

Precisão de avaliadores na avaliação da criatividade por meio da produção de metáforas¹

Ricardo Primi – Universidade São Francisco²
 Fabiano Koich Miguel – Universidade São Francisco
 Gleiber Couto – Universidade São Francisco
 Monalisa Muniz – Universidade São Francisco

Resumo

A criatividade é um dos temas da psicologia cuja mensuração está circundada de dificuldades, já que é uma área carente de bons instrumentos. Metáforas como “O camelo é o barco do deserto” são meios de expressão de aspectos diferente de algum conceito (camelo) por meio da associação de idéias. A elaboração de metáforas desse tipo pressupõe a execução de processos cognitivos básicos, tais como o raciocínio analógico e as associações remotas, processos esses que estão associados aos componentes cognitivos da criatividade. Objetivou-se nesta pesquisa estudar a precisão de critérios de pontuação de um instrumento de avaliação da criatividade por meio de produção de metáforas que avalia respostas metafóricas a partir de um estímulo do tipo “O camelo é o _____ do deserto”. Participaram deste estudo 19 sujeitos e nove juízes. O teste de metáforas é composto por nove itens, aos quais os sujeitos deram um total de 513 respostas. Cada resposta foi avaliada independentemente pelos juízes, em uma escala de 0 a 3, indicando o nível de elaboração da metáfora. A precisão foi calculada a partir do modelo de Rasch, assumindo cada idéia como um caso e cada juiz como um item de um teste hipotético, aplicando-se um procedimento chamado rede de juízes ancorados. A precisão de avaliadores variou de 0,52 a 0,83, com média 0,74 (DP=0,08), indicando uma boa precisão de avaliadores.

Palavras-chave: Teste de criatividade; Modelo de Rasch; Teoria de resposta ao item.

Inter rater reliability in the creativity assessment using metaphor production

Abstract

The measure of creativity is a difficult task in psychology and consequently there is a shortage of good quality instruments. Metaphors, like “The camel is the boat of the desert”, are means of expressing different characteristics of a concept (camel) through association of ideas. The creation of such metaphors assumes the accomplishment of basic cognitive processes, like analogical reasoning and remote associations, which are linked to the creativity cognitive processes. The goal of this research was to study the reliability of scoring system for a creativity test based on metaphors creation, making use of items like “The camel is the _____ of the desert”. The participants were 19 people and nine raters. The metaphor test is made of nine items, to which the participants gave 513 answers. Each answer was independently measured by the raters using a scale from 0 to 3, indicating the metaphor’s elaboration. Reliability was calculated by a Rasch model assuming every idea as a case and each judge as an item of a hypothetical test in procedure called judge-linking network. The inter-rater reliability varied from .52 to .83 with a mean of .74 (SD=.08) resulting in a acceptable inter-rater reliability.

Keywords: Creativity test; Rasch model; Item response theory.

Introdução

A explicação psicológica das grandes realizações humanas em várias áreas, como, por exemplo, grandes inventos, descobertas científicas, matemática, obras artísticas e literárias, é buscada, em parte, nos estudos e teorias sobre os níveis mais altos do desenvolvimento da inteligência. Ackerman (1996) e Ackerman e Heggstad (1997) citam duas tradições de teorização e medida da inteligência: uma que busca medir processos cognitivos de maneira mais pura e outra mais centrada na organização

dos conteúdos ou conhecimentos nos quais operam os processos mentais. Essa distinção coincide similarmente com os construtos inteligência fluida (processos mentais de raciocínio) e cristalizada (extensão e profundidade dos conhecimentos adquiridos) (Cattell, 1943). Ackerman e Beier (2005) salientam que, tradicionalmente, a abordagem processual tem sido mais influente no desenvolvimento dos testes de inteligência, mas a avaliação da inteligência do adulto deve incorporar a investigação das estruturas e organização de conhecimentos mais especializados sobre o que as pessoas sabem e conseguem realizar.

¹ As atividades de pesquisa do primeiro autor que deram origem a este artigo são financiadas pelo CNPq e pela FAPESP.

² Endereço para correspondência:

Universidade São Francisco – Laboratório de Avaliação Psicológica e Educacional (LabAPE) – Mestrado em Psicologia
 Rua Alexandre Rodrigues Barbosa, 45 – 13251-900 – Itatiba-SP – Telefone (11) 45348118
 E-mail: ricardo.primi@saofrancisco.edu.br – rprimi@uol.com.br.

Tal argumentação faz sentido, já que as realizações profissionais acadêmicas de mais alta complexidade dependem, além, evidentemente, da capacidade de raciocínio (inteligência fluida), de um conjunto extenso e profundo de conhecimentos organizados, necessários à solução de problemas requeridos nas áreas de atuação profissional altamente especializadas. Também faz sentido que a consolidação desses conhecimentos dependa, entre outras coisas, de capacidades fluidas que entram em ação quando novas informações são adquiridas, uma vez que atuam na sua organização, facilitando a posterior recuperação, tese central da teoria do investimento de Cattell (1943). Duncan, Burgess e Emslie (1995) também enfatizam essa idéia quando afirmam que os testes de inteligência cristalizada podem refletir o fator *g* no momento em que os conhecimentos foram aprendidos. Assim, é certo que realizações de alto nível dependem de aplicação de conhecimentos especializados e, ao mesmo tempo, de capacidades fluidas de organização de novas informações.

Uma dimensão não tão estudada quanto medidas de inteligência fluida e cristalizada, mas que também tem papel fundamental nas realizações humanas complexas, é a criatividade. Embora ainda não haja consenso sobre a natureza da relação entre inteligência e criatividade, vários modelos a incluem como dimensões da inteligência ou relacionada a esse construto. Há poucos modelos que a definem como um construto separado da inteligência. Como afirmam Sternberg e O'Hara (2000),

a inteligência tem sido concebida como crucial na adaptação a ambientes existentes, enquanto que a criatividade, a qual envolve a produção de idéias ou produtos novos e úteis, tem sido vista como crucial na modificação e modelagem do ambiente. (p. 611)

Sternberg e Lubart (1991) propuseram um modelo no qual seis elementos convergem para formar a criatividade: inteligência, conhecimento, estilos cognitivos, personalidade, motivação e ambiente.

Assim, no nível cognitivo, a criatividade envolve processos específicos associados à produção de idéias novas a partir de uma base de conhecimentos já adquirida (Guilford & Christensen, 1973). Envolve, portanto, a reorganização de conhecimentos adquiridos em novas recombinações incomuns e originais. Além disso, um dos critérios também considerados é que a nova idéia solucione um problema reconhecido em uma comunidade (Mackinnon, 1962). Nesse sentido, uma compreensão mais abrangente das realizações humanas em seu nível mais alto de expressão deve envolver aspectos cognitivos fluidos, conhecimento especializado e também a criatividade.

Dada a escassez de instrumentos e estudos sobre a avaliação da criatividade, um projeto foi iniciado no

Laboratório de Avaliação Psicológica e Educacional (LabAPE) da Universidade São Francisco para o desenvolvimento de um novo instrumento para avaliação de componentes cognitivos da criatividade. Tal projeto originou o Teste de Criação de Metáforas – TCM (Primi e cols., 2006), objeto deste artigo. Esse instrumento se fundamenta nas idéias sobre o raciocínio analógico e metafórico de Sternberg e Nigro (1983) e Tourangeau e Sternberg (1981, 1982). O pensamento metafórico é um tipo particular de raciocínio, que utiliza idéias conhecidas para criar sentidos novos para outras idéias. Alguns exemplos de metáfora são: “O camelo é o barco do deserto”, “O cabide é a espinha da roupa”, “O bigode é a antena do gato”, “O cavalo é o pasto do carrapato” (os dois primeiros exemplos foram retirados de protocolos do TCM e os outros dois, da música *Cultura*, de Arnaldo Antunes).

As metáforas podem ser decompostas em três elementos estruturais: o teor, o veículo e a base representacional. Por exemplo, as palavras camelo/cabide/bigode/cavalo são teores que receberão um novo sentido por intermédio dos veículos barco/espinha/antena/pasto. A base representacional constitui-se nos domínios semânticos de atributos e relações, particulares de cada teor e veículo, no qual eles estão inseridos. Nos quatro exemplos, algumas relações entre os teores em sua base representacional ficam explícitas (camelo do deserto, cabide da roupa, bigode do gato, cavalo com carrapato). Já em relação aos veículos, essa base representacional está implícita e precisa ser inferida para que seja possível entender as metáforas. Por exemplo, podemos estabelecer as seguintes relações paralelas: barco/mar, indicando que, analogamente, o camelo é um meio de transporte no deserto; espinha/peixe, sugerindo que o cabide dá sustentação para a roupa; antena/casa, significando que o bigode do gato é um meio de receber sinais do ambiente; e pasto/vaca, apontando que o cavalo serve de local e alimento para o carrapato.

Sternberg e Nigro (1983) argumentam que a concepção de uma idéia nova em razão de idéias antigas forma a base tanto do pensamento metafórico quanto do pensamento analógico. Propõem que as metáforas sejam vistas como baseadas em uma analogia subjacente, pela qual alguns termos ficam implícitos ou subentendidos. Na Tabela 1, os exemplos citados são apresentados de forma esquemática demonstrando que esses casos deixam implícito o termo “D” das analogias. Assim, os processos cognitivos do pensamento metafórico seriam semelhantes aos do processamento das analogias (Primi, 1998; Sternberg, 1977, 1983), isto é, envolveriam a codificação dos termos (A, B, etc.), a inferência de relações entre termos A:B e o mapeamento de relações mais abstratas (similaridade e diferença das relações inferidas) entre o par A:B com o par C:D. A diferença, no caso das metáforas, é a necessidade da descoberta dos termos e domínios subjacentes e o peso

maior do componente de mapeamento para perceber as relações paralelas entre os domínios (Sternberg, 1983).

Tabela 1 – Representação estrutural comparativa das analogias e metáforas

Analogias	Metáforas
Estrutura	
A está para B assim como C está para D	A é o C de B
A:B // C:D	A (// C:?) : B
Exemplos	
Camelo:Deserto // Barco:Mar	O camelo é o barco do deserto
Cabide:Roupa // Espinha:Peixe	O cabide é a espinha da roupa
Bigode:Gato // Antena:Casa	O bigode é a antena do gato
Cavalo:Carrapato // Pasto:Vaca	O cavalo é o pasto do carrapato

Tourangeau e Sternberg (1982) propõem um modelo de interação de domínios cuja idéia central para explicação das metáforas é a de que os veículos – relacionados ao seu espaço semântico particular de propriedades e relações específicas àquele domínio – são os modelos de base para ver o teor de uma nova maneira. A interação ocorre já que não somente as idéias (teor e veículo), mas também seus domínios, são vistos de maneira diferente em razão da associação feita pelas metáforas. Os autores citam como exemplo a metáfora “Os homens são lobos” para ilustrar esse ponto. As metáforas envolvem ver uma idéia (homem) que é inserida em seu espaço semântico como outra idéia (lobos), de um segundo espaço semântico, resultando não só na visão diferente da primeira, mas também na mudança das visões de ambos os domínios. Os lobos podem ser caracterizados como ferozes, predadores, carniceiros e esfomeados. Essas características pertencem ao domínio de representações das características dos animais, mas precisam ser transformadas para que haja o entendimento do que se quer dizer sobre os homens. Essa metáfora procura salientar o aspecto competitivo “feroz” das relações sociais entre homes. Assim, ao se aplicarem as características do domínio do veículo (lobo) ao universo das relações sociais humanas (teor), as características do domínio animal (veículo) são transformadas para dar sentido à metáfora e salientar algo que não era tão óbvio antes.

Um estudo de Tourangeau e Sternberg (1981) operacionalizou essa idéia em dois critérios mais objetivos, estudando a compreensibilidade e adequação das metáforas, chamados de equivalência, ligados às relações das idéias dentro de seus domínios, e remotividade, tratando das relações entre os domínios. A equivalência diz respeito ao grau em que a relação C:D é análoga à relação A:B, isto é, se as relações que a idéia C tem com o termo D (implícito nas metáforas) são equivalentes às relações A:B. No exemplo, barco (veículo) e camelo (teor) têm uma relação de equivalência com o elemento de seu universo semântico (mar e deserto), que é o fato de serem meios de

transporte. Em suma, a equivalência retrata a qualidade ou o paralelismo das associações de idéias explícitas (nesse caso: deserto) e implícitas (nesse caso: mar) na metáfora. Indica que as idéias da metáfora facilmente compõem um mosaico de associações analógicas. Portanto, a equivalência é um indicador da qualidade das associações entre os termos dentro do domínio de cada um dos termos da metáfora (teor e veículo).

A remotividade diz respeito à distância dos universos semânticos A:B//C:D. Quanto mais distantes ou remotos, mais interessante parecerá a metáfora. Mas, se for extremamente distante, pode-se perder o sentido da nova idéia e a metáfora passa a não ser tão boa. A remotividade está associada à originalidade, pois a distância dos campos semânticos associa-se também ao incomum, pouco freqüente, e à novidade. É evidente que o novo deve ser compreensível, por isso há uma relação conjunta entre novidade e compreensibilidade. Por intermédio desses critérios, Sternberg e Tourangeau (1981) mostraram que as metáforas são julgadas como mais compreensíveis e adequadas quanto mais haja um paralelismo das relações intradomínio dos elementos teor e veículo e quanto mais distantes sejam os domínios, desde que não passem de um ponto considerado ótimo.

Há outros aspectos também importantes ligados à qualidade das metáforas. Um deles está associado ao reconhecimento consensual da qualidade de uma idéia. Esse critério pode envolver a avaliação e concordância quanto ao nível, qualidade e compreensibilidade de vários julgamentos independentes. Outro ponto ligado à compreensão é o fato de que as características do veículo usadas para redefinir o teor são normalmente características salientes (conhecidas ou freqüentes) em seu domínio, facilitando assim a compreensão da metáfora, pois podem ser facilmente interpretadas. O que é novo e infreqüente nas metáforas é resultante da associação de domínios distantes, trazendo à tona uma característica incomum do teor por meio dos processos associativos implícitos.

Em suma, nas metáforas há uma interação dinâmica de associações de equivalência dentro dos dois sistemas conceituais ligados ao teor e ao veículo, e de paralelismo e distância ótima entre domínios de forma a criar sentidos diferentes para o teor, quando visto por intermédio do veículo. Esse modelo fundamentou a criação do instrumento que é objeto deste estudo. O trabalho de Morais (2001) foi o primeiro a aplicar o modelo teórico de Sternberg e Tourangeau (1981) na construção do Teste de Pensamento Metafórico para avaliar a criatividade. O instrumento elaborado por Morais é composto por tarefas de pensamento convergente de raciocínio metafórico em que os sujeitos analisam um estímulo e devem escolher a idéia correta dentre cinco alternativas de resposta (por exemplo: O camelo é _____ do deserto? A. o vitral, B. o burro, C. o barco, D. o armazém, E. o rato). Cada alternativa era pontuada diferencialmente em razão dos graus de equivalência e remotividade que possuíam. Nogueira, Dias e Primi (2003) fizeram uma primeira adaptação desse instrumento,

transformando os itens em tarefas de pensamento divergente no qual, ao invés de ter que escolher respostas apresentadas, os sujeitos deveriam criar respostas metafóricas. Dias (2005) fez um primeiro estudo de validade e precisão dessa nova versão (ver também Dias, Primi & Couto, 2007).

Com base nesses estudos foi criada uma nova versão aprimorando a anterior, com alterações na maneira que o sujeito deve responder, acrescentando informações por meio de um inquérito, de forma a facilitar a pontuação das respostas. Também foram criados itens novos, e outros foram substituídos. Assim, o TCM é uma terceira geração dos estudos iniciados por Morais (2001). Na Figura 1, apresenta-se o formato dos itens do TCM. Esse item preenchido é mostrado na parte inicial do teste em que são dadas as instruções, explicando como o sujeito deve responder ao teste. Sua tarefa consiste em criar até quatro metáforas por item, explicar as relações pensadas e, com cada idéia criada, estabelecer uma analogia (o termo D, que geralmente, nas metáforas, está implícito).

O camelo é o/a _____ do deserto

	<i>Relação</i>	<i>Analogia</i>
1) <u>barco</u>	<u>Porque o camelo balança como o barco no mar quando está em movimento.</u>	<u>mar</u>
2) <u>trenó</u>	<u>Porque o camelo é um transporte do deserto e o trenó é um transporte da neve.</u>	<u>Antártida</u>
3) <u>transporte</u>	<u>Porque o camelo serve de transporte.</u>	<u>veículo</u>
4) <u>árvore</u>	<u>Porque o camelo está no deserto e a árvore está na floresta.</u>	<u>floresta</u>

Figura 1 – Exemplo do formato do item no TCM

Em resumo, a tarefa cognitiva do sujeito nesse teste é (a) codificação e inferência das relações das idéias apresentadas, isto é, a do teor “A” que é apresentado em um contexto semântico minimamente estruturado em razão da apresentação do segundo termo “B” (A:B ou

Camelo _____ do deserto); e (b) a produção de idéias que expressem algo novo do teor, isto é, a produção de veículos (termo C) tais que redefinam o teor. Além disso, pede-se ao sujeito mais detalhes de sua resposta, tais como a “relação” que indica o que ele quis expressar com aquela

idéia e também uma analogia, verificando-se se ele possui clareza dos termos implícitos das analogias que estão subentendidos nesses tipos de metáfora. Sternberg e Nigro (1983) estudaram vários formatos combinando os termos analógicos implícitos nas metáforas (A:B//C:D) e descobriram que o formato adotado no TCE “encoraja os sujeitos a formar uma imagem interativa relacionando o teor ao veículo da metáfora [...] e que isso contribui mais para a criação de metáforas consideradas adequadas” (p. 23). Portanto, esse tipo de formato é propício à estimulação do pensamento divergente criativo implícito na produção de veículos e, por isso, foi escolhida para o TCM. Supõe-se que a criação de metáforas possa ser usada com um indicativo da capacidade do sujeito de criar idéias novas a partir da combinação associativa de idéias conhecidas.

A correção desse instrumento envolve a avaliação de três aspectos associados à qualidade das idéias apresentadas como boas metáforas. Analisa-se se as idéias produzidas pelos sujeitos apresentam equivalência e remotividade, isto é, se preservam estrutura clara de relações entre os termos e se essas relações são distantes. Além disso, avalia-se o quanto as idéias apresentadas conseguem ser facilmente compreendidas por um grupo de juízes. Muitas idéias candidatas a metáfora falham em algum dos três aspectos citados. Portanto, o processo de correção do TCM envolve a avaliação de juízes que atribuem pontos seguindo esses critérios. Um problema nesse teste é que o processo de avaliação de uma idéia como sendo metafórica ou não envolve aspectos da subjetividade da pessoa que faz essa avaliação. Por isso, esse processo pode ser considerado mais vulnerável a erros de classificação. Esse foi o problema enfocado neste estudo.

Uma tentativa de minimizar esses erros é o estabelecimento de critérios de correção suficientemente claros e objetivos para serem compreendidos e utilizados pelos avaliadores. Não obstante, mesmo com critérios adequadamente definidos, é necessário verificar se tais padrões para correção funcionam. A forma de se verificar o grau em que o instrumento se encontra livre de erros é pela estimação de coeficientes de precisão. Tradicionalmente, em psicometria, o conceito de precisão é interpretado por uma relação estabelecida com o erro de medida, uma vez que o resultado de avaliações repetidas de uma mesma característica podem não ser os mesmos em razão dos erros embutidos nos processos de medida. Questiona-se se o escore observado é aquele que representa a verdadeira posição do sujeito no construto. A resposta para essa questão, no modelo da psicometria clássica, é que o escore verdadeiro se encontra em algum lugar em torno da média de escores observados a partir de

repetidas medidas, sendo a faixa do entorno calculada a partir do erro padrão da medida.

Como descrito, a tarefa das pessoas ao responderem ao TCM é produzir metáforas condizentes com o estímulo apresentado. Cada idéia é pontuada por juízes treinados em uma escala de 0 (não metáfora) até 3 (metáfora bem elaborada e original). Portanto, um problema psicométrico nesse caso refere-se à precisão dessas pontuações. Num certo sentido, está-se lidando com a precisão de avaliadores, ou melhor, com a estimativa da fonte de erro ligada ao subjetivismo que diferentes avaliadores teriam ao pontuar as idéias dos itens. A metodologia clássica para estudar essa precisão é por meio da estimação da correlação das pontuações entre dois juízes para as mesmas idéias. Essa análise foi realizada em um estudo anterior, resultando valores bastante adequados (Barros e cols., 2007). Nesse estudo, entretanto, apresenta-se um procedimento alternativo empregando o modelo de Rasch. Esse método será detalhado na seção procedimento. O objetivo geral foi verificar a precisão dos avaliadores, aplicando-se o modelo de Rasch de créditos parciais (Wright & Masters, 1982), e também, em razão de outras potencialidades do modelo de Rasch, estimar a comparabilidade das pontuações.

Método

Participantes

Participaram desta pesquisa nove juízes (três doutorandos, três mestrandos, um professor doutor e dois graduandos em iniciação científica), treinados no método de pontuação desenvolvido para a avaliação da criatividade por meio de metáforas. Desses, três eram homens e seis, mulheres. Os juízes avaliaram 513 idéias de 19 pessoas, das quais 11 responderam à Forma A e oito à Forma B do TCM (Primi e cols., 2006). Todos os participantes concordaram explicitamente em fazer parte deste trabalho, o que foi formalmente registrado na assinatura dos termos de consentimento livre e esclarecido.

Material e procedimento

Material e critérios de pontuação

O instrumento utilizado foi o Teste de Criação de Metáforas – TCM, formas A e B (Primi e cols. 2006), sendo cada forma composta por nove itens, com os dois primeiros comuns às duas formas. Cada item é composto por uma frase com uma lacuna que deve ser utilizada para a construção de uma metáfora (ver Figura 1). Um exemplo de item é “O cabide é o/a _____ da roupa”. Em cada item existem quatro espaços em branco para o sujeito escrever suas idéias, acompanhado de espaço para a

explicação da relação e para a identificação do termo “D” da analogia a ser composta com a resposta metafórica.

Foram utilizados 19 protocolos do TCM previamente respondidos. Esses protocolos foram distribuídos entre os nove juízes, tomando o cuidado para que alguns protocolos fossem corrigidos por mais de um juiz (e um subconjunto dos protocolos foi corrigido pelos nove juízes) de forma independente, isto é, sem que um juiz soubesse a pontuação atribuída pelos outros juízes. Todos foram treinados nos critérios gerais utilizados para julgar metáforas, isto é, avaliando sua equivalência e sua remotividade para que preservassem uma estrutura clara de relações entre os termos e com relações distantes. Nesse sentido, foi criado um sistema de pontuação gradual conforme a descrição a seguir (Primi, 2006):

Pontuação 0

Uma idéia C que não é metáfora. Ex.: O camelo é o meio de transporte do deserto (idéia óbvia, remotividade zero).

Uma idéia C que é um adjetivo de A. Ex.: O camelo é o marrom do deserto.

Uma idéia C que represente uma associação com somente algum dos termos das idéias apresentadas (A ou B). Ex.: O camelo é o *sbeik* do deserto.

Uma idéia C cuja relação proposta não explicita a metáfora ou que esteja errada.

Pontuação 1

Uma idéia C que represente uma metáfora correta, isto é, é equivalente ($r(A:B)=r(C:D)$) e medianamente remota. A remotividade foi julgada a partir da distância entre os campos semânticos da resposta C com as idéias A:B. Não é preciso que o D esteja correto para que o sujeito receba pontuação 1. Basta que a idéia C e a relação sejam corretas.

Pontuação 2

Uma idéia C que atinja o critério para ser pontuação 1 e que também possua remotividade avançada e resposta D da analogia correta. Embora o critério de remotividade possa ser baseado na originalidade da idéia (baixa frequência de ocorrência), nesse momento não foi julgada por esse critério.

Pontuação 3

Uma idéia C que atinja o critério para ser pontuação 2 e que, além disso, possua remotividade muito mais avançada. Uma idéia que se apresente como muito mais nova e criativa que as respostas que recebem pontuação 2 ou idéias com duas ou mais relações claras que utilizam idéias concretas e não vagas.

Procedimento de análise baseado no modelo de Rasch de créditos parciais.

Os dados coletados foram organizados em uma matriz cujas linhas continham as idéias produzidas e as colunas os nove juízes, portanto, uma matriz 513 x 9. Em cada célula foi colocada pontuação do juiz i para a idéia n . Note-se que dessa forma cada juiz é análogo a um item de teste, em uma situação tradicional, e cada idéia análoga a um sujeito. Assim temos uma matriz de 513 idéias que foram pontuadas por 9 juízes (itens). Cada idéia pode receber uma pontuação total pela soma de pontos dos vários juízes que a corrigiram. Na matriz de dados nem todos os juízes pontuaram todas as idéias. De fato: 131 (25,5%) idéias foram pontuadas por dois juízes, 41 (8%) por três juízes, 179 (34,9%) por quatro juízes, 97 (18,9%) por cinco juízes e 65 (12,7%) pelos nove juízes. Utilizou-se um delineamento discutido por Linacre (1998) chamado rede de juízes ancorados (*judge-linking network*). Nesse esquema, uma proporção das respostas foi corrigida por todos os juízes e o restante por pares, trios e assim por diante. Buscou-se maximizar as diferentes combinações de juízes. Esse conjunto de correções conjuntas funciona como “itens âncora” em delineamentos clássicos de equiparação de notas (Wolfe, 2000). A essa matriz foi aplicado o modelo de Rasch de créditos parciais implementado pelo programa WINSTEPS (Linacre & Wright, 1991). O modelo de créditos parciais é dado pela seguinte fórmula (Wright & Masters, 1982):

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

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

Os escores no item têm a notação $x = 0, \dots, m_i$ para um item com $K_i = m_i + 1$ categorias de resposta. Assim,

$$P_{nix}(\theta)$$

indica a probabilidade do sujeito n ter o escore x no item i . Os valores

$$\delta_{ij} \quad (j = 1, \dots, m_i)$$

apontam os valores dos limiares de transição entre a categoria $j-1$ e a categoria j , que indicam o ponto de intersecção entre as curvas da categoria $j-1$ e j . Em situações ideais, esse ponto indicará o momento a partir do qual a j passa a ser a mais provável, portanto, o passo de transição entre a categoria menor $j-1$ e a categoria em consideração à qual esse parâmetro se refere. Esse modelo tem uma série de vantagens cuja discussão foge ao escopo deste artigo (ver mais detalhes em Nunes e cols., 2007). Mas uma das vantagens importantes de serem notadas é que o modelo acomoda os dados ausentes (células de juízes que não pontuaram uma determinada resposta) e atribui uma pontuação equiparável às idéias, isto é, na mesma escala, mesmo que esta tenha sido obtida por diferentes combinações de juízes.

Portanto, essa análise estima uma medida Rasch (teta) para cada idéia a partir dos escores brutos atribuídos pelos juízes. Como neste trabalho não se pretendeu verificar a precisão do instrumento, mas sim do sistema de correção dos itens do instrumento, um coeficiente central foi a correlação juiz-total, indicando o quanto as pontuações de cada juiz se correlacionam com a medida

Rasch, que é uma medida equiparável, a qual leva em conta as pontuações dos outros juízes. Esses coeficientes são equivalentes ao índice de consistência dos avaliadores. Além desses, o WINSTEPS oferece outros índices analíticos e gráficos que permitem depurar o estudo da precisão da aplicação do sistema de pontuações feito pelos juízes, que serão tratados a seguir na seção resultados.

Resultados e discussão

Para a realização das análises desta pesquisa, cada idéia foi corrigida independentemente pelos juízes, atribuindo pontuações seguindo a escala de 0 a 3, indicando o nível de elaboração da metáfora. Os juízes levaram em consideração os critérios de equivalência e remotividade propostos por Sternberg e Tourangeau (1981) e operacionalizados por Primi (2006). Dessa maneira, o valor do teta estimado segundo o modelo de Rasch de créditos parciais corresponde a um índice de metaforicidade da idéia. Na Tabela 2, podem-se verificar as estatísticas descritivas sumarizadas da metaforicidade das idéias avaliadas.

Tabela 2 – Estatísticas descritivas dos valores de teta atribuídos às idéias e índices de ajuste do modelo

	N	Teta	Erro padrão	Infit	Outfit
Média	3,7	-1,09	1,01	0,91	0,94
Desvio padrão	1,9	2,55	0,23	1,21	1,29
Máximo	9	9,68	1,81	9,90	9,90
Mínimo	2	-5,43	0,63	0,05	0,05

O N mostra que cada idéia foi avaliada em média por 3,7 juízes, algumas por no mínimo 2 e no máximo 9. No que diz respeito ao índice de metaforicidade, a média obtida foi de -1,09, isto é, um logit abaixo das médias dos juízes, que são centradas em zero. Isso pode indicar que a maioria as pontuações brutas estão entre 0 e 2, e o valor 3 é bem mais infrequente de se encontrar. Complementando, o desvio padrão de 2,55 indica uma elevada variação na qualidade das metáforas.

Os índices que avaliam a correspondência entre os valores esperados e observados das estimativas teta para as idéias, ou simplesmente os índices de ajuste infit e outfit, mostraram-se adequados, conforme os parâmetros sugeridos por Linacre e Wright (1994a, 1994b), qual seja, inferiores a 1,20. Porém, os valores máximos de infit e outfit foram 9,90, indicando que a pontuação de algumas idéias não se ajustou adequadamente ao que é esperado pelo modelo. Contudo, observando os valores da média e o desvio padrão dos índices de ajuste, pode-se dizer que a maioria das idéias se encontra em torno do limite

aceitável (cerca de 74% dos índices de *infit* e *outfit* são menores do que 1,20, portanto, de acordo com o esperado).

A precisão das estimativas de teta das idéias (índice de metaforicidade) calculada pelo modelo de Rasch foi 0,67 (precisão real) e 0,73 (precisão dos escores modelados). Assumindo cada idéia como um caso e cada juiz como um item de um teste hipotético, o número de juízes (itens) avaliando cada idéia variou de dois a nove ($M=3,7$), um número reduzido, o que leva a uma diminuição no coeficiente de precisão. Essa situação poderia ser comparada à aplicação de um teste contendo em média quatro itens para a avaliação de um determinado sujeito, isto é, uma amostra reduzida de informação. Mesmo assim, o coeficiente de fidedignidade das estimativas de metaforicidade das idéias está ao redor de 0,67, o que indica uma precisão moderada, conforme esperado. Utilizando-se a fórmula de profecia de Spearman-Brown, seria necessário o dobro de juízes (portanto, uma média de 7,4 juízes para cada idéia) para que se obtivesse um coeficiente de 0,80. Isso seria requerido caso fossem necessárias estimativas

muito precisas de cada idéia apresentada pelos sujeitos. Mas de fato, o que se irá desenvolver em trabalhos futuros será a estimativa da precisão das pontuações dos sujeitos, agregando os tetras que um determinado sujeito recebeu nas idéias que produziu. Certamente a estimativa de teta agregado atingirá um índice de precisão mais elevado do que as estimativas baseadas em uma única idéia.

Na Tabela 3 encontram-se as estatísticas descritivas sumarizadas das notas dos juizes. O parâmetro b , estimado segundo o modelo de Rasch, pode ser interpretado como o grau de severidade com que cada um dos juizes avalia as idéias. Esse valor representa a média dos três valores na escala teta dos limiares de transição das categorias dos escores brutos 0-1, 1-2 e 2-3.

Tabela 3 – Estatísticas descritivas dos índices de dificuldade e ajuste dos juizes

	N	b	Erro padrão	<i>Infit</i>	<i>Outfit</i>
Média	145,2	0,00	0,20	0,97	0,96
Desvio padrão	81,4	1,28	0,07	0,25	0,29
Máximo	268,0	2,44	0,32	1,42	1,41
Mínimo	47,0	-2,12	0,12	0,58	0,44

Os valores de N mostram que cada juiz avaliou em média 145,2 idéias. A média do b obtida foi de 0,00 e desvio padrão 1,28. Os índices que avaliam o ajuste geral dos juizes (*infit* e *outfit*) mostraram-se adequados. As médias indicam que houve uma alta consistência dos padrões de atribuição de nota pelos juizes, uma vez que, levando-se em conta os desvios padrão, tendem a se encontrar abaixo de 1,20. Os valores máximos de 1,42 para *infit* e 1,41 para *outfit*, nesse caso, podem ser interpretados como a ocorrência de uma leve incongruência em um número baixo de juizes.

A Tabela 4 apresenta os valores médios de b para cada juiz, indicando o índice médio de severidade correspondente. Cada juiz possui um valor médio de b porque o modelo Rasch de créditos parciais identifica valores dos limiares, ou seja, o valor equivalente em teta da transição de uma nota para outra, como, por exemplo, da nota 0 para nota 1. Como foi dito, o índice b representa a média dos três valores de transição entre as categorias, calculados para cada juiz. Pode-se notar que o juiz mais leniente foi o de número 8, ao passo que o mais austero foi o de número 4.

Tabela 4 – Índices de dificuldade, ajuste e precisão dos juizes

	b	Erro padrão	<i>Infit</i>	<i>Outfit</i>	Correlação Juiz-Total
Juiz 1	1,26	0,12	0,84	0,82	0,81
Juiz 2	-0,74	0,16	0,92	0,86	0,75
Juiz 3	1,01	0,27	1,42	1,41	0,61
Juiz 4	2,44	0,14	0,92	1,22	0,72
Juiz 5	-0,73	0,17	1,21	1,20	0,69
Juiz 6	0,07	0,14	1,18	1,19	0,68
Juiz 7	-0,61	0,26	0,73	0,67	0,78
Juiz 8	-2,12	0,32	0,58	0,44	0,87
Juiz 9	-0,57	0,18	0,92	0,85	0,74

Com relação ao ajuste, por um lado pode-se notar que o juiz de número 3 apresenta índices inadequados, ou seja, acima de 1,20, e os juizes 5 e 6 apresentam valores limítrofes. Por outro lado, os juizes com índices mais adequados foram os de número 8, 7 e 1. No que respeita à concordância entre os juizes, pode-se notar que os coeficientes de correlação juiz-total mostraram-se bastante apropriados, sendo aqueles que melhor se correlacionaram com todos os demais, ou seja, que demonstraram maior concordância, os juizes 8, 1 e 7.

Os resultados até aqui encontrados atestam a adequada precisão do sistema de avaliação da criatividade por meio de metáforas. O coeficiente de precisão mostrou-se adequado, levando-se em conta o número de juizes. Além disso, com exceção do juiz 3, todos os coeficientes de concordância (correlação juiz-total) foram acima de 0,68. Além dessas análises, a utilização da Teoria de Resposta ao Item, em especial o modelo de Rasch, permite uma exploração mais detalhada dos dados de desajuste e padrões de pontuação adotados pelos juizes, como exposto a seguir.

Os resultados expressos na Figura 2 representam exemplos de inconsistência na utilização dos critérios para atribuição das notas pelos juízes. Estes foram ordenados em forma crescente de severidade e as idéias estão ordenadas de forma decrescente pelo valor de teta. As primeiras três linhas (acima da linha tracejada *high*) apresentam o número da

idéia em consideração. Note que os números estão em posição vertical (idéia 138, 7, 137, e assim por diante). Isso é repetido nas últimas três linhas. No corpo do texto estão apresentados escores observados não correspondentes ao que prevê o modelo de Rasch a partir dos índices de severidade do juiz e de metaforicidade da idéia em consideração.

		1	1211121111111112211	1222222	1111	122	1111	2
		3	303220333218214376650875221433126010122209631289					
		8	7782520310329469629398255627865560295998760877007					
			high-----					
Juiz 8	-2.12	a	0
Juiz 2	-.74	d 0	222
Juiz 5	-.73	C 0	0	2.2200.0	.2.2000
Juiz 9	-.57	E	. . 0 . 0 0 2 20
Juiz 6	.07	D	11 . . 0	0002222200222	.02.22.21
Juiz 3	1.01	A	. . . 0 . 0 . 00 . 3
Juiz 1	1.26	c	3331	2
Juiz 4	2.44	B	. 3 . . . 0 2 111.1	1
			-----low-					
		7	1211121111111112211	165122222241111	1612211111	19231289		
		3	303220333218214376390875221833120010922200677007					
		8	782520310329469629	8255627	6556	295	9876	8

Figura 2 – Respostas mais inconsistentes dos juízes

As inconsistências podem ser explicadas pela relação entre os valores de teta das idéias e os respectivos valores de severidade dos juízes ao atribuir as notas. Pode-se observar, por exemplo, que na idéia 138 o juiz 1 atribuiu nota 3, quando o modelo matemático, considerando a severidade desse juiz e o teta da idéia, esperava que ele aferisse nota 2. Já no extremo oposto, mais direito do gráfico, o juiz 4 atribuiu nota 1 à idéia 97, quando o esperado era 0. Essa figura apresenta um nível detalhado identificando os casos mais inconsistentes (ressalta-se que o juiz de número 7 não apresentou inconsistência na atribuição de notas que fizessem com que ele constasse na tabela) e permitindo, por exemplo, que o grupo de juízes possa posteriormente discutir cada caso, buscando aprimorar os critérios de pontuação. Portanto, esse gráfico é muito útil no treinamento dos critérios de pontuação.

A Tabela 5 descreve a estatística dos escores de cada item conforme atribuídos pelos nove juízes. Nessa tabela pode-se observar, da segunda à quarta coluna, a frequência e porcentagem correspondente

dos escores que cada juiz atribuiu. Na quinta coluna (média de teta), observam-se os valores médios de teta ou a medida da metaforicidade das idéias correspondentes às pontuações brutas (valor do escore). Por exemplo, para o juiz 1, as idéias para as quais ele atribuiu nota 0, ou seja, considerou inadequadas a resposta do indivíduo, a relação, a remotividade e a analogia, a metaforicidade recebe em média o valor -3,75 calculado pelo modelo de Rasch. Ou seja, de maneira geral para esse juiz, o valor médio de -3,75 em teta indica que a idéia não se constitui como metáfora. Da mesma maneira, o teta -0,63 foi o escore médio das idéias que receberam escore 1, ou seja, uma metáfora que apresenta alguma remotividade e adequada relação. Como se pode perceber, alguns juízes não atribuíram nota 3 para nenhuma idéia, o que pode ser interpretado como um alto índice de severidade. É esperado que as médias de teta cresçam consecutivamente ao aumento dos escores observados. Se isso não ocorre, há um problema de consistência no sistema de pontuação adotado pelo eventual juiz.

Tabela 5 – Estatísticas descritivas dos escores dos itens para cada juiz

	Valor do escore	N	%	Média de teta	Erro padrão	<i>Outfit</i>	Correlação Juiz-Total
Juiz 1	0	50	17	-3,75	0,19	0,7	-0,67
	1	123	43	-0,63	0,11	0,8	-0,13
	2	110	38	1,51	0,15	0,9	0,59
	3	4	1	4,91	0,84	1,0	0,26
Juiz 2	0	31	19	-3,08	0,30	0,8	-0,62
	1	78	48	-0,35	0,14	0,7	-0,07
	2	52	32	1,81	0,24	1,0	0,61
	3	1	1	7,13			0,18
Juiz 3	0	10	20	-1,97	1,05	3,3	-0,40
	1	21	42	-1,00	0,21	0,4	-0,29
	2	18	36	1,48	0,21	0,6	0,62
	3	1	2	0,53*		1,4	0,05
Juiz 4	0	78	45	-2,07	0,23	1,0	-0,64
	1	54	31	0,00	0,18	1,8	0,11
	2	42	24	2,10	0,17	0,7	0,60
	3	1	1	2,62		1,1	0,10
Juiz 5	0	33	19	-3,31	0,37	1,5	-0,60
	1	99	57	-0,35	0,17	1,1	0,03
	2	43	25	1,70	0,21	1,1	0,51
Juiz 6	0	71	25	-2,99	0,22	0,9	-0,66
	1	165	57	0,22	0,12	1,2	0,26
	2	50	17	1,78	0,25	1,5	0,41
	3	1	0	7,13			0,18
Juiz 7	0	13	26	-2,70	0,59	0,8	-0,68
	1	16	32	-0,54	0,28	0,6	-0,09
	2	21	42	1,45	0,19	0,6	0,68
Juiz 8	0	5	10	-4,64	0,99	0,9	-0,68
	1	17	34	-1,34	0,15	0,2	-0,36
	2	28	56	1,16	0,20	0,6	0,76
Juiz 9	0	26	16	-2,77	0,45	1,3	-0,50
	1	97	60	-0,53	0,13	0,6	-0,19
	2	38	23	2,49	0,24	0,6	0,65
	3	1	1	7,13			0,18

De modo geral, os padrões são semelhantes para todos os juízes. A única exceção é idéia pontuada como 3 pelo juiz 3, que não possui um teta mais elevado que as idéias pontuadas por ele com valor 2 (essa inversão é marcada automaticamente pelo WINSTEPS com um asterisco). Entretanto, como se trata de uma única idéia, isso não compromete a precisão geral do sistema. Isso reflete uma idiosincrasia que naturalmente existe entre os juízes na aplicação dos critérios de pontuação. Outro índice importante para compreender a concordância entre os juízes é a análise dos valores médios do índice de ajuste *outfit* (sétima coluna). Esse índice será tanto mais alto quanto mais houver discrepância entre os escores observados e os preditos pelo modelo. Pode-se notar que, em geral, os juízes mostraram-se coerentes ao

atribuir os escores para cada idéia conforme os valores de teta. Os poucos desajustes encontrados foram para o juiz 3 atribuir nota 0 e 3, o juiz 4 ao atribuir nota 1, e os juízes 5, 6 e 9 atribuírem nota 0, 2, 0, respectivamente. Em conjunto com as informações apresentadas na Figura 2, esses índices podem identificar problemas localizados de alguns juízes em atribuir determinadas pontuações. Mais uma vez, isso pode ser motivo para entendimento entre os juízes para se tentar aprimorar os critérios e a concordância. Em resumo, essa tabela é muito informativa para identificar quais problemas podem estar ocorrendo nas aplicações dos critérios. A fim de ilustrar os valores de teta que representam a transição de uma nota para a outra, a Figura 3 apresenta os valores dos limiares (*thresholds*) para cada juiz.

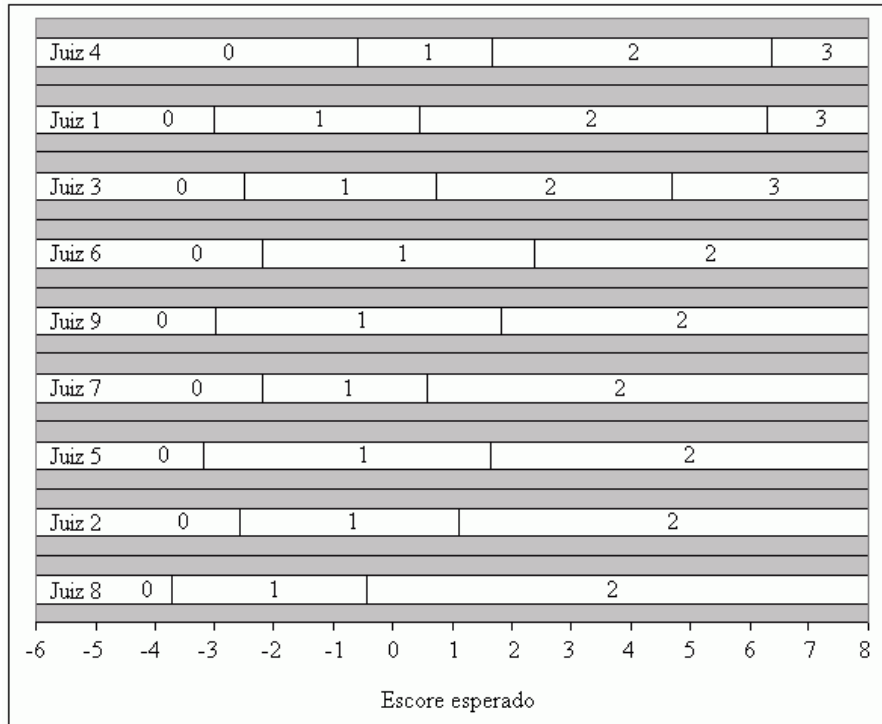


Figura 3 – Valores dos limiares de cada juiz

Pode-se verificar que juízes com limiares mais próximos se agrupam, como, por exemplo, os casos dos juízes 9, 7, 5 e 2, e dos juízes 1 e 3. O juiz 4 apresentou o limiar mais alto de transição entre as notas 0 e 1, ao passo que o juiz 8 mostrou o limiar mais baixo. Esse fato indica que o juiz 4 é mais severo e o juiz 8 é mais leniente para esse limiar. Já no caso da mudança da nota 1 para a 2, o juiz 6 foi o mais severo, sendo o mais leniente novamente o juiz 8. Apesar dessas diferenças, observa-se que em geral os juízes não atribuem nota 1 para tetras inferiores a -3,73; 2 para tetras inferiores a -0,51 e 3 para tetras inferiores a 4,57. Vale ressaltar que, em relação à nota 3, nem todos os juízes aferiram essa nota.

A Figura 4 ilustra, a título de exemplo, a distribuição dos limiares (*thresholds*) do juiz 1. No gráfico, pode-se perceber

que o limiar de transição da nota 0 para a 1 equivale ao valor -2,99 de teta, da nota 1 para a 2 equivale ao teta 0,47 e da nota 2 para a 3 equivale a 6,29. Neste exemplo, fica clara a representação das notas de 0 a 3 em faixas distintas de teta. Cada linha representa uma nota atribuída pelo juiz à idéia. A não-sobreposição dos picos das curvas, isto é, a separação destes em regiões distintas mostra o uso adequado dos critérios de avaliação dos níveis de qualidade da metáfora. O uso inadequado estaria representado pela sobreposição de picos, ou seja, a ausência de regiões claras e distintas na escala de teta para as diferentes pontuações do teste, o que comprometeria a medida. Esse mesmo padrão foi observado para todos os oito juízes restantes.

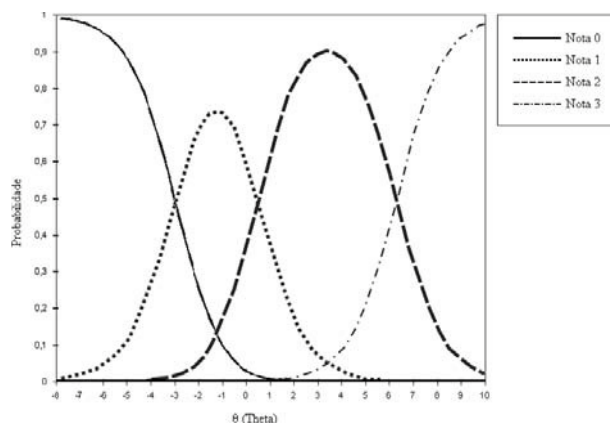


Figura 4 – Curvas características das pontuações e limiares do juiz 1

No que respeita à precisão de avaliadores, as notas dos juízes, bem como seus limites de mudança, estão apresentados numa escala comum que permite a calibração dos escores atribuídos. Essa calibração permite a comparação das notas dos juízes. Ou seja, quando um juiz atribui uma certa nota, esta apresenta um valor correspondente na escala de teta. Por exemplo, uma idéia pode receber uma nota 2 de um juiz e uma nota 1 de outro juiz, porém ambas podem compartilhar um mesmo valor de teta ou um valor muito próximo. Isso permite equiparar as pontuações dos juízes e se chegar a um valor mais preciso em razão do equilíbrio dos níveis de exigência entre eles.

Considerações finais

O objetivo do presente estudo foi verificar a precisão dos critérios de correção de um instrumento de medida de criatividade por meio da produção de metáforas. Em termos de precisão de avaliadores, objeto principal deste estudo, em geral os índices de correlação juiz-total foram bem adequados, com média 0,74 (DP=0,08). O coeficiente de precisão de cada idéia estimado utilizando-se o modelo de Rasch foi de 0,67. Dadas as circunstâncias do método de avaliação, qual seja, um número médio reduzido de juízes por idéia, o coeficiente mostrou-se bastante adequado. Em situações de análise mais favoráveis, com número maior de juízes, o valor tenderia a aumentar, conforme verificado pela fórmula da profecia de Spearman-Brown.

A aplicação do modelo de Rasch permitiu ainda realizar estudos mais detalhados quanto à atribuição de notas por parte dos juízes. Os resultados demonstraram que os juízes podem variar no que diz respeito ao grau de severidade de pontuação de uma idéia, ou seja, uma idéia que receba nota 3 de um juiz pode receber nota 2 de outro juiz. No entanto, essas possíveis idiosincrasias são solucionadas na medida em que esses valores são ordenados em um escala comum, na qual as notas dadas por cada juiz são convertidas para um valor equivalente de teta.

Os resultados encontrados são promissores no que diz respeito aos critérios desenvolvidos operacionalizados a partir de Sternberg e Nigro (1983) e como método de avaliação da criatividade. Assim sendo, novas pesquisas são sugeridas, ampliando o número de juízes por idéia, o que aumentaria a precisão da estimação da metaforicidade. Outro procedimento que poderia aumentar a credibilidade do sistema de correção seria o treinamento de novos juízes, a fim de se verificar a replicabilidade do método.

Referências

- Ackerman, P. L. (1996). A theory of intellectual development: Process, personality, interests, and knowledge. *Intelligence*, 22, 227-257.
- Ackerman, P. L. & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, 121, 219-245.
- Ackerman, P. L. & Beier, M. E. B. (2005). Knowledge and Intelligence. Em: O. Wilhelm & R. W. Engle (Orgs.). *Handbook of understanding and measuring intelligence*. (pp. 125-139). Thousand Oaks, CA: EUA.
- Anastasi, A. & Urbina, S. (2000). *Testagem psicológica*. Porto Alegre: Artes Médicas.
- Barros, D. P., Miguel, F. K., Cunha, T. F., Cruz, M. B. Z., Couto, G., Muniz, M. & Primi, R. (2007). Precisão e validade de um teste de criatividade por meio da produção de metáforas. *Livro de Resumos – Painéis do III Congresso Brasileiro de Avaliação Psicológica e XII Conferência Internacional de Avaliação Psicológica: Formas e Contextos*, 2007, João Pessoa.
- Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin*, 40, 153-193.
- Cronbach, L. J. (1996). *Fundamentos da testagem psicológica*. Porto Alegre: Artes Médicas.
- Dias, A. R. (2005). *Avaliação da criatividade por metáforas*. Dissertação de Mestrado. Itatiba, SP: Universidade São Francisco – Programa de Pós-Graduação *Stricto Sensu* em Psicologia.
- Dias, A. R., Primi, R. & Couto, G. (2007). Avaliação da criatividade por meio da produção de metáforas. Manuscrito submetido à publicação.
- Duncan, J., Burgess, P. & Emslie, H. (1995). Fluid intelligence after frontal lobe lesions. *Neuropsychologia*, 33(3), 261-268.
- Guilford, J. P. & Christensen, P. R. (1973). The one-way relation between creative potential and IQ. *Journal of Creative Behavior*, 7, 247-252.
- Linacre, J. M. (1998). Rating, judges and fairness. *Rasch Measurement Transactions*, 12(2), p. 630-1. Disponível: <http://www.rasch.org/rmt>. Acessado em 22/06/2007.
- Linacre, J. M. & Wright, B. D. (1991). *WINSTEPS - Rasch Model Computer Programs*. Chicago: MESA Press.
- Linacre, J. M., & Wright, B. D. (1994a). Chi-square fit statistics. *Rasch Measurement Transactions* [On-line] 8(2), 350. Obtido em 31 de janeiro de 2001 do World Wide Web: <http://www.rasch.org/rmt/rmt82a.htm>.

- Linacre, J. M. & Wright, B. D. (1994b). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(2), 370. Obtido em 31 de janeiro de 2001 do World Wide Web: <http://www.rasch.org/rmt/rmt83.htm>.
- Mackinnon, D. W. (1962). The nature and nurture of creative talent. *American Psychologist*, 17, 484-495.
- Morais, M. F. (2001). *Definição e avaliação da criatividade: uma abordagem cognitiva*. Braga: IEP: Universidade do Minho.
- Nogueira, B. T. B., Dias, A. R. & Primi, R. (2003). *Criando metáforas: estudo piloto*. Itatiba, SP: Universidade São Francisco – Laboratório de Avaliação Psicológica e Educacional.
- Nunes, C. H. S. S., Primi, R., Nunes, M. F. O., Muniz, M., Cunha, T. F. & Couto, G. (2007). *Análise de instrumentos com escalas tipo Likert por teoria de resposta ao item*. Manuscrito submetido à publicação.
- Primi, R. (1998). *Desenvolvimento de um instrumento informatizado para avaliação do raciocínio analítico*. Tese de Doutorado. São Paulo: Universidade de São Paulo – Instituto de Psicologia.
- Primi, R. (2006). *Teste de Criação de Metáforas – critérios de pontuação e interpretação*. Itatiba, SP: Universidade São Francisco – Laboratório de Avaliação Psicológica e Educacional.
- Primi, R., Miguel, F. K., Cruz, M. B. Z., Couto, G., Barros, D. P., Muniz, M. & Cunha, T. F. (2006). *Teste de Criação de Metáforas – formas A, B e C*. Itatiba, SP: Universidade São Francisco – Laboratório de Avaliação Psicológica e Educacional.
- Sternberg, R. J. (1977). A component process in analogical reasoning. *Psychological Review*, 84(4), 353-378.
- Sternberg, R. J. (1983). Components of human intelligence. *Cognition*, 15, 1-48.
- Sternberg, R. J. & Nigro, G. (1983). Interaction and analogy in the comprehension and appreciation of metaphors. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 35, 17-38.
- Sternberg, R. J. & Lubart, T. I. (1991). An investment theory of creativity and its development. *Human Development*, 34, 1-31.
- Sternberg, R. J. & O'Hara, L. A. (2000). Intelligence and creativity. Em R. J. Sternberg (Ed.). *Handbook of intelligence* (pp. 611-630). New York, NY: Cambridge University Press.
- Tourangeau, R. & Sternberg, R. J. (1981). Aptness in metaphor. *Cognitive Psychology*, 13, 27-55.
- Tourangeau, R. & Sternberg, R. J. (1982). Understanding and appreciating metaphors. *Cognition*, 11, 203-244.
- Tourangeau, R. & Rips, L. (1991). Interpreting and evaluating metaphors. *Journal of Memory and Language*, 30, 452-472.
- Wolfe, E. W. (2000). Understanding rasch measurement: equating and item banking with rasch model. *Journal of Applied Measurement*, 1(4), 409-434.
- Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA.
- Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: MESA.

Recebido em fevereiro de 2007
Reformulado em agosto de 2007
Aprovado em outubro de 2007

Sobre os autores:

Ricardo Primi é psicólogo pela PUCCampinas, doutor em Psicologia Escolar e do Desenvolvimento Humano pela Universidade de São Paulo – com parte desenvolvida na Yale University (EUA) –, coordenador do Laboratório de Avaliação Psicológica e Educacional (LabAPE) e recebe financiamento da Fapesp e de produtividade em pesquisa do CNPq. É diretor acadêmico de Pós-Graduação da Universidade São Francisco e professor do Mestrado e Doutorado em Avaliação Psicológica da mesma universidade, membro da Comissão Consultiva em Avaliação Psicológica do Conselho Federal de Psicologia e da Comissão da Área de Psicologia INEP/MEC.

Fabiano Koich Miguel é psicólogo pela Universidade Presbiteriana Mackenzie, mestre e doutorando em Avaliação Psicológica pela Universidade São Francisco e bolsista Fapesp vinculado ao Laboratório de Avaliação Psicológica e Educacional – LabAPE

Gleiber Couto é psicólogo pela PUC-MG, doutorando em Avaliação Psicológica pela Universidade São Francisco – USF e bolsista Capes vinculado ao Laboratório de Avaliação Psicológica e Educacional – LabAPE.

Monalisa Muniz é psicóloga e doutoranda em Avaliação Psicológica pela Universidade São Francisco – USF e bolsista FAPESP vinculada ao Laboratório de Avaliação Psicológica e Educacional – LabAPE e ao Laboratório de Avaliação Psicológica em Saúde Mental – LAPSaM.