	r_{pm}	SC
1	1.00	1.08	.52	.07	.64	.20
2	1.01	.98	.79	.07	.75	-.57
3	.71	.93	-.62	.08	.60	.61
4	.96	1.10	-.47	.08	.59	.66
5	1.01	1.01	-.17	.07	.69	-.12
6	1.16	1.14	.14	.07	.70	-.60
7	.95	1.21	-.70	.08	.59	.47
8	1.27	1.41	1.87	.07	.64	-.15
9	.94	.79	-1.02	.08	.70	-.33
10	.75	.75	-.34	.07	.73	.06
Mean	.98	1.04	.00	.07	---	---
SD	.16	.19	.82	.00	---	---

Note. Di = item location; Infit = information-weighted mean square statistic; MNSQ = mean square fit indices; Outfit = outlier-sensitive mean square statistic; rpm = point-measure correlation; SC = structure coefficients from the standardized residual PCA; SE = Standard error.

Analysis of Table 1 relative to the content validity shows that the average *infit* and *outfit* values are practically equal to the expected one (1.0), which denotes perfect item fit. The individual values of each of the 10 scale items lay within the interval [.5 – 1.5] that Wright and Linacre (1994) established as productive for the assessed measure and that indicates absence of redundant items and homogeneity among RSES-P items. That first indicator of the data unidimensionality tends to be corroborated by other estimators included in Rasch modeling to analyze the contribution of the items to the definition of a central construct for the scale's internal structure. Indeed, the values of the point-measure correlations (r_{pm}), which are similar to item-CTT total score correlations, varied from .59 to .79, thus indicating the absence of unmodeled noise or data dependence (Linacre, 2011). Those results indicate that each and every item contributes to the definition of a common construct (e.g., self-esteem). Furthermore, the SE value of the items varied from .07 to .08, which also denotes high item reliability. Indeed, the excellent value of the *item separation reliability* (.99) (Fisher, 2007) indicates that the RSES-P items were measured with high precision.

To investigate the scale's structural validity, principal component analysis (PCA) of the standardized residuals was performed after controlling for the Rasch dimension to establish the extent to which the scale items correspond to the defined construct (Smith, 2004). Due to the lack of consensus on the criteria indicative of a secondary dimension (Chou & Wang, 2010), the authors of the present study established *eigenvalues* lower than 2.0 (Linacre, 2011), and up to 10% of the variance was explained by the first component of the residuals as indicators of RSES-P unidimensionality. The last column of Table 1 lists the structural coefficients of the 10 RSES-P items and shows that the values of five of them are higher than the cutoff point (.40), whereas only items 3 and 4 slightly surpass the .60 threshold. In addition, the *eigenvalue* of the first component is 1.90, corresponding to 8.8% of the residual variance, which might be considered good according to Fischer (2007) and which suggests that the standardized residuals do not exhibit additional systematic information. As the variance explained by the measure (53.5%) allows classifying it as a strong measurement dimension (Reckase, 1979), these results might be interpreted as indicative of RSES-P unidimensionality, which is a basic assumption required by the Rasch model.

The substantive validity concerns the diagnosis of the empirical functioning of a scale's response categories to establish whether they comply with the author's expectations at the time he or she formulated the corresponding items (Wolfe & Smith, 2007). Within the RSM context, Linacre (2002) suggested a set of criteria to establish the efficiency of response categories: (a) regular distribution of the observation frequency across categories with at least 10 observations of each category; (b) monotonic advancement of the observed average measures (*Bn*) and of the step calibration (*Fk*) for all the response categories; and (c) *outfit* mean-squares (MNSQ)—which exhibit greater sensitivity to unexpected responses compared to *infit*—of the response categories below 2.0. Table 2 summarizes the statistics necessary to assess the extent in which the four response categories of RSES-P comply with these criteria. Analysis of these results allows concluding that the structure of the response scale meets the criteria formulated by Linacre (2012) because the observation frequency of each category is higher than 10 and because the *outfit* value is lower than 2.0. In addition, the observed average measures and step calibration advance monotonically along the four response categories.

Table 2. RSES Response Category Statistics

	Observed			MNSQ		
Category	Count	%	Average B_n	Infit	Outfit	F_k
1	299	6	-1.33	1.16	1.32	---
2	992	20	-.13	.95	1.06	-1.92
3	1914	38	1.24	.91	.94	-.06
4	1875	37	2.60	.99	.99	1.98

Note. B_n = person trait; F_k = step calibration; Infit = information-weighted mean square statistic; MNSQ = mean square fit indices; Outfit = outlier-sensitive mean square statistic.

The adequacy of the response scale as established by the proper functioning of its four categories is corroborated by the results of the person fit analysis. The mean and SD of the fit statistics were 1.03 and .61 (*infit*) and 1.04 and .73 (*outfit*), respectively. As the percentage of individuals with an *infit* and/or *outfit* over 1.5—which is the upper threshold of psychometric acceptability in Rasch modeling (Wright & Linacre, 1994)—is low (17.0%), one might infer that the model fit is adequate (Fisher, 2007). Indeed, the parameters were estimated with high precision in most individuals. This inference is confirmed by the value of the *person separation reliability* (.82), which is a reliability estimator similar to Cronbach's alpha that measures the proportion of person variance that is not explained by the measurement error. Therefore, the RSES-P enables discriminating the sample in two or three levels regarding the measured attribute (e.g., self-esteem).

Figure 1 depicts the person-item joint representation on an identical metric scale (i.e., logits). The RSES-P's 10 items are organized in decreasing order of difficulty, which here is established based on the amount of attribute measured in the person (i.e., self-esteem). Consequently, item 8 poses the greatest choice difficulty, while item 9 is easier to agree with. Individuals with higher levels of self-esteem exhibit greater odds of selecting category 3 = *agree* or 4 = *strongly agree* in item 8 compared with individuals with lower levels of the measured attribute.

Therefore, the estimates relative to person attribute and item difficulty should expectably exhibit substantial overlapping for the set of items to be considered as adequately representing the sample. When the difference between such estimates is less than one logit (Bond & Fox, 2007), the information contained in the items facilitates discriminating the individuals in a much more precise manner regarding the

construct that the instrument is intended to measure. In the present study, this difference was 1.39, which denotes a high level of self-esteem, which explains the fact that most participants are above the point of the variable where the items are located. As the average of the individuals in this attribute is high (i.e., low item difficulty), which corresponds to individuals with low self-esteem level (± 1 logit) who tend to coincide with the scale items, the latter do not allow for a highly precise measurement of participants with average and high self-esteem levels (see Figure 1).

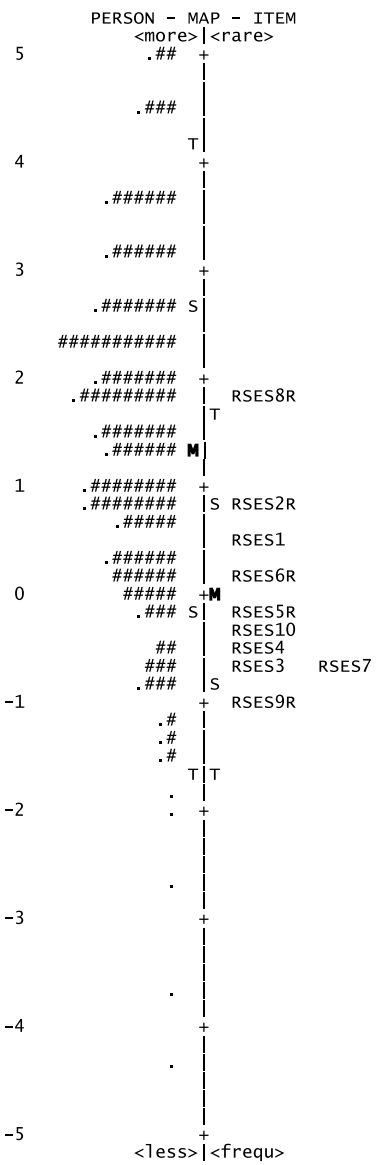

Figure 1. Joint Person and Item Representation along RSES Variable

In addition, the items were subjected to DIF analysis to assess the validity of the RSES-P results relative to gender. The standardized difference between the locations of the male and female parameters was calculated following adjustment for possible gender-related differences in the distribution of career decision-making self-efficacy. The Bonferroni procedure was used for this purpose, which corrected the selected significance level as a function of the number of comparisons (.05/10) (Linacre, 2011). Based on the application of that conservative criterion, none of the 10 RSES-P items exhibited locations above .50, which is the value established by Wright & Douglas (1975) as the cutoff point for DIF contrast, in the male or female genders (see Figure 2).

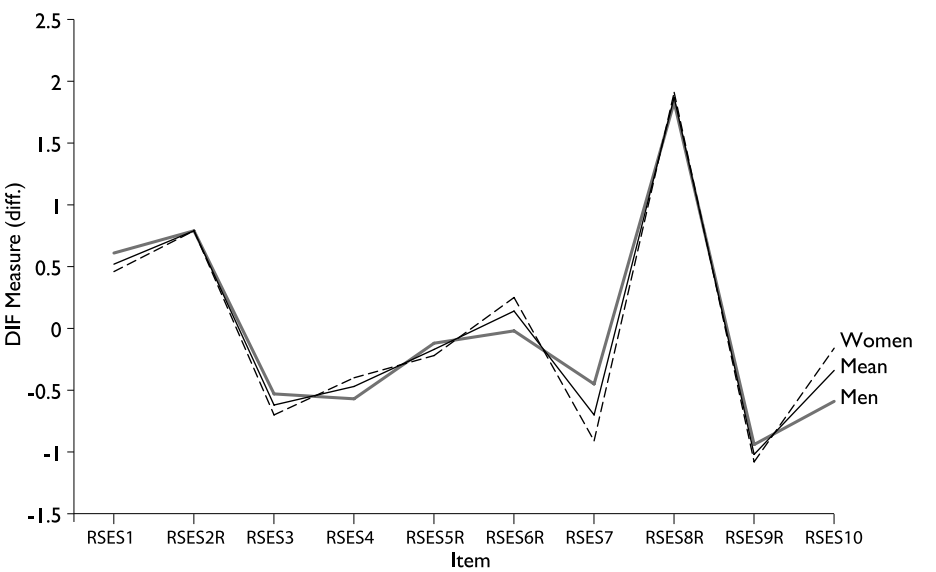

Figure 2. RSES differential item functioning (DIF) by gender

The average self-esteem score of the males (1.67) was significantly higher than in the females (1.16), although the effect size was low (d = .33) according to Cohen's (1988) conventional criteria. Those results agree with the findings of other studies that investigated self-esteem in adolescent samples using the RSES (Hendricks et al., 2001; Saigal, Lambert, Russ, & Hoult, 2002; Santos & Maia, 2003; Turner, Pickering, & Johnson, 1998).

To establish the convergent and discriminant validity of RSES-P in logits, its correlation with two affect dimensions (PA and NA), as measured by the PANAS (Watson et al., 1988) and in the logit scale (Miguel, Silva & Prieto, submitted), was

assessed. The results showed a strong positive correlation between self-esteem and PA (r = .57) and a moderate negative correlation with NA (r = -.27). All the correlations were statistically significant at a significance level of .001. These results provide further support to the already established association between high PA and low NA with self-esteem (Brown & Marshall, 2001) and thus agree with the theoretical expectations.

3. Discussion

Rasch analyses were performed to examine the psychometric properties of the RSES-P. Data were analyzed using the RSM (Wright & Masters, 1982), which (as a Rasch model) has ideal metric properties for ranking item difficulty and a person's ability (e.g., the level of the attribute measured) on a common scale. In addition to the joint measurement of people and items, and provided that the data are fitted to the model's requirements, Rasch modeling also allows for comparing people (items) regardless of the items (people) used in the measurement, which is property designated as *specific objectivity* by Rasch (Andrich, 1988).

Once the fit of the data to the Rasch model proved to be adequate, thus enabling a highly precise estimation of the item and person parameters, the RSES-P psychometric properties were analyzed relative to its dimensionality. The main purpose was to test the basic assumption of unidimensionality required by the Rasch model. The results of the present study show that the scale's 10 items measure a single latent dimension, to wit, self-esteem. Therefore, the application of Rasch analysis provided psychometric evidence supporting the hypothesis of the unidimensionality of the RSES-P.

The values of item and person reliability were high. The reliability of the RSES-P items is excellent (Fisher, 2007), thus allowing for adequate hierarchical ranking of the items' difficulty (Linacre, 2011), which results in a distribution that reflects a large interval of individual differences regarding the measured level of self-esteem. The average degree of difficulty of the items used to assess that person attribute was lower than the average level of self-esteem exhibited by the participants.

The person fit to the model was adequate, and the global person reliability was good (Fisher, 2007). This reliability indicator, which is similar to Cronbach's alpha in CTT, is in line with the results of studies assessing RSES reliability based

on the CTT (Pullmann & Allik, 2000; Rosenberg, 1965; Santos & Maia, 2003; Shapurian et al., 1987; Schmitt & Allik, 2005; Vasconcelos-Raposo, et al., 2012). However, although this reliability indicator allows discriminating the sample in a sufficient number of attribute levels (low, average, and high self-esteem), because the average self-esteem level of the participants was higher than the average difficulty of the items used to assess it, the latter do not enable highly precise measurements of individuals with average and high self-esteem levels. This limitation has previously been identified in IRT-based studies (Gray-Little et al., 1997; Quintão et al., 2011), according to which RSES can only measure with high precision the individuals located at the lowest construct level.

The results show that the four categories of the response scale used in RSES-P yield an adequate performance, thus psychometrically validating the CTT-based approach (Rosenberg; 1965) that is now confirmed by IRT in secondary school students. The Rasch model provided substantive validity for using this response scale and allowed for a thorough description of its functioning based on the precise manner by which the calibration of each of the four response categories was performed.

The DIF analysis that was based on gender and that applied a conservative criterion showed that the functioning of the RSES-P items was constant in both genders, being equally difficult for both subgroups. The items are sufficiently robust to allow for assessing self-esteem independently from the respondents' gender. Therefore, one might conclude that the responses merely quantify the level of the construct possessed by the assessed person and is exclusively measured as a function of the item difficulty, rather than by additional gender-related competences.

3.1. Conclusion

Overall, the present study confirms that the Portuguese version of the RSES is a valid instrument for assessing self-esteem in students attending secondary school in Portugal. The results support the hypothesis of unidimensionality posited by original author of the RSES (Rosenberg, 1965) and by other researchers (Marsh, 1996; Owens, 1994; Sheasby, Barlow, Cullen & Wright, 2000). Use of Rasch modeling reveals that the self-esteem level of individuals and the degree of difficulty of the RSES-P items measuring that construct are hierarchically ranked on the same scale continuum, with a high degree of precision in both cases. Thus, the adequacy of the

response scale with five alternatives is demonstrated, and so is the lack of gender-related bias.

The main implication of the present study, as its results show, is associated with the possibility of demonstrating that the RSES-P is an interval rather than ordinal measure. Indeed, the fit of the data to the Rasch model allows for generating a linear scale (logits) for items and people who yield values with high probabilities of being expressed in equal units. The demonstration that RSES-P items and the sample used to calibrate them are fully independent, thus allowing for direct comparison between the location of the person attribute in the latent variable (e.g., self-esteem) and the degree of difficulty of the item used to measure that attribute, confirms the presence of a simple structure, which is crucial for the performance of the invariant measures (Engelhard, 2008) that provide the basis for useful measurement models (Rasch, 1960). The interval nature of the RSES measure has several advantages that manifest in the valid use of parametric tests (which had not been previously ascertained) and that is justified by the measure invariance (people/items).

Regarding clinical practice relative to Portuguese adolescents, the results of the present study confirm the applicability of the RSES-P as a function of the respondents' level of self-esteem and the intended measurement error criterion, which is an approach that can help streamline the duration of intervention and minimize the costs of the process by fitting the items to the consultant's level of self-esteem (Vispoel, Boo, & Bleiler, 2001).

The present study exhibited limitations that should be overcome in future studies, particularly relative to the assessed sample, which although reproducing the demographic characteristics of the targeted population, cannot be considered representative of Portugal's reality. For the results to be generalizable, future studies must employ random sampling methods.

Finally, once the psychometric robustness of the RSES-P is confirmed by Rasch modeling, exploration of the network of connections between self-esteem and other psychosocial constructs (e.g., math self-efficacy, depression, or anxiety) using IRT-based approaches might prove useful. Moreover, further development of the present line of research through the performance of cross-cultural studies comparing the global assessment of self-esteem performed by adolescents from Portugal and the United States might also prove useful. Once the cross-cultural equivalence of the

various versions of this self-report instrument for self-esteem assessment is demonstrated by Rasch analysis, other researchers might use the results to map the assessed construct in other groups (e.g., adults) and application domains (e.g., workplaces and health environments). Further investigation of the self-esteem construct including, for example, its level of overlap with social desirability and the locus of control, is an interesting prospect.

References

Andrich, D. (1988). *Rasch models for measurement*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-068. Newbury Park, CA: SAGE.

Blascovich, J., & Tomaka, J. (1991). Measures of self-esteem. In J. P. Robinson, P.R. Sharver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 115-160). San Diego: Academic Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, J. D., Collins, R. L., & Schmidt, G. W. (1988). Self-esteem and direct versus indirect forms of self-enhancement. *Journal of Personality and Social Psychology*, *59*, 538-549.

Brown, J. D. & Marshall, M. A. (2001). Self-esteem and emotion: Some thoughts about feelings. *Personality and Social Psychology Bulletin*, *27*, 575-584.

Byrne, B. M. (1996). *Measuring self-concept across the lifespan: Issues and instrumentation*. Washington, DC: American Psychological Association.

Byrne, B. M., & Shavelson, R. J. (1996). On the structure of social self-concept for pre-, early and late adolescents: A test of the Shavelson, Hubner and Stanton model. *Journal of Personality and Social Psychology*, *70*, 599-613.

Cheng, S. T., & Hamid, P.N. (1995). An error in the use of translated scales: The Rosenberg Self-esteem Scale for Chinese. *Perceptual and Motor Skills*, *81*, 431-434.

Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residual. *Educational and Psychological Measurement*, *70*, 717-731.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Corwyn, R. F. (2000). The factor structure of global self-esteema mong adolescents and adults. *Journal of Research in Psychology*, *34*, 357-379.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*, 155-189.

Fisher, W. P. Jr. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*, p. 1095.

Franck, E., De Raedt, R., Barbez, C., & Rosseel, Y. (2008). Psychometric properties of the Dutch Rosenberg Self-esteem scale. *Psychologica Belgica*, *48*, 25-35.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology*, *45*, 30-45.

Gray-Little, B., William, V. S. L., & Hancock, T. D. (1997). An Item Response Theory analysis of the Rosenberg Self-esteem scale. *Personality and Social Psychology Bulletin*, *23*, 443-451.

Greenberger, E., Chen, C., Dimitrieva, J., & Farragia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-esteem Scale: Do they matter? *Personality and Individual Differences*, *35*, 1241-1254.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hendricks, C. S., Tavakoli, A., Hendricks, D. L., Harter, N. R., Campbell, K. P., L'Ecuver, R. I., Geddings, A. A. Hackett, D., & Mathis, D. (2001). Self-esteem matters: Racial and gender differences among rural southern adolescents. *Journal of National Black Nurses' Association*, *12*, 15-22.

Kamakura, T., Ando, J., & Ono, Y. (2001). Genetic and environmental influences on self-esteem in a Japanese twin sample. *Twin Research*, *4*, 439-442.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85-106.

Linacre, J. M. (2011). *Winsteps Rasch measurement computer program, version 3.73.0*. [Computer program.] Chicago, Il: Winsteps.com.

Marsh, H. W. (1993). Relations between global and specific domains of the self: The importance of individual importance, certainty and ideals. *Journal of Personality and Social Psychology*, *65*, 975-992.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantial meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, *70*, 810-819.

Martín-Albo, J., Nuñez, J. L., Navarro, J. G., & Grijalvo, F. (2007). The Rosenberg Self-esteem Scale: Translation and validation in university students. *The Spanish Journal of Psychology*, *10*, 458-467.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.

Miguel, J. P., Silva, J. T., & Prieto, G. (submitted). Positive and Negative Affect Schedule—Portuguese European version: A Rasch analysis. *International Journal of Testing*.

O'Brien, E. J. (1985). Global self-esteem scales: Unidimensional or multidimensional? *Psychological Reports*, *57*, 383-389.

Owens, T. J. (1994). Two dimensions of self-esteem: Reciprocal effects of positive self-worth and self-deprecation on adolescents problems. *American Sociological Review*, *59*, 391-407.

Prezza, M., Trombaccia, F. R., & Armento, L. (1997). La scala dell'autoestima di Rosenberg: traduzione e validazione italiana [The Rosenberg self-efficacy scale: italian translation and validation]. *Bolletino di Psicologia Applicata*, *223*, 35-44.

Prieto, G., & Delgado, A. R. (2007). Measuring math anxiety (in Spanish) with the Rasch rating scale model. *Journal of Applied Measurement*, *8*, 149-160.

Pullmann, H., & Allik, J. (2000). The Rosenberg Self-Esteem Scale: its dimensionality, stability and personality correlates in Estonian. *Personality and Individual Differences*, *28*, 701-715.

Quintão, S., Prieto, G., & Delgado, A. R. (2011). Avaliação da escala de auto-estima de Rosenberg mediante o modelo de Rasch [Evaluation of Rosenberg self-esteem scale using Rasch model]. *Psicologia*, *25*, 87-101.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research. [Expanded edition, 1980. Chicago: University of Chicago Press.]

Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics*, *4*, 207-230.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rosenberg, M. (1979). *Conceiving the self*. New York: Basic Books.

Rosenberg, M., Schooler, C., Schoenbach, C., & Rosenberg, F. (1995). Global self-esteem and specific self-esteem: Different concepts, different outcomes. *American Sociological Review*, *60*, 141-156.

Rost, J. (2000). The growing family of Rasch models. In A. Boomsma, A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on Item Response Theory* (pp. 25-42). New York: Springer-Verlag.

Roth, M., Decker, O., Herzberg, P. Y., & Brähler, E. (2008). Dimensionality and norms of the Rosenber Self-esteem Scale in a German population sample. *European Journal of Psychological Assessment*, *24*, 190-197.

Saigal, S., Lambert, M., Russ, C., & Hoult, L. (2002). Self-esteem of adolescents who were born prematurely. *Pediatrics*, *109*, 253-268.

Santos, P. J., & Maia, J. (2003). Análise factorial confirmatória e validação preliminar de uma versão portuguesa da escala de auto-estima de Rosenberg [Confirmatory factor analysis and preliminary validation of a Portuguese version of the Rosenberg Self-esteem scale]. *Psicologia: Teoria, Investigação e Prática, 8*, 253-268.

Shavelson, J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, *46*, 407-442.

Shevlin, M. E., Bunting, B. P., & Lewis, C. A. (1995). Confirmatory factor analysis of the Rosenberg Self-esteem Scale. *Psychological Reports*, *76*, 707-710.

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, *89*, 623-642.

Shapurian, R., Hojat, M., & Nayerahmadi, H. (1987). Psychometric characteristics and dimensionality of Persian version of Rosenberg Self-esteem Scale. *Perceptual and Motor Skills*, 65, 27-34.

Sheasby, J. E., Barlow, J. H., Cullen, L. A., & Wright, C. C. (2000). Psychometric properties of the Rosenberg Self-esteem Scale among people with arthritis. *Psychological Reports*, *86*, 1139-1141.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120.

Sinclair, S. J., Blais, M. A., Gansler, D. A., Sandberg, E., Bistis, K., & LoCicero, A. (2010). Psychometric properties of the Rosenberg-Self-esteem Scale: Overall and across demographic groups living within the United States. *Evaluation & the Health Professions*, *33*, 56-80.

Šmídová, J., Hátlová, B., & Stochl, J. (2008). Global self-esteem in a sample of Czech seniors and adolescents. *Acta Universitatis Palackianae Olomucensis*, *Gymnia*, *38*, 31-37.

Smith Jr, E. V. (2004). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In: E. V. Smith Jr & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 575-600). Maple Grove, MN: JAM Press.

Tabachnick, B., & Fidell, L. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon / Pearson.

Tomás, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factor sor method effects? *Structural Equation Modeling*, *6*, 84-98.

Turner, L. A., Pickering, S., & Johnson, R. B. (1998). The relationship of attributional beliefs to self-esteem. *Adolescence*, *33*, 477-484.

Vallieres, E. F., & Vallerand, R. J. (1990). Traduction et validation Canadienne-Française de l'échelle de l'estime de soi de Rosenberg [French-Canadian translation and validation of the Rosenber's Self-esteem Scale]. *International Journal of Psychology*, *25*, 305-316.

Vasconcelos-Raposo, J., Fernandes, H. M., Teixeira, C. M., & Bertelli, R. (2012). Factorial validity and invariance of the Rosenberg Self-Esteem Scale among Portuguese youngsters. *Social Indicators Research*, *105*, 483-498.

Vispoel, W. P., Boo, J., & Bleiler, T. (2001). Computerized and paper-and-pencil versions of the Rosenberg self-esteem scale: A comparison of psychometric features and response preferences. *Educational and Psychological Measurement*, *61*, 461-474.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063-1070.

Waugh, R. F., & Chapman, E. S. (2005). An analysis of dimensionality using factor analysis (true-score) and Rasch measurement: What is the difference? Which method is better? *Journal of Applied Measurement*, *6*, 80-99.

Wolfe, E. W., & Smith Jr, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II – Validation activities. In E. V. Smith Jr & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 243-290). Maple Grove, MN: JAM Press.

Wright, B. D., & Douglas, G. A. (1975). A better procedure for sample-free item analysis. *Research Memorandum*. Statistical Laboratory. Department of Education. University of Chicago.

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

Wright, B D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, p. 370.

Wright, B. D., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. In E. V. Smith Jr & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 1-24). Maple Grove, MN: JAM Press.

Conclusão

O modelo de Rasch é o modelo mais simples da teoria de resposta ao item (TRI) e constitui uma abordagem útil ao nível da construção e refinamento psicométrico de instrumentos de avaliação psicológica. Trata-se de um modelo logístico de um parâmetro da TRI no qual a quantidade de traço latente existente na pessoa e a quantidade do mesmo traço latente reflectido nos vários itens do instrumento podem ser estimados de forma independente e comparados directamente entre si, uma vez que sujeitos e itens foram medidas numa mesma métrica comum, a escala logit.

É adequado usar uma análise Rasch sempre que se pretende usar o resultado total de uma prova de avaliação psicológica para fazer inferências acerca do nível de traço latente existente no sujeito avaliado, uma vez que o modelo parte matematicamente da exigência de invariância entre pessoas e itens (Andrich, 2004). Embora a teoria clássica dos testes (TCT) também use o resultado total para caracterizar cada sujeito, preconiza-o como a estatística relevante, menorizando as anomalias verificadas nos itens ou nas respostas dos sujeitos aos itens. A utilização do modelo de Rasch, em contrapartida, concorre para explicar estas anomalias, proporcionando um resultado muito mais informativo. O propósito da mensuração Rasch é análogo à construção de uma régua (Engelhard,1992), numa tentativa de estabelecimento da medida correcta através do mapeamento do construto ao longo de um mesmo contínuo escalar (Andrich & Luo, 2003).

O modelo de Rasch usa a soma das respostas aos itens como estatística suficiente para estimar as probabilidades de resposta, baseadas na disponibilidade individual do sujeito para escolher um conjunto de itens e na dificuldade de escolha desses itens. A dificuldade em escolher o item é assumida como a principal caracte-

rística que influencia as respostas (Linacre, 2011). A análise Rasch apresenta estimativas da disponibilidade de escolha do item e da dificuldade de escolhê-lo ao longo de uma mesma escala de probabilidades logarítmicas (logit[1]) que é uma escala intervalar na qual a unidade dos intervalos entre as localizações na qual a unidade dos intervalos entre as localizações conjuntas pessoas-itens têm um valor ou um significado consistente (Bond & Fox, 2007). Os autores explicam que a utilização de técnicas Rasch permite a ordenação dos respondentes ao longo deste contínuo de disponibilidade de escolha dos itens e ordena os itens ao longo de um contínuo relativo à dificuldade de os escolher.

O modelo de Rasch produz uma imagem abrangente e informativa do construto medido, bem como dos sujeitos nos quais esse traço latente foi avaliado. O modelo permite que a o traço existente nos sujeitos e o nível de dificuldade dos itens usados nessa avaliação sejam interligados de modo a indicarem a ocorrência de uma determinada resposta como uma probabilidade e não como um determinismo. Por outro lado, o modelo preserva a ordem na medida em que a probabilidade de ser dada uma certa resposta define uma ordem para os sujeitos e para os itens (Wright & Masters, 1982). Dito de outra forma, o modelo de Rasch é um modelo hierárquico implicativo. É esperado que itens difíceis sejam escolhidos apenas por sujeitos com grande quantidade do traço medido e que pessoas com baixos valores nesse atributo apenas escolham itens fáceis. Os itens formam, pois, um estrutura hierárquica na qual as respostas positivas a itens difíceis implicam respostas positivas a itens fáceis, embora a inversa não seja verdadeira, isto é, respostas positivas a itens fáceis não implicam necessariamente respostas positivas a itens difíceis. A estrutura implicativa, subjacente ao modelo de Rasch, é direccional no sentido do difícil para o fácil, contrariamente à análise factorial que, por se tratar de um modelo correlacional, quando os itens saturam num factor também terão que correlacional com todos os itens que designam essa dimensão, ou seja, a escolha de um item implica que o sujeito tem probabilidade de escolher todos os itens que integram essa escala. A classificação hierárquica dos itens pode ser usada como um teste empírico à validade de construto. Se os itens medem um único traço latente, as estatísticas de ajuste deverão indicar que os itens ajustam ao modelo e a ordenação dos itens deverá fazer qualitativamente senti-

[1] O acrónimo resulta da expressão anglo-saxónica *log odds unit*.

do em função do conhecimento que se tem acerca do construto que está a ser medido (Prieto& Delgado, 2003).

A conceptualização original de Rasch para itens dicotómicos (Rasch, 1960) foi estendida por Andrich (1978) a itens politómicos, quando este autor propôs que as respostas numa escala tipo de Likert podem ser ordenadas e usadas de um modo semelhante aos das respostas dicotómicas, de modo a permitir inferir a quantidade de atitude ou de atributo psicológico existente num sujeito que responde a um instrumento de avaliação. Esta extensão do modelo a escalas de classificação através do Rasch Rating Scale Model (RSM; Wright & Masters, 1982) reconceptualiza a dificuldade do item como a resistência à escolha de uma categoria da escala de resposta. O mapeamento que o modelo faz da variável ajuda a visualizar o construto e o modo como os itens do instrumento o definem, constituindo por isso uma ferramenta útil para a validação e progressiva compreensão do construto avaliado.

Tendo em consideração os propósitos norteadores da investigação que se sintetiza na presente Dissertação, apresentar as potencialidades do modelo de Rasch enquanto abordagem útil para o refinamento psicométrico de instrumentos de avaliação psicológica e subsequentemente contribuir para o garante da ética e deontologia da prática profissional neles radicada (Simões, Almeida & Gonçalves, 1999), importa salientar, dos resultados obtidos, aqueles que se revestem de maior importância para os objectivos estipulados. No âmbito dos três estudos realizados com o Rasch RSM, os resultados alcançados permitiram reunir evidência psicométrica que corrobora, para as versões Portuguesas dos instrumentos de avaliação utilizados (e.g., CDSE-SF, PANAS e RSES), a respectiva: (1) validade de conteúdo (i.e., o bom ajustamento dos dados ao modelo permitiu que a parametrização dos sujeitos e a calibração dos itens tenha sido feita com elevada precisão na); (2) validade estrutural (i.e., unidimensionalidade e ausência de DIF), e (3) validade substantiva (i.e., as categorias da escala de resposta funcionam de forma adequada).

Globalmente, os estudos confirmam que as versões da CDSE-SF, da PANAS e da RSES constituem instrumentos válidos para a avaliação dos respectivos construtos junto de amostras da população de alunos do ensino secundário Portugueses. O Rasch RSM permite que os níveis dos atributos medidos nas pessoas e o nível de dificuldade associado à escolha dos itens que medem esses construtos sejam hierarquizados num mesmo contínuo de traço latente, fazendo-o com elevados graus de pre-

cisão, o que por sua vez prova a adequação das respectivas escalas de resposta, e sem enviesamento quanto ao género de pertença. Da sua aplicação resulta que, no caso específico das versões Portuguesas da CDSE-SF, da PANAS e da RSES, os seus respectivos itens são representativos e relevantes para os domínios dos construtos avaliados (e.g., validade de conteúdo), têm correspondência com os construtos definidos em cada um dos instrumentos (e.g., validade estrutural) e são avaliados através de escalas de resposta cujo diagnóstico ao funcionamento empírico das respectivas categorias revelou adequação das mesmas (e.g., validade substantiva).

A principal implicação dos estudos realizados com o Rasch RSM resulta, tal como os resultados comprovam, na possibilidade de demonstrar que os três instrumentos de avaliação utilizados configuram medidas em intervalos, não ordinais. De facto, o ajustamento dos dados ao modelo de Rasch permitiu a criação de escalas lineares de natureza intervalar, em logits, para itens e pessoas que proporcionam valores com elevada probabilidade de serem expressos nas mesmas unidades. Ao demonstrar-se que os itens de cada um dos instrumentos de avaliação e as amostras usadas nas respectivas calibrações são totalmente independentes, permitindo comparar a localização do atributo do sujeito na variável latente (e.g., auto-eficácia de decisão de carreira, afecto positivo, afecto negativo, auto-estima) directamente com o nível de dificuldade do item usado para medir esse atributo, está-se também a confirmar a existência de uma estrutura simples que é essencial na realização de medidas invariantes (Engelhard, 2008) que formam a base para modelos de medida úteis (Rasch, 1960). A natureza intervalar da medida traduz-se em vantagens que se concretizam na utilização válida de testes paramétricos, o que até aqui não era possível afirmar, justificada pela invariância da medida (pessoas/itens).

Embora haja autores que, no processo de validação de um instrumento, utilizam a análise factorial como estratégia para a identificação das subescalas existente no espaço latente dos dados, seguida de análises Rasch para avaliarem a qualidade dessas subescalas, optou-se por não replicar este procedimento. A razão prende-se com o facto das diferenças relativas aos requisitos de cada uma destas abordagens poderem causar problemas quando usados conjuntamente (Wright, 1996). A análise factorial tende a favorecer itens que caem dentro de um intervalo estreito de dificuldade, bem como itens que são redundantes ou que carecem de independência. A pré selecção de itens para uma análise Rasch a partir dos resultados da análise factorial,

ou em alternativa, aplicar análise de Rasch a um instrumento que foi originariamente desenvolvido com recurso à análise factorial, pode resultar em escalas cujos itens revelem sobre ajuste e intervalos reduzidos de dificuldade.

Pelas razões acabadas de expor optou-se, no caso da PANAS, por realizar dois estudos independentes. Os resultados do estudo baseado na abordagem da TCT com recurso a técnicas de análise factorial exploratória e confirmatória, revelaram dificuldade de ajustamento do modelo aos dados que só parcialmente conseguiram replicar a estrutura factorial original do instrumento. De facto, grandes diferenças ao nível da dificuldade de escolha dos itens configurou-se problemático para a análise factorial, uma vez que itens de escolha difícil revelaram correlações moderadas com itens fáceis de escolher, mesmo quando ambos são indicativos do mesmo traço. Em alguns casos, itens fáceis e itens difíceis não saturaram conjuntamente na mesma dimensão, resultado que conflitua com a calibração precisa que o Rasch RSM havia alcançado. Se quando aplicado a dados binários, o modelo factorial (TCT) constitui uma aproximação linear do modelo de Rasch (TRI), por vezes tida como satisfatória (McDonald, 1999), com dados politómicos tal não se verifica dada a natureza logística, não linear, do modelo.

Tendo em consideração a natureza exploratória dos estudos realizados com a CDSE-SF, com a PANAS e com a RSES, a principal limitação decorre do facto das amostras não serem representativas da realidade Portuguesa. Futuros estudos, visando a produção de normas para estes instrumentos de avaliação psicológica, irão reunir grupos amostrais representativos da realidade Portuguesa em função de variáveis sociodemográficas tão importantes como a idade e a situação ocupacional, para além do género de pertença.

References

Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrica*, *43*, 561-573.

Andrich, D. A. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E. V. Smith Jr. and R. M. Smith (Eds.), Introduction to Rasch measurement (pp. 143-166). Maple Grove, MN: JAM Press.

Andrich, D. A., & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement*, *4*, 205-221.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement*, *6*, 155-189.

Engelhard, G., Jr. (1992). Historical views of invariance: Evidence from the Measurement Theories of Thorndike, Thurstone, and Rasch. *Educational and Psychological Measurement*, *52*, 275-291.

Linacre, J. M. (2011). *Winsteps Rasch measurement computer program, version 3.73.0* [*Manual*]. Chicago, Il: Winsteps.com.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Prieto, G. & Delgado, A. R. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema*, *15*, 94-100.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Simões, M. R., Almeida, L. S., & Gonçalves, M. M. (1999). Testes e provas psicológicas em Portugal: Roteiro de alguma questões que atravessam a utilização de instrumentos de/na avaliação psicológica. In M. R. Simões, M. M. Gonçalves, & L. S. Almeida (Eds.), *Testes e provas psicológicas em Portugal* (Vol. 2, pp. 1-12). Braga: APPORT/SHO.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, 3, 3-24.

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

CURRICULUM VITÆ

José Manuel Pacheco Miguel

José Manuel Pacheco Miguel

Coimbra, Dezembro de 2013

· U C ·