Universidade de Brasília

Instituto de Psicologia

Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações

Tese de Doutorado

Adaptação de uma Medida de Inteligência Emocional para um Teste Adaptativo Computadorizado para o Contexto Brasileiro

Adaptation of a Measure of Emotional Intelligence into a Computer-Adaptive Test for the Brazilian Context

Victor Vasconcelos de Souza

Brasília – DF

Junho de 2023

Universidade de Brasília

Instituto de Psicologia

Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações

Adaptação de uma Medida de Inteligência Emocional para um Teste Adaptativo Computadorizado para o Contexto Brasileiro

Victor Vasconcelos de Souza

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações como requisito parcial à obtenção de grau de Doutor em Psicologia Social, do Trabalho e das Organizações.

Orientadora: Prof. Dra. Cristiane Faiad

Brasília – DF

Junho de 2023

Tese de doutorado defendida diante e avaliada pela banca examinadora constituída por:

Prof.ª Dr.ª Cristiane Faiad (Orientadora)

Instituto de Psicologia

Universidade de Brasília

Prof. Dr. Fabio Iglesias

Instituto de Psicologia

Universidade de Brasília

Prof.ª Dr.ª Ana Carolina Zuanazzi

Instituto Ayrton Senna

Prof. Dr. Mauricio Sarmet

Unidade de Gestão e Negócios

Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Prof.ª Dr.ª Girlene Ribeiro (Suplente)

Faculdade de Educação

Universidade de Brasília

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Abreviações

Material Suplementar

Resumo

Este trabalho teve como objetivo adaptar dois testes de inteligência emocional (IE) para um método de administração computadorizado adaptativo e para uma amostra brasileira: o Teste Situacional de Compreensão Emocional (*Situational Test of Emotional Understanding* [STEU]) e o Teste Situacional de Gerenciamento Emocional (*Situational Test of Emotional Management* [STEM]). O primeiro manuscrito estabeleceu a perspectiva teórica adotada no decorrer do trabalho. No segundo manuscrito, esta perspectiva serviu de base para a tomada de decisão editorial no decorrer do procedimento de retrotradução dos testes. Em seguida, foram analisadas as evidências de validade das versões adaptadas do STEU e STEM. Os testes foram aplicados em 688 participantes adultos que também responderam à Escala de Satisfação com a Vida (SWLS) e ao Teste Reduzido dos Fatores de Personalidade (ER5FP), em uma replicação parcial do estudo original de validação de MacCann e Roberts (2008). Os resultados revelaram que as versões finais dos testes adaptados, com 32 (STEU) e 30 itens (STEM), obtiveram bons índices de ajuste globais na análise fatorial confirmatória, e globais e por item nas modelagens de teoria de resposta ao item (TRI). Os coeficientes de correlação com o SWLS e os cinco fatores do ER5FP fundamentaram as evidências de validade com base em relações com variáveis externas. No terceiro manuscrito foram discutidas as vantagens, desvantagens, e aspectos práticos da testagem adaptativa computadorizada (CAT). No quarto manuscrito, apresentou-se a adaptação dos testes STEU e STEM para uma CAT. Para isso, foram realizadas simulações de administrações adaptativas dos testes utilizando os padrões de resposta previamente coletados. Algoritmos de seleção de itens (ISA) baseados na maximização da informação coletada foram capazes de obter estimativas de habilidade e erro padrão equivalentes com a administração de menos itens para 368 (53,5%) participantes. Os algoritmos baseados na informação de Fisher foram mais eficientes, mas também mais enviesados do que aqueles baseados na informação de Kullback-Leibler. Em conjunto, e levando em consideração o princípio de *validity by design* proposto por

Mislevy (2007), as evidências de validade com base no conteúdo do teste acumuladas por meio do procedimento de adaptação transcultural e de construção do teste com base na teoria de Roseman (2001) e no paradigma do teste de julgamento situacional formam um argumento convincente de validade. Além disso, o resultado das modelagens estatísticas revelou adequação dos índices de ajuste, enquanto os resultados das simulações revelaram a pertinência da utilização dos testes como CATs. Além de contribuir para o desenvolvimento da pesquisa em IE, a disponibilização de um teste psicológico como CAT tem desdobramentos importantes. O desenvolvimento de medidas neste formato tem benefícios para validade e precisão da testagem. Para os benefícios do CAT que exigem grandes bancos de item, a disponibilização dos itens do STEU e do STEM com os parâmetros psicométricos permitirá o emprego de delineamentos de blocos incompletos balanceados, facilitando a multiplicação de formas dos testes. Isto é particularmente interessante para estes testes, que tiveram métodos sistemáticos de construção de itens com base em teorias causais objetivas. Espera-se que os testes adaptados sejam utilizados na testagem futura do modelo da teoria Mayer–Salovey–Caruso no Brasil, mas também que novos testes psicológicos sejam construídos no modelo de CAT.

Palavras-chave: inteligência emocional, teoria Mayer–Salovey–Caruso, testagem adaptativa computadorizada, teoria de resposta ao item, algoritmos de seleção de itens.

Abstract

This study aimed to adapt two emotional intelligence (EI) tests for a computerized and adaptive administration method and for a Brazilian sample, the Situational Test of Emotional Understanding (STEU) and the Situational Test of Emotional Management (STEM). The first manuscript established the theoretical perspective adopted throughout the work. In the second manuscript, this perspective served as the basis for editorial decision-making during the backtranslation procedures of the tests. Then, the evidence of validity of the adapted versions of STEU and STEM was analyzed. The tests were applied to 688 adult participants, who also answered the Life Satisfaction Scale (SWLS) and the Reduced Test of Personality Factors (ER5FP), in a partial replication of the original study by MacCann and Roberts (2008). The results revealed that the final versions of the adapted tests, with 32 (STEU) and 30 items (STEM), obtained good global fit indices in the confirmatory factor analysis, and global and per-item fit in the IRT models. The correlation coefficients with the SWLS and the five factors of the ER5FP supported the evidence of validity based on relationships with external variables. In the third manuscript, the advantages, disadvantages, and practical aspects of computer-adaptive testing (CAT) were discussed. In the fourth manuscript, the adaptation of the STEU and the STEM for a CAT was presented. For this, simulations of CAT administrations of the tests were carried out using the response patterns previously collected from the participants. Item selection algorithms (ISA) based on maximizing collected information were able to obtain equivalent ability estimates and standard error levels with the administration of fewer items for 368 (53.5%) participants. Algorithms based on Fisher information were more efficient but also more biased than those based on Kullback-Leibler. In sum, considering the principle of validity by design proposed by Mislevy (2007), validity evidence accumulated through cross-cultural adaptation procedure and test construction based on Roseman's (2001) theory and situational judgment testing paradigm form a convincing argument for validity. In addition, results from statistical models revealed

adequate fit indices, and simulation results revealed adequacy of using adapted tests as CATs. In addition to contributing to development of EI research, availability of a psychological test such as CAT has important consequences. Developing measures in this format has benefits for test validity and reliability. Availability of items along with their IRT parameters also allows use of balanced incomplete block designs, facilitating multiplication of test forms. This is particularly interesting for STEU and STEM, which had systematic item construction based on objective theory. It is hoped that these adapted tests will be used in future testing of Mayer-Salovey-Caruso theory model in Brazil but also that new psychological tests will be built using CAT methodology.

Keywords: emotional intelligence, situational judgment tests, computer-adaptive testing, item response theory, item selection algorithms.

Apresentação

A inteligência emocional (IE) é um conjunto de habilidades mentais de raciocínio envolvidas no reconhecimento de emoções nas pessoas, na arte e nos contextos sociais, no entendimento do significado das palavras que descrevem emoções, e no gerenciamento e utilização das próprias emoções e das emoções dos outros para a resolução de problemas (Mayer et al., 2016). O construto tem recebido interesse crescente devido ao reconhecimento de sua importância para a vida humana (Mayer et al., 2008; Schneider & McGrew, 2018). Este reconhecimento tem ocorrido de duas formas. Primeiro, por meio da construção de uma literatura que investiga o papel que a inteligência emocional tem nas áreas mais importantes da vida humana, incluindo a vida pessoal, educacional, e profissional das pessoas (Mayer et al., 2008; Schneider & McGrew, 2018), bem como a possibilidade de treiná-la. Segundo, pela introdução do construto no modelo de inteligência geral, lado a lado com outras habilidades de raciocínio tradicionalmente identificadas como cruciais para a resolução de problemas (Schneider & McGrew, 2018).

Este último desenvolvimento é consequência de um esforço de mais de duas décadas. Mayer et al. (2000) se empenharam na construção de uma teoria de habilidades que fosse capaz de superar as reservas que pesquisadores da área de inteligência geral tinham com este construto (Mayer et al., 2008; Woyciekoski & Hutz, 2009). Apesar de o reconhecimento de aspectos socioemocionais na inteligência datar do início do século XX (Thorndike, 1920), a introdução do construto em um contexto em que muitas afirmações eram feitas sobre a possibilidade do uso da IE como autoajuda, com base em pouca ou nenhuma evidência, afastou pesquisadores tradicionais da área de inteligência geral de estudos mais profundos sobre o tema (Drigas & Papoutsi, 2018).

Desde 2014, esforços renovados têm sido feitos para demonstrar a viabilidade da IE como uma dimensão da inteligência geral com base em estudos de modelagem por equações

estruturais (Evans et al., 2019; MacCann et al., 2014). Esses estudos mostram consistentemente a pertinência de inclusão da IE como uma habilidade ampla no modelo de inteligência geral.

A introdução probatória do modelo de habilidade de IE da teoria Mayer–Salovey–Caruso (MSC; Mayer et al., 2016) na teoria Cattell–Horn–Carroll (CHC; Schneider & McGrew, 2018) de inteligência geral evidenciou as dificuldades associadas à dimensionalidade da IE (Schneider & McGrew, 2018). A escassez de instrumentos para medir a IE na perspectiva de habilidade da teoria MSC tem sido apontada como uma das dificuldades para elucidar esta questão (Bru-Luna et al., 2021; O'Connor et al., 2019; Santos et al., 2015).

No Brasil, apesar da existência de importantes instrumentos de mensuração de habilidades descritas pela teoria MSC como percepção (Miguel & Primi, 2014) e regulação emocional (Miguel et al., 2016), esta lacuna é ainda maior, tendo em vista a ausência do teste dos próprios autores da teoria, o *Mayer–Caruso–Salovey Emotional Intelligence Test* (MSCEIT; Mayer et al., 2002), que é um dos poucos testes tipicamente utilizados internacionalmente (Bru-Luna et al., 2021). Considerando que o teste de Miguel et al. (2016) é de autorrelato, há uma escassez de testes de desempenho de regulação emocional no Brasil, observada na revisão dos testes de IE utilizados no Brasil por Gonzaga e Monteiro (2011). Isto é um problema, visto que há diversas recomendações pelo uso de testes de desempenho na medição da IE na perspectiva de habilidade (Grubb & McDaniel, 2007; Mayer et al., 2016; Schneider & McGrew, 2018), e que a regulação emocional é justamente a área da IE sobre a qual há dúvidas (Gignac, 2005; Palmer et al., 2005; Rossen et al., 2008).

Com o intuito de identificar testes prospectivos para adaptação para o contexto brasileiro, foram estudadas as medidas de desempenho de IE utilizadas internacionalmente. O Teste Situacional de Compreensão Emocional (*Situational Test of Emotional Understanding* [STEU]) e o Teste Situacional de Gerenciamento Emocional (*Situational Test of Emotional Management* [STEM]) foram identificados como medidas atrativas para adaptação, tendo em

vista que além de mensurarem dimensões da teoria MSC, os testes foram desenvolvidos por meio de um método sistemático de construção de itens baseado numa teoria substantiva de emoção (Roseman, 2001) e no paradigma do teste de julgamento situacional (Motowidlo et al., 1990). Além disso, desenvolvimentos recentes na área de uso do computador em psicologia têm possibilitado a discussão de metodologias atualizadas de administração de testes. A presença ubíqua de computadores na sociedade (Instituto Nacional de Estudos e Pesquisas em Avaliação Educacional Anísio Teixeira, 2023) tem tornado mais eficiente o recrutamento de uma amostra minimamente adequada para realização de pesquisa e o crescimento do poder computacional tem tornado possível a execução dos algoritmos estatísticos mais avançados.

Neste contexto, a ferramenta que tem se destacado na testagem psicológica e educacional é a testagem adaptativa computadorizada (CAT; Becker & Bergstrom, 2013). O CAT é contrastado com a testagem de formato fixo e é caracterizado pelo emprego de um algoritmo que seleciona itens para uma sessão de testagem de forma responsiva ao nível de habilidade dos participantes. O objetivo é maximizar a eficiência da estimativa da habilidade utilizando informações provenientes da teoria de resposta ao item (TRI) sobre a dificuldade e a capacidade de discriminação entre diferentes níveis de habilidade que os itens possuem (Luecht, 2016).

Os benefícios do emprego do CAT para a testagem são diversos (Luecht & Sireci, 2011). Primeiro, como um teste de formato fixo precisa, tipicamente, discriminar tanto entre os níveis de habilidade mais alto quanto os mais baixos, todo participante precisa responder a itens que trazem pouca ou nenhuma informação sobre si. Omitindo a aplicação destes itens, a aplicação adaptativa economiza o tempo e o investimento cognitivo dos participantes (Luecht, 2016). Outros benefícios possíveis incluem aumento da validade, precisão e segurança da testagem e redução de vieses de reaplicação de testes (Wise, 2018), além da melhora do engajamento do examinando com o teste (Martin & Lazendic, 2018).

Tendo isso em vista, o objetivo deste trabalho foi avaliar a possibilidade de aplicar de forma eficiente uma versão brasileira do STEU e do STEM num formato de CAT sem prejuízo para validade e precisão. Os testes, construídos em inglês por MacCann e Roberts (2008) avaliam compreensão emocional e gerenciamento emocional, dois fatores da teoria MSC.

Com o objetivo de realizar uma organização conceitual para orientar os próximos passos da adaptação dos testes, o primeiro manuscrito é responsável por realizar uma revisão do modelo de habilidades da IE. Outras perspectivas do estudo da IE proeminentes incluem, por exemplo, o modelo de traço de personalidade, encabeçado pela teoria do traço emocional de Petrides e Furnham (2001), e o modelo misto, que é diverso em sua composição, mas possui a teoria da inteligência socioemocional de Bar-On (1997) como contemporânea das teorias dos demais modelos.

Há ainda ouras formas de estudar a IE, não necessariamente com essa nomenclatura, como por exemplo pela neuropsicologia social e a cognição social. A neuropsicologia social se ocupa de construtos basilares para o desenvolvimento da IE, como empatia e teoria da mente, mas de forma mais básica, com foco na explicação neural destes fenômenos (Ward, 2017). Na cognição social, a teoria da autorregulação aborda o tema com o emprego de conceitos da psicologia cognitiva e da psicologia social (Bandura, 1991).

Neste trabalho, estes modelos não serão abordados. Ao invés disso, optou-se por adotar unicamente o modelo de habilidade como base teórica para adaptação do teste. Esta decisão não é incomum; a escolha do modelo de IE a ser adotado em um projeto de pesquisa comumente representa a perspectiva adotadas na tradição de pesquisa da psicometria: modelos de habilidade adotam testes de desempenho enquanto modelos de traço de personalidade adotam escalas de autorrelato (Bru-Luna et al., 2021).

Esta distinção é importante porque até mesmo os autores da teoria MSC concordam com a posição relativamente consensual que a IE é descrita por habilidades de raciocínio, mas está

muito próxima de aspectos de personalidade (Mayer et al., 2016). No entanto, na mensuração dos fatores da teoria MSC, a opção pelo modelo de habilidade é uma recomendação de autores de diversas perspectivas (Mayer et al., 2016; Santos et al., 2015; Schneider & McGrew, 2018).

Revisando as bases paradigmáticas da teoria MSC, Mayer et al. (2016) afirmam definitivamente que a IE é mais bem mensurada por um teste de desempenho. Schneider e McGrew (2018) fazem a mesma recomendação quando revisam o estudo do raciocínio emocional como um fator de inteligência. Em uma revisão de literatura sobre as competências socioemocionais, Santos et al. (2015) repetem essa recomendação, alertando que as estimativas de IE costumam ser enviesadas quando baseadas em escalas de autorrelato.

Tendo tido como objetivo a realização de uma limpeza conceitual, o Manuscrito 1 analisará os aspectos centrais do modelo de habilidade de IE, afastando os problemas que devem ser superados na literatura para estabelecer recomendações de pesquisa para elucidar os problemas reais desta perspectiva.

Tendo estabelecida a base teórica, o segundo manuscrito descreve o procedimento de adaptação para o contexto brasileiro e a coleta de evidências de validade dos testes STEU e STEM (MacCann & Roberts, 2008). O procedimento de adaptação de retrotradução de Brislin (1970) foi seguido, utilizando a validação por especialistas tanto da versão traduzida quanto de uma versão retraduzida para a língua original. Este procedimento é realizado para garantir confiabilidade à qualidade da adaptação. Também foram seguidas recomendações de Sireci et al. (2006) e da *International Test Commission* (2017) ao procedimento original.

Similarmente ao trabalho original de MacCann e Roberts (2008), neste trabalho foram coletadas evidências de validade com base na relação com variáveis externas. Neste paradigma, foram avaliadas as relações de convergência e discriminação (Campbell & Fiske, 1959) dos escores dos testes adaptados com outros testes. Os construtos adotados para relação discriminante incluem os cinco grandes fatores de personalidade (por meio da medida de Passos

e Laros, 2015), e, para relação convergente, uma medida de satisfação com a vida (por meio da escala de Oliveira et al., 2009) e os próprios testes, que são fatores de um mesmo construto.

O processo de adaptação em si configura coleta de evidências de validade adicional com base no conteúdo do teste (*American Educational Research Association* [AERA] et al., 2014). Na medida em que o processo decisório de aceitação de itens segue um protocolo de avaliação por especialistas, conforme recomendado por Sireci et al. (2006), o argumento pela validade dos testes perpassa também a discussão de "validade desde a concepção" (*validity by design*) de Mislevy (2007).

Desta forma, o segundo manuscrito revelará as evidências de validade dos testes adaptados, sejam aquelas produzidas no procedimento de adaptação ou no estudo empírico. Uma versão final dos testes deverá ser produzida e aplicada e, após análise estatística, deverão ser obtidos escores fatoriais e medidas de ajustes que mostrem a adequação dos testes ao modelo de IE da teoria MSC.

Além da adaptação dos testes para o contexto brasileiro, o objetivo desta tese foi avaliar se os testes adaptados poderiam ser aplicados de forma eficiente por meio da CAT sem prejuízo para validade e precisão. Esta proposta é inovadora, tendo em vista que, apesar dos benefícios já conhecidos do CAT, ainda há um número pequeno de esforços de construção de CATs para mensuração de construtos psicológicos no Brasil (Peres, 2019), que não tem nenhum teste disponível para aplicação nesta modalidade. Além de descrever as vantagens do emprego do CAT, o terceiro manuscrito teve como objetivo apontar os desafios que desenvolvedores de testes educacionais e psicológicos têm na construção de CATs e elucidar como os CATs podem ser configurados com propriedades de forma a superar estes desafios.

O manuscrito descreve as vantagens dos CATs descritas por Luecht (2016), como a redução do erro de mensuração em todos os níveis de aptidão, a redução do tempo de testagem, a melhora na segurança contra fraude na testagem, a detecção de respostas de baixa qualidade e

o uso de itens inovadores. As diferentes decisões que precisam ser tomadas sobre o nível de flexibilidade do CAT, a utilização de testlets, a escolha do critério para o algoritmo de seleção de itens (*item selection algorithm*, ISA) e da regra de parada são alguns dos aspectos discutidos detalhadamente.

Uma das vantagens mais importantes da aplicação do CAT é o efeito positivo na validade do teste (Wise, 2018). O emprego de itens inovadores e a realização de mensuração por meio de simulações para melhorar a validade ecológica é uma das formas pelas quais isso ocorre, mas também existem aspectos paralelos que têm efeito significativo. Por exemplo, Martin e Lazendic (2018) revisaram a medida em que emprego do CAT melhora a motivação e o engajamento na testagem, reduzindo a variância irrelevante ao construto no teste causada pelo tédio e desânimo dos participantes. Isso ocorre pela redução da procrastinação, que pode ocorrer pela leitura das próximas questões, entre outros motivos, e pela redução do tempo que o participante precisa dedicar ao teste (Matin & Lazendic, 2018).

Este benefício é particularmente importante considerando que o impacto do tédio e da motivação reduzida nos escores não é homogêneo entre os participantes (Balart & Oosterveen, 2019). Por exemplo, Balart e Oosterveen (2019) apresentaram evidências que o impacto destas variáveis no desempenho nos testes possui magnitude diferente dependendo do sexo do participante. Isso resulta na introdução de erro de aplicação não-sistemático (Hogan, 2007), o que ameaça a validade dos escores do teste.

Além dos aspectos envolvidos com o momento de resposta, a motivação também pode ser melhorada pela utilização de feedback imediato (Ling et al., 2017). Um estudo que comparou testes comuns a testes que revelavam imediatamente ao participante se ele havia acertado a questão mostrou que os participantes tiveram desempenho significativamente melhor nos testes com este tipo de feedback.

A validade da testagem também é beneficiada com o emprego do CAT por meio da redução do efeito da exposição. Há evidências que conhecer as tarefas e itens aplicados de antemão é prejudicial para a validade dos escores (AERA et al., 2014). Isso pode ser evitado de duas formas: primeiro, pela construção de um banco de itens grande o suficiente para que não seja prático memorizá-lo. Segundo, pela possibilidade de utilização de formas de teste diferentes. Apesar de isso ser possível na testagem linear com a construção manual de formas de teste, o CAT automatiza este processo mesmo sem qualquer planejamento (Luecht, 2016).

Estes benefícios possuem desdobramentos práticos. Por exemplo, na área de avaliação educacional, o emprego do CAT torna possível disponibilizar aos estudantes informações detalhadas sobre suas habilidades, dando feedback mais preciso ao final das atividades avaliativas; reduzir a carga de trabalho dos professores do ensino básico ao ensino superior; aumentar o engajamento dos estudantes com finalidade não só somativa, mas também formativa, entre outros (Luecht, 2016).

Na testagem psicológica, o CAT possui o potencial de impactar várias áreas de interface da psicologia. Por exemplo, na área da saúde mental, o CAT pode ser empregado para construção de um paradigma de testagem voltado para o paciente (Carlo et al., 2021). A remediação do efeito de exposição dos itens, por exemplo, permite o controle da evolução do paciente por meio da aplicação de formas de teste inéditas, que não são enviesadas pelo pré-conhecimento dos itens que ocorre quando se reaplicam testes de forma fixa. Além disso, a possibilidade de construção de escalas curtas de depressão que são precisas para o indivíduo permite o diagnóstico rápido para pessoas cuja baixa motivação disposicional poderia impedi-las de responder a escalas mais longas (Carlo et al., 2021).

Na área de recrutamento e seleção, a metodologia tem sido empregada, em pequena escala, para resolver problemas da aplicação não supervisionada de testes pela Internet

(*unsupervised Internet testing*, UI; Kantrowitz et al., 2011). Isso tem possibilitado o alcance de um número maior de candidatos, aumentando a qualidade do processo de seleção.

Apesar das vantagens serem atrativas, adaptar um teste para o formato de CAT não é uma tarefa trivial (Huff & Sireci, 2001). Para adaptar o teste para aplicação como CAT, para além de realizar uma aplicação adaptativa—real ou simulada—foi necessário refletir como oferecer subsídios para que outros pesquisadores e profissionais realizassem a aplicação dessa forma. Nesta tese, utilizou-se como critério de viabilidade de aplicação de um determinado teste como CAT a capacidade do ISA utilizado de alcançar um erro padrão (EP) que caracterize uma estimativa de IE equivalente àquela estimativa que seria feita pelo teste no formato fixo. Naturalmente, além do EP, a própria estimativa da habilidade precisaria ser comparável àquela que havia sido feita pelo teste no formato fixo.

Para cumprir este objetivo, o quarto manuscrito descreve dois estudos de simulações com os algoritmos de seleção de itens mais frequentemente descritos na literatura (Choi & Swartz, 2009). As simulações utilizaram os padrões de resposta coletados empiricamente para registrar quais itens os ISAs iriam selecionar, qual seria a estimativa do nível de habilidade a cada resposta do participante simulado, bem como o erro padrão dessa estimativa, até que se chegasse a um de dois objetivos—o que divide os grupos de simulação.

O objetivo das simulações do primeiro estudo foi alcançar a regra de parada descrita na literatura como padrão (Chalmers, 2016): a diferença de erro padrão entre a estimativa de habilidade após a resposta do penúltimo item e a estimativa de habilidade após resposta do último item de 0,001. Para o segundo estudo, o objetivo das simulações era administrar um número fixo de itens que variou entre um item até o número total de itens para o teste. As simulações foram realizadas da mesma forma para ambos os testes, variando apenas o número total de itens de cada teste, que era 32 para o STEU e 30 para o STEM. Neste sentido, o segundo estudo foi composto por simulações que foram realizadas de forma relativamente

artificial com o objetivo de coletar medidas específicas. Isso se contrasta com o primeiro estudo, que foi composto por simulações que representam uma aplicação em CAT convencional para testagem psicológica com objetivo de calcular as habilidades dos participantes.

Isso permitiu que o objetivo deste manuscrito fosse alcançado com base na avaliação de três medidas. Como o primeiro estudo ocorreu da forma como uma aplicação convencional em CAT ocorreria, as estimativas de habilidade finais dos participantes representam as estimativas que se teria nessa situação. Desta forma, a primeira medida avaliada foi a correlação entre o nível de habilidade estimado por cada ISA, calculado nas simulações, e o nível de habilidade estimado com base no teste em forma fixa, que já havia sido calculado para o segundo manuscrito.

Ainda neste estudo, como a regra de parada utilizada foi convencional, os ISAs estudados tiveram a mesma capacidade de reduzir o número de itens necessário para alcançar o nível de EP adequado que teriam em uma aplicação real do teste como CAT. Assim, a segunda medida analisada foi a quantidade de números de itens que cada ISA precisou aplicar para alcançar estes níveis de EP adequados para cada participante.

O segundo estudo estabeleceu como regra de parada uma quantidade fixa de itens, de modo que foi possível obter a terceira medida avaliada no manuscrito, os níveis de erro padrão médio atingidos por cada ISA para cada quantidade de itens. Esta medida permitiu que fossem avaliadas situações hipotéticas em que os participantes com os níveis de habilidade presentes no banco tivessem respondido menos ou mais itens, por exemplo, por terem pulado algum item, ou pelo teste especificar a aplicação de um número mínimo de itens. Além disso, esta medida, apesar de ser mais abstrata—por não dizer respeito a situações reais de aplicação—, permite que se entenda o desempenho do algoritmo para os mesmos sujeitos para os mesmos números de itens. Isto representa uma informação complementar visto que nas medidas anteriores cada participante só havia sido testado até o número de itens necessário para alcançar a regra de

parada, de forma que só havia informação do desempenho dos algoritmos para cada participante para aquele número de itens que foi aplicado.

Essa estrutura do método do quarto manuscrito é refletida nos resultados, que também são analisados medida-a-medida. Para a primeira medida, será examinado se as estimativas de habilidades realizadas com o emprego dos algoritmos foram parecidas com as estimativas de habilidade calculadas com todo o teste. Será, ainda, avaliado, se os ISAs estudados precisaram de aplicar menos itens para alcançar os níveis de EP adequados do que o algoritmo aleatório, que foi utilizado como controle negativo. Isto é particularmente importante para os algoritmos baseados na informação de Kullback-Leibler, tendo em vista que, apesar de terem sido comparados com uma situação controle, não há comparação do desempenho deste algoritmo com algoritmos que usam a informação de Fisher na literatura.

Em conjunto, estes quatro manuscritos estabelecem uma base teórica para trabalho, descrevem a adaptação de um instrumento de inteligência emocional para o contexto brasileiro e as evidências de validade coletadas que revelaram boas medidas de ajuste, explicam os benefícios e os desafios associados ao CAT, e exploram o desempenho psicométrico dos instrumentos adaptados numa administração em CAT.

MANUSCRITO 1

Teoria e Testagem do Modelo de Habilidade da Inteligência Emocional

Theory and Assessment of The Ability Model of Emotional Intelligence

Resumo

No modelo de habilidades da inteligência emocional (MHIE), a inteligência emocional (IE) é definida como um conjunto de habilidades envolvidas no reconhecimento, uso, compreensão e gerenciamento dos processos mentais de si e dos outros. O objetivo deste trabalho foi organizar a base conceitual do MHIE nacional e internacionalmente, descrevendo a evolução do construto e os desafios atuais para o desenvolvimento teórico da área. A teoria de Mayer, Salovey e Caruso (2012; teoria MSC) descreve a IE em termos de quatro fatores: percepção emocional, facilitação emocional, compreensão emocional e gerenciamento emocional. Apesar desta estrutura também ser incluída no modelo de inteligência geral de Cattell–Horn–Carroll (Schneider & McGrew, 2018), há evidências de que a facilitação emocional e o gerenciamento emocional compõem um só fator, chamado regulação emocional. Elfenbein e MacCann (2017) utilizaram dados de décadas de pesquisa fatorial com instrumentos do MHIE para propor uma reorganização das habilidades dentre os fatores da teoria MSC. Além de designar um fator específico de expressão emocional, antes presente no fator de percepção emocional, as autoras formalizaram a junção das habilidades de facilitação e gerenciamento em fatores de regulação emocional, que seriam divididos em três, de acordo com o objeto. Após análise da literatura, conclui-se que a IE é um tipo de raciocínio que existe independentemente do comportamento inteligente, pois este depende de fatores motivacionais. Superando isto, o principal desafio atual do MHIE é a falta de confiabilidade na estrutura fatorial do gerenciamento emocional e da facilitação emocional. Novas pesquisas devem se dedicar a preencher essa lacuna para investigar a relação do gerenciamento e da facilitação emocional com a regulação emocional.

Palavras-chave: inteligência emocional, regulação emocional, teoria Mayer-Salovey-Caruso, testagem psicológica, teoria Cattell–Horn–Carroll

Abstract

In the literature tradition that considers emotional intelligence to be an ability (ability EI), emotional intelligence (EI) is defined as a set of skills involved in recognizing, utilizing, understanding, and managing the mental processes of the self and of others to solve problems and regulate behavior (Mayer & Salovey, 1997). The present study aimed to organize the conceptual base of the MHIE nationally and internationally, describing the evolution of the construct and the current challenges for the theoretical development of this area of study. The Mayer–Salovey–Caruso (MSC) theory (Mayer et al., 2012) of EI describes it in terms of four factors: emotion perception, emotion facilitation, emotion understanding, and emotion management. Although this structure is also included in the Cattell–Horn–Carroll model of general intelligence (Schneider & McGrew, 2018), there is evidence that emotion facilitation and emotion management make up a single factor, called emotion regulation. Elfenbein and MacCann (2017) used data from decades of factorial research with ability EI instruments to propose a reorganization of skills among the factors of the MSC theory. In addition to designating a specific emotion expression factor, previously merged with the emotion perception factor, the authors joined the facilitation and management abilities into three emotional regulation factors, classified according to their object. The result of the study reveals that EI, both because of its ontological nature and because of the MSC theory, is a type of reasoning that should be understood separately from intelligent behavior. Overcoming this, the major difficulty of the EI ability model is the unreliability in the factor structure of emotional management and emotional facilitation. New research should be dedicated to investigating the relationship between emotional management and facilitation and emotional regulation.

Keywords: ability emotional intelligence, emotion regulation, Mayer-Salovey-Caruso theory, psychological testing, Cattell-Horn-Carroll theory

Teoria e Testagem do Modelo de Habilidades da Inteligência Emocional

O modelo da inteligência emocional (IE) como habilidade é um paradigma que afirma que este construto é melhor entendido como a habilidade de raciocinar de forma válida com as emoções e com informações relacionadas às emoções, de resolver problemas com conteúdo emocional, e de utilizar as emoções para estimular o pensamento (Mayer et al., 2016). As habilidades subjacentes ao construto incluem a percepção da emoção nas mais variadas situações da vida real e da arte, o entendimento dos antecedentes e das consequências das emoções, o gerenciamento das próprias emoções e das emoções dos outros, e a utilização das emoções para regular o próprio pensamento (Mayer et al., 2016).

A teoria Mayer–Salovey–Caruso (MSC) de IE, inicialmente encabeçada por Mayer e Salovey (1997), se destacou ao longo do tempo como a teoria fatorial mais influente no estudo da IE por um modelo de habilidade (*ability EI*; Gonzaga & Monteiro, 2011; Vieira-Santos et al., 2018). Os autores consideram que a emoção é uma resposta mental organizada que inclui aspectos fisiológicos, experienciais e cognitivos (Salovey & Mayer, 1990). Tendo isso em vista, o modelo da teoria conta com quatro fatores que se desenvolvem de forma cumulativa e interconectada, razão pela qual foram chamados ramos (Mayer et al., 2016). Estes fatores, que foram organizados pelos autores na ordem em que são desenvolvidos, foram inicialmente descritos com os nomes percebendo emoções (*p*erceiving emotions*),* facilitando pensamentos usando emoções (*facilitating thoughts using emotions*)*,* compreendendo emoções (*understanding emotions)* e gerenciando emoções (*managing emotions;* Mayer & Salovey, 1997).

O desenvolvimento da teoria MSC ocorreu paralelamente à construção do instrumento de mensuração do construto, a Escala Multifatorial de Inteligência Emocional (*Multifactor Emotional Intelligence Scale,* MEIS; Mayer & Salovey, 1997). No entanto, o estudo de apresentação de evidências de validade da MEIS revelou dificuldades na separação dos dois

últimos fatores, que foram interpretados em um estudo empírico de Roberts et al. (2001) como um único fator de regulação emocional. A MEIS foi sucedida pelo *Mayer–Salovey–Caruso Emotional Intelligence Test* (MSCEIT; Mayer et al., 2002), no qual se insistiu em manter escores para os quatro fatores teóricos, decisão que foi eventualmente argumentada com base em evidências apresentadas por Brackett e Salovey (2006). Mesmo assim, esta decisão foi controversa, visto que ainda há estudos que sugerem que a separação da regulação emocional em gerenciamento emocional e facilitação emocional é potencialmente inviável (Fan et al., 2010).

Em 2012, Mayer et al. renomearam os quatro ramos da teoria MSC para percepção emocional, facilitação emocional, compreensão emocional e gerenciamento emocional. Os autores descreveram as habilidades específicas associadas a cada um destes fatores. Segundo eles, a percepção emocional seria a capacidade de:

- Identificar expressões emocionais honestas e desonestas.
- Discriminar expressões emocionais precisas de imprecisas, como, por exemplo, em uma montagem criada por um computador.
- Entender como as emoções são expressas dependendo do contexto e da cultura.
- Expressar emoções de forma correta quando elas forem consideradas pertinentes para a situação.
- Perceber conteúdo emocional no ambiente, nas artes visuais e na música.
- Perceber emoções nos outros por meio de pistas na voz, na expressão facial, na linguagem e no comportamento.
- Identificar emoções nos próprios estados, as emoções ligadas aos próprios sentimentos e pensamentos.

A segunda dimensão, a facilitação emocional, seria a capacidade de:

- Priorizar a solução de problemas para os quais o estado emocional em curso pode facilitar a cognição.
- Utilizar as mudanças do próprio humor para gerar diferentes perspectivas cognitivas.
- Priorizar o pensamento direcionando a atenção de acordo com o sentimento presente.
- Gerar emoções com o intuito de facilitar o próprio julgamento e a própria memória.
- Gerar emoções como meio de se relacionar com experiências de outras pessoas.

Ao passo que o primeiro e o segundo ramo são habilidades que se desenvolveriam até mesmo de forma instintiva, o terceiro ramo, compreensão emocional, passa a exigir consciência dos aspectos emocionais de si e dos outros. Este ramo envolveria:

- Reconhecer as diferenças culturais na avaliação das emoções.
- Compreender como uma pessoa pode se sentir em decorrência de certas situações.
- Reconhecer e prever a evolução das emoções por meio das transições entre emoções mais prováveis.
- Compreender emoções complexas e mistas.
- Diferenciar entre humor e emoções.
- Avaliar os aspectos das situações para identificar o seu potencial de provocar emoções.
- Determinar os antecedentes, significados, e consequências das emoções.
- Nomear as emoções e reconhecer as relações entre elas.

O quarto e último ramo, gerenciamento emocional, seria o mais complexo. Os comportamentos componentes deste ramo incluem:

- Gerenciar efetivamente as emoções próprias ou alheias para alcançar um resultado desejado.
- Avaliar estratégias para manter, reduzir, ou intensificar uma resposta emocional.
- Monitorar as reações emocionais para determinar sua razoabilidade.
- Engajar-se com as emoções se elas forem úteis e afastá-las se não.
- Ficar aberto a sentimentos agradáveis e desagradáveis conforme necessário e às informações que eles transmitem.
- Gerar emoções como auxílio ao julgamento e à memória.

Além de listar as habilidades associadas à IE, os autores também apresentaram evidências do aspecto de desenvolvimento e complexidade cumulativos dos fatores com base em dados coletados com o MSCEIT (Cabello et al., 2016). Em um estudo transversal, Cabello et al. (2016) mostraram que o primeiro ramo é o primeiro a ser desenvolvido no começo da vida e é o que melhor se conserva ao final da vida. Os demais ramos, no entanto, são desenvolvidos até a meia-idade. A pesquisa mostrou, ainda, a possibilidade das habilidades emocionais se degenerarem em idades mais avançadas. Outras habilidades de raciocínio que compõem a inteligência geral também possuem esta caraterística (Mayer et al., 2000).

Este argumento introduz uma das grandes contribuições da teoria MSC, que propõe que a IE é um raciocínio intelectual como qualquer outro (Mayer et al., 2001). Para defender esta tese, Mayer et al. (2000) apresentaram um argumento em duas etapas. Primeiro, eles propuseram três critérios para a aceitação de um construto em um modelo de inteligência geral. Em seguida, argumentaram que a IE atende a estes critérios. Para ser reconhecido como um raciocínio, o construto ter o potencial de ser operacionalizado em uma série de itens, deveria se correlacionar com outros tipos de inteligência moderadamente—mas ter variância específica—, e deveria evoluir ao longo do desenvolvimento natural do ser humano.

Os autores demonstraram, com base em dados da escala MEIS, que a IE como habilidade cumpria estes requisitos. No entanto, Roberts et al. (2001), respondendo a Mayer et al. (2000), questionaram a validade do MEIS como operacionalização da IE. Embora tenham reconhecido que as evidências de validade convergente e discriminante pareciam convincentes, os autores argumentaram que os métodos de correção levavam a resultados contraditórios. Este mesmo estudo foi responsável por indicar a inadequação da estrutura de quatro fatores e propor a fusão dos fatores de facilitação emocional e gerenciamento emocional em um único fator de regulação emocional. Em conjunto, estas críticas desempenharam um papel importante na contraposição ao argumento de introdução da IE como habilidade no modelo de inteligência geral.

Apesar do reconhecimento de que o construto tinha se maturado com MSCEIT, Maul (2012) renovou as críticas anteriormente dirigidas ao MEIS ao novo instrumento. Após analisar os resultados de uma análise fatorial confirmatória, Fan et al. (2010) havia sugerido que a melhor estrutura para o MSCEIT seria de três fatores, espelhando a sugestão que Roberts et al. (2001) havia feito ao MEIS. Relatando os achados de Fan et al. (2010), Maul (2012) ainda criticou baixos índices de fidedignidade e a falta de um procedimento de construção de itens sistemático baseado numa teoria consolidada de emoções.

Em resposta a essas críticas, Mayer et al. (2012) apresentaram novas evidências de validade baseadas na reanálise do banco de normatização do teste, que possuía 5000 participantes, e incluíram estimativas de fidedignidade atualizadas e relatos de quatro outros estudos com evidências de validade com base na estrutura interna do teste e na relação com variáveis externas. Posteriormente, para responder às críticas acerca da pertinência da IE como habilidade de raciocínio intelectual, Mayer et al. (2016) publicaram um tratado que teve como objetivo reformular os aspectos teórico-metodológicos inicialmente propostos em 1990 e 1997.

Princípios do Modelo de Habilidades da Inteligência Emocional

No tratado de Mayer et al. (2016), os autores buscaram esclarecer os conceitos basilares da teoria, incluindo a especificação de aspectos comportamentais e ontológicos e consolidando argumentos e respostas a questionamentos passados, como os de Maul (2012). Os autores propuseram sete princípios do modelo de habilidades da IE, que evoluíram de básicos a mais complexos e que são condições para o entendimento do construto.

1. A IE é uma habilidade mental. Se a inteligência, em termos gerais, é a capacidade de executar o pensamento abstrato, entender sentidos, similaridades, e generalizações, então a IE deve ser considerada uma inteligência, pois circunscreve essas mesmas habilidades no contexto emocional individual e interpessoal (Brand, 1985).
2. A IE é melhor medida como uma habilidade. Como qualquer outra habilidade, a melhor forma de avaliar a IE seria não por meio de escalas de autorrelato, mas por meio de testes objetivos de desempenho com respostas certas e erradas decididas por especialistas (Schneider & McGrew, 2018) ou por consenso de pares (Mayer et al., 2016). A metodologia de escalas de autorrelato consistentemente leva a estimativas exageradas da inteligência dos sujeitos (Paulhus et al., 1998).
3. A inteligência é uma capacidade mental que não pode ser reduzida aos comportamentos destinados à resolução de problemas (Mayer et al., 2016). A capacidade de solucionar problemas efetivamente não é a mesma coisa que estar disposto a executar o comportamento inteligente. No contexto da IE, isto é ainda mais relevante tendo em vista que em situações sociais os indivíduos produzem comportamentos resultantes parcialmente das suas habilidades de IE mas também de uma série de aspectos situacionais e disposicionais que não têm correlação com a IE (DeYoung, 2011). Em termos práticos, isso significa que é possível ter IE alta e não resolver de forma eficaz os problemas emocionais da própria vida.

4. O conteúdo de um teste deve ser claramente especificado como condição para a mensuração de habilidades mentais humanas. Testes psicológicos só podem ser classificados como de inteligência e só produzem escores válidos quando definem de forma precisa seu conteúdo (AERA et al., 2014).
5. Testes com evidências de validade favoráveis possuem conteúdos bem definidos que evocam habilidades mentais humanas. Consequentemente, o conteúdo do teste deve ser circunscrito dentro do conjunto de conteúdos emocionais e a resposta aos itens deve evocar as habilidades relevantes.
6. A IE é uma forma ampla de inteligência: assim como defendido nos artigos de MacCann et al. (2014) e Evans et al. (2019), estudos fatoriais defendem a inclusão da IE como uma habilidade de segundo estrato, isto é, subserviente ao fator geral de inteligência, mas composta de variância independente do fator *g* e das demais habilidades amplas (Schneider & McGrew, 2018).
7. A IE é focada no processamento de informações "quentes". Apesar do modelo CHC não trazer explicitamente a diferenciação entre habilidades *hot* (quentes) e *cold* (frias), Brand (1985) já havia teorizado a possível existência dessa divisão. As inteligências frias seriam aquelas que lidam com conteúdo que evocam respostas fundamentalmente apáticas, enquanto as "quentes" tratariam de objetos com os quais o indivíduo de fato se investe (Mayer et al., 2016).

Estes princípios resumiram respostas a questionamentos levantados ao longo da história do MEIS e do MSCEIT. Por exemplo, críticas de autores como Petrides e Furnham (2001) expressavam a preocupação de que não seria possível construir itens que tivessem repostas certas ou erradas com base no princípio de que o desenvolvimento emocional é individual. Os autores defendiam que não faz sentido dizer que um determinado padrão de respostas aprendido nativamente por um indivíduo no decorrer de seu desenvolvimento é errado. No entanto, o

trabalho de Mayer et al. (2016) defendeu que nas definições de inteligência as habilidades mentais são subservientes a um propósito utilitário. No caso da inteligência emocional, o propósito utilitário que deve servir como parâmetro para distinguir respostas corretas de respostas incorretas é a comunicação efetiva das emoções. Neste sentido, assim como não se faz juízo de valor quando se diz que uma questão de matemática está certa ou errada independentemente da forma como o aluno aprendeu a resposta que deu, definir um gabarito para um teste de desempenho de IE também não é realizar um juízo de valor (Mayer et al., 2016). O importante, tanto para a prova de matemática quanto para o teste de IE, é que o critério de correção esteja claro.

Frente a mais críticas sobre o cálculo dos escores, os autores também esclareceram que o MSCEIT já havia resolvido problemas que ainda eram alvos de críticas que só eram pertinentes na época do MEIS (Mayer et al., 2016). Por exemplo, antecipando a crítica que a definição do gabarito do MEIS por meio do consenso dos participantes com mais IE configurava um argumento circular—isto é, que não era justificável utilizar a resposta de um grupo com alta IE como parâmetro de seleção das respostas corretas antes de haver um instrumento independente que pudesse garantir que os membros do grupo tinham alta IE—os autores empregaram um gabarito alternativo construído por especialistas na área de pesquisa de emoções. Dentre os itens construídos no estudo de construção do MSCEIT (Mayer et al., 2003), só foram adicionados aqueles sobre os quais havia concordância entre o gabarito dos especialistas e o gabarito montado pelas respostas dos indivíduos com alta IE (Mayer et al., 2016).

Por fim, os autores reafirmaram a resposta a outro questionamento acerca da pertinência do modelo de habilidade da IE ser incluído em teorias de inteligência (Mayer et al., 2012). Husin et al. (2013) inquiria se a IE já estaria contemplada pela inteligência cristalizada, de forma que não seria necessário propor um fator específico para a IE.

A inteligência cristalizada é um dos dois tipos de raciocínio da teoria *Gf–Gc* de Cattell

(1963). Neste modelo admite-se apenas a inteligência fluida (*Gf*), que daria conta de todo o raciocínio executado frente a situações desconhecidas e a inteligência cristalizada (*Gc*), utilizada na resolução de problemas sobre os quais há conhecimento prévio.

Na crítica de Husin et al. (2013), o raciocínio em IE seria orientado pela inteligência cristalizada com base no conhecimento prévio de emoções. No entanto, Mayer et al. (2012) apontaram que esta sugestão apenas fazia sentido no contexto do modelo *Gf–Gc* de Cattell (1963). Se a inteligência tivesse que ser restrita aos fatores *Gf* e *Gc*, a IE, de fato, seria melhor entendida como parte da *Gc*. Mas quando se passou a admitir a existência de outros fatores, a inclusão de um fator de inteligência seria sim pertinente quando fosse identificado um raciocínio que, além de ter correlação positiva com os demais raciocínios, possuísse conteúdo e processos específicos não associados a estes outros raciocínios. Para Mayer et al. (2012), especialmente quando se considerava os fatores de percepção e regulação emocional, a IE cumpria este requisito tão bem quanto os demais fatores que Horn (1988) propôs com a teoria *Gf–Gc* expandida, que adicionava raciocínios como raciocínio verbal ou abstrato, e era igualmente digna de um fator independente.

Além disso, já havia precedentes para o reconhecimento dos aspectos socioemocionais do raciocínio. Mesmo no início da psicometria Thorndike (1920) previu que a disciplina iria eventualmente reconhecer um fator de inteligência social com grande sobreposição ao que seria conhecido hoje como IE, mas não chegou a adicioná-lo no seu modelo. Na literatura mais recente, quando Carroll (1993) baseou-se no modelo de Horn (1988) para construir o seu modelo hierárquico que posicionava habilidades de raciocínio específicas abaixo do fator geral de inteligência—o modelo de três estratos—, o autor previu a existência de um fator geral socioemocional. No final, Carroll (1993) não adicionou o fator ao seu modelo (Schneider & McGrew, 2012). Mais tarde, McGrew e Evans (2004) expandiram os fatores de Horn (1988) e Carroll (1993) na teoria que ficou conhecida como Cattell–Horn–Carroll (CHC), mas ainda

julgaram que o construto ainda não havia maturado o bastante para figurar no modelo.

Mesmo assim, os autores defenderam que o CHC seria um modelo flexível que se atualizaria com os desenvolvimentos da literatura, visto que, da mesma forma que diversas habilidades novas haviam sido reconhecidas desde Horn (1988), provavelmente ainda havia habilidades de raciocínio legítimas a serem reconhecidas (Schneider & McGrew, 2012).

A demanda pela inclusão da IE nos modelos de inteligência geral eclodiu com a publicação *Emotional intelligence as a second stratum factor of intelligence* (MacCann et al., 2014). Neste trabalho, análises de modelos unidimensionais, oblíquos, hierárquicos e bifatoriais baseados numa amostra de 688 alunos universitários sugeriram que a IE seria bem integrada ao modelo de inteligência como uma habilidade ampla subjacente ao fator geral de inteligência e composta pelos quatro fatores do MSCEIT.

A atualização da teoria CHC realizada em 2018 finalmente reconheceu a IE como um dos tipos de raciocínio (Schneider & McGrew, 2018). Apesar de especificar que a inclusão é "probatória", os autores aceitaram a contribuição da teoria de quatro fatores de Mayer et al. (2016), apesar de terem adotado definições constitutivas próprias da IE e de seus quatro fatores. De acordo com Schneider e McGrew (2018), a percepção emocional é a habilidade de reconhecer emoções com precisão no ambiente, nas faces, na voz e no comportamento, enquanto a compreensão emocional é o conhecimento sobre os antecedentes da emoção e as consequências da expressão emocional. Estas duas formam o componente majoritário da IE, sendo definidas como "habilidades principais". Esta classificação diz respeito não a uma precedência teórica, mas sim estatística, sendo conferida aos fatores cuja confiabilidade estatística é maior. A lógica inversa se aplica na classificação dos dois fatores remanescentes como "habilidades menores". O terceiro fator é o gerenciamento de emoções, que é a capacidade de regular as emoções de forma deliberada e adaptativa. Por fim, a utilização emocional é a capacidade de fazer uso adaptativo das emoções, especialmente para facilitar o

raciocínio.

Novos desenvolvimentos desde então evidenciaram a correção dessa decisão. Numa replicação do trabalho de MacCann et al. (2014), Evans et al. (2019) adicionaram ao crescente corpo de evidências (e.g., Fernández-Berrocal et al., 2017; MacCann, 2010, 2012; MacCann et al., 2014) que o modelo da teoria MSC teria mérito para inserção no modelo da teoria CHC. Os autores realizaram uma análise fatorial confirmatória num banco de dados com resposta de 830 indivíduos. Os resultados confirmaram a adequação de um modelo hierárquico de inteligência com três fatores, que chamaram de percepção, compreensão e gerenciamento emocional. No ano seguinte, Bryan e Mayer (2020) realizaram uma meta-análise, utilizando uma série de estudos acerca da relação entre os tipos de inteligência, que mostrou que a correlação média entre a IE e a inteligência geral foi de 0,58, na mesma faixa de correlações dos demais raciocínios amplos. Segundo os autores, este resultado evidenciou a similaridade da IE aos demais tipos de raciocínio e reforçou a adequação da inserção da inteligência emocional na teoria CHC.

A limitação demonstrada em diversos trabalhos sobre a adequação dos fatores de gerenciamento e de utilização emocional, que perdura desde as primeiras críticas da escala MEIS (Roberts et al., 2001), ainda é alvo de questionamentos. Apesar de replicar as evidências favoráveis a IE como uma inteligência, o trabalho de Evans et al. (2019) também replicou a recomendação do modelo de três fatores (Roberts et al., 2001), e o caráter probatório da inserção da IE no modelo CHC tem relação direta com esta dificuldade (Schneider & McGrew, 2018). Os próprios autores da teoria, agora, reconhecem a importância do conteúdo destes fatores para a descrição da inteligência emocional em si, mas admitem que as evidências disponíveis na literatura mostram que a estrutura fatorial da IE pode ser diferente daquela que eles apresentaram (Mayer et al., 2016). Há demanda para novos trabalhos que desenvolvam esta temática.

Inteligência Emocional e Regulação Emocional

Uma publicação recente que teve como objetivo desenvolver a teoria MSC e elucidar as dúvidas acerca da dimensionalidade foi publicado por Elfenbein e MacCann (2017). As autoras apresentam um modelo de IE que faz algumas modificações, incrementações, e remoções à teoria MSC com base no conhecimento acumulado desde a publicação do modelo da teoria. Apesar de não ter sido amplamente utilizado, o trabalho delas reflete em grande medida os desenvolvimentos recentes da literatura de IE como habilidade.

O modelo é organizado da seguinte forma: além de dois fatores equivalentes aos de Mayer et al. (2012), percepção emocional e compreensão emocional, as autoras propuseram a divisão do gerenciamento emocional em três fatores (Elfenbein & MacCann, 2017). Estes fatores também acumulariam os comportamentos associados à facilitação emocional. Uma comparação entre os quatro fatores da teoria MSC e os seis fatores de Elfenbein e MacCann (2017) pode ser visualizada na Tabela 1.

Tabela 1

Dimensionalidade das Teorias de Mayer et al. (2012) e de Elfenbein e MacCann (2017).

Modelo Mayer–Salovey–Caruso	Modelo Elfenbein–MacCann
Compreensão Emocional	Compreensão Emocional
Percepção Emocional	Percepção Emocional
	Expressão Emocional
Gerenciamento Emocional Facilitação Emocional	Regulação Emocional de Si
	Regulação Emocional dos Outros
	Regulação Emocional da Atenção

A reorganização ocorreu de duas formas: primeiro, conforme recomendações dos estudos fatoriais do MSCEIT, os aspectos teóricos e comportamentais do gerenciamento e da facilitação emocional foram unificados (Roberts et al., 2001) no construto de regulação emocional. Desde a recomendação inicial de Roberts et al. (2001), o crescimento da tradição de pesquisa e a consolidação deste construto (McRae, 2016) influenciou esta alteração.

Além disso, estudos fatoriais revelaram que o fator pode ser dividido em três,

organizado no fator de *regulação emocional de si*, *dos outros*, ou *da atenção*, dependendo se o a regulação realizada é das próprias emoções, das emoções alheias, ou é das próprias emoções, mas especificamente realizado para facilitar o controle da própria atenção. Por fim, do fator de percepção emocional, que antes incluía a expressão emocional (Mayer et al., 2012), foi extraído um novo fator específico desta habilidade com o nome *expressão emocional* (Elfenbein & MacCann, 2017). Esta é a inclusão mais inovadora do modelo. Segundo as autoras, a expressão emocional é posterior à emoção e diz respeito a forma como um determinado indivíduo comunica as emoções que sente ou finge sentir, geralmente pelo uso de expressões faciais e linguagem, seja verbal ou não-verbal. A Tabela 2 compara os modelos.

Apesar das diferenças em dimensionalidade, as autoras não rejeitam o corpo teórico da teoria MSC (Mayer et al., 2012). Na verdade, foram as mesmas autoras responsáveis pela recomendação de introdução do modelo de quatro fatores dessa teoria ao modelo CHC (MacCann et al., 2014). Além disso, o modelo das autoras consta exclusivamente com as habilidades descritas por Mayer et al. (2016), e apenas as reenquadra de forma a responder aos estudos fatoriais que vêm sendo realizados.

Dentre as mudanças realizadas, a nomenclatura utilizada por Elfenbein e MacCann (2017) inclui o termo "regulação emocional", que por si só tem sido objeto de interesse em um número crescente de estudos (McRae, 2016). A regulação emocional é definida como o processo de utilização de ferramentas do executivo central—particularmente o controle inibitório—, com o objetivo de gerenciar as emoções que são sentidas. Apesar do termo independer do objetivo, é geralmente aceito que seres humanos tendem a utilizá-la com o intuito de manter as emoções positivas, e interromper as negativas. Na teoria MSC, ela é representada nos fatores de gerenciamento emocional e facilitamento emocional do pensamento, que não só possuem definições constitutivas e operacionais em termos de regulação emocional (Mayer et al., 2016), como são tratados, conjuntamente, como um só fator de regulação emocional em

diversos estudos (Evans et al., 2019; Fan et al., 2010).

Apesar dos conceitos serem equivalentes, as tradições de pesquisa são reconhecidamente diferentes. Em uma revisão de literatura, Peña-Sarrionandia et al. (2015) buscaram estudos escritos em inglês que relacionavam IE e estratégias de regulação emocional em adultos saudáveis. Ao comparar as estratégias de pesquisa utilizadas, os autores sugeriram que as tradições de pesquisa seriam beneficiadas com uma integração teórico-metodológica, tendo em vista que isso resultaria em arsenal de métodos e técnicas mais completo. Isto porque, apesar do termo "regulação emocional" ser usado nas duas tradições de pesquisa, a pesquisa em regulação emocional tradicionalmente foca nos detalhes da regulação como um processo psicológico básico (Gross, 1998), ao passo que a pesquisa em IE utiliza técnicas fatoriais para focar na mensuração do construto, com um menor foco em explicar os processos, com poucas exceções (e.g., Joseph & Newman, 2010).

Peña-Sarrionandia et al. (2015) ainda tiveram como objetivo revisar artigos que relacionaram as variáveis por meio de definições e medidas diferentes, partindo do modelo de processo da regulação emocional para avaliar a literatura de IE. Eles ainda apontaram duas conclusões: primeiro, que indivíduos com alta IE regulam suas emoções desde o início da trajetória emocional e têm muitas estratégias à sua disposição. E, segundo que estes mesmos indivíduos têm sucesso quando regulam suas emoções, mas o fazem de forma flexível, isto é, quando necessário e sem impedir que outras emoções surjam.

Modelo de Habilidade da Inteligência Emocional no Brasil

Uma caracterização da pesquisa sobre IE no Brasil foi realizada por Gonzaga e Monteiro (2011). Os autores listam todos os artigos produzidos até a publicação acerca dos temas de IE e emoção, bem como os instrumentos psicométricos utilizados nos estudos. Dos 19 trabalhos identificados, 13 mensuraram a IE com um teste baseado na teoria MSC.

Diversos trabalhos dedicaram-se ao estudo de um dos fatores mais estatisticamente

relevantes da teoria MSC, a percepção emocional. Miguel et al. (2016a) e Zuanazzi et al. (2015), por exemplo, estudaram o funcionamento de testes de desempenho de percepção emocional. Miguel et al. (2017) investigaram os aspectos de percepção emocional pelo método de Rorschach.

Também foram utilizadas em estudos brasileiros medidas de IE baseadas em outros modelos de IE, como a IE como traço de personalidade (*EI trait*) de Petrides e Furnham (2000; 2001). Recentemente, Zuanazzi et al. (2022) apresentaram evidências de validade do Questionário do Traço Emocional (originalmente *Trait Emotional Questionnaire* [TEIQue]; Petrides & Furnham, 2001).

Apesar de alguns autores, incluindo Petrides e Furnham (2000; 2001), terem inicialmente defendido o modelo da IE exclusivamente como traços de personalidade, o consenso na literatura internacional é que os modelos devem ser vistos de forma complementar (Ciarrochi et al., 2000). Até mesmo Mayer et al. (2016) apresentam uma visão similar, defendendo a utilização dos testes para a mensuração das habilidades intelectuais do IE, mas admitindo que a complexidade do construto significa que há outros fatores envolvidos, se não na IE, pelo menos na tomada de decisão de executar um comportamento tipicamente considerado emocionalmente inteligente. Internacionalmente, uma visão que também adota ambas as perspectivas, chamada IE mista (*mixed EI*), é encabeçada por Bar-On (2000; 2006), e utiliza majoritariamente o Teste Informatizado do Quociente Emocional (*EQ-i*; Bar-On, 2004).

A perspectiva mista é majoritária também no Brasil, mas é relativamente independente. A literatura de competências socioemocionais adota perspectivas dos modelos de traço e de habilidade simultaneamente (Santos et al., 2015), e se interessa não só no raciocínio emocional isoladamente, mas também nos mecanismos que motivam o comportamento emocionalmente inteligente.

Primi et al. (2021a), por exemplo, descreveram as evidências de validade do inventário

SENNA (Primi et al., 2021b), que conta com avaliação tanto de aspectos de habilidades quanto de traços que incluem, mas não são limitados ao construto de IE. Os autores descrevem a mensuração de 18 habilidades, além das escalas de traço de personalidade que se agrupam em fatores chamados de “cinco grandes fatores socioemocionais” (Primi et al., 2021a).

A regulação emocional também tem tido atenção crescente em estudos no Brasil. As versões completa e reduzida da Escala de Dificuldades em Regulação Emocional foram adaptadas em um estudo que encontrou parâmetros de validade convergente e fidedignidade adequados, exceto para um dos fatores (Miguel et al., 2016b). Em outro estudo, aspectos de regulação emocional foram estudados em testes que utilizam técnicas projetivas, como o Zulliger e o Pfister (Miguel et al., 2017).

Rede Nomológica da Inteligência Emocional como Variável Preditora

Desde o início dos estudos de construtos por meio da psicometria, mas principalmente a partir das contribuições de Cronbach e Meehl (1955), a rede nomológica consiste numa fonte de informação importante acerca dos construtos. A rede nomológica compõe-se do resultado de estudos empíricos que estendem o entendimento do construto por meio da expressão quantitativa das relações com construtos da mesma área de estudo, ou de diferentes áreas.

O interesse acerca do impacto, isto é, da validade preditiva da IE no desempenho no trabalho (Fahr et al., 2012), no sucesso nas organizações (Gregory & Levy, 2011), nos comportamentos estudados pela psicologia social (Kross & Grossman, 2012) e até mesmo na personalidade (Mayer et al., 2008) sempre representou a maior parte das pesquisas realizadas acerca do construto (Vieira-Santos et al., 2018). Recentes avanços na abordagem psicométrica da IE permitiram a realização de diversos estudos do poder preditivo do construto, inclusive revisões sistemáticas e meta-análises sobre o tema (e.g., Mayer et al., 2008; Schneider & McGrew, 2018; Van Rooy & Viswesvaran, 2003).

No *Annual Review of Psychology* de 2008, Mayer et al. foram convidados a explicar o que a inteligência emocional é capaz de prever. Segundo os autores, a maior capacidade preditiva da IE está associada com a qualidade das relações desenvolvidas, desde a infância até a vida adulta. Na adolescência, estudantes com melhores escores na MSCEIT foram mais frequentemente classificados como amigos pelos seus pares (Mestre et al., 2006). Na adultícia, indivíduos com alta IE como medida pelo MSCEIT, mas não por uma medida de autorrelato, mais frequentemente se importavam com os eventos de vida das pessoas que identificavam como amigos, e eram menos críticos em relação ao sucesso deles.

Considerando o impacto da qualidade das relações na satisfação com a vida, não surpreende que a IE também esteja associada ao bem-estar geral—com o qual Brackett e Mayer (2003), e depois Brackett et al. (2006) encontraram correlações de r = 0,16 a 0,28. A qualidade das relações com os irmãos e os pais, na infância, e com os filhos e amigos, na juventude e adultícia, são considerados os melhores preditores de bem-estar (Chanfreau et al., 2013). Este tipo de confirmação das hipóteses formuladas por meio do entendimento da relação entre os construtos na rede nomológica caracteriza a evidência de validade com base na relação com variáveis externas da mensuração da EI pelo MSCEIT.

Sánchez-Álvarez et al. (2016), em uma meta-análise, também encontraram evidências de que a EI está associada com a qualidade de vida subjetivamente avaliada, apesar dos resultados terem revelado que medida de IE de autorrelato haviam obtido maiores correlações do que as medidas de desempenho de IE. Além disso, uma revisão demonstrou que a depressão é negativamente correlacionada com medidas de IE de desempenho (Fernández-Berrocal & Extremera, 2016). Neste último estudo, medidas de autorrelato não foram testadas.

Brackett et al. (2006), em um estudo com o MSCEIT, mostraram que indivíduos que se identificaram no sexo feminino tendem a ter escores mais altos em três dos quatro ramos mensurados, apesar de não ter sido essa a pergunta de pesquisa dos autores. No entanto, Mayer

et al. (2012) sugerem cautela ao interpretar os escores dos quatro ramos, visto que possuem fidedignidade mais baixa do que o desejável; apenas o escore final alcança esse quesito. De qualquer forma, esta diferença também foi estatisticamente significativa no escore total, t (284) = 5,80, η^2 = 0,105, p < 0,001 (Brackett et al., 2006).

Além disso, um estudo que teve como objetivo avaliar se havia diferenças de IE entre sexos confirmou as observações destes autores (Cabello et al., 2016). No entanto, esta diferença foi marcadamente menor, e foi demonstrada nas quatro dimensões da teoria MSC.

A compreensão emocional em crianças de cinco anos, conforme mensurado por uma tarefa de reconhecimento de emoções e uma tarefa de nomeação de emoções, mediou a relação entre conhecimento verbal e o desempenho escolar, r = 0,43 (Izard et al., 2001). Um estudo brasileiro com o Teste de Conhecimento de Emoções de Denham et al. (1990) revelou que o desempenho acadêmico de crianças com idade média de 6 anos (*DP* = 0,5) também estava associado à percepção emocional, o primeiro ramo da teoria MSC.

Os efeitos positivos na vida social se estendem à faculdade, à universidade e ao trabalho. Em contrapartida à relação do bem-estar subjetivamente avaliado, que é mais bem explicado pela IE de autorrelato, a existência de habilidades emocionais comprovada por testes de desempenho é pré-requisito para o bom desempenho na faculdade. Em uma meta-análise, MacCann et al. (2020) procuraram distinguir o efeito das medidas de IE serem de autorrelato ou de desempenho, e encontraram que o autorrelato apenas predizia o desempenho acadêmico no nível de ρ = 0,12, ao passo que a medida com base na habilidade prediz o desempenho em ρ = 0,20. Considerando a variância que é devida ao complexo processo de aprendizagem dos alunos, esta magnitude de predição não pode ser considerada baixa; ao contrário, consistiu no terceiro melhor preditor do desempenho na academia, depois da inteligência geral e da conscienciosidade (MacCann et al., 2020).

Os autores sugeriram três mecanismos que seriam responsáveis por este efeito: a regulação das emoções associadas à academia, a construção de relações sociais na universidade, e até mesmo o conteúdo associado à IE (MacCann et al., 2020). Este último mecanismo está associado a um outro achado do estudo: a IE foi melhor preditora do desempenho nas disciplinas humanas (artes, literatura, história etc.) do que nas ciências (exatas, biológicas, da terra, social e psicológica).

Uma categoria de estudos acerca do impacto da IE na academia e no trabalho utiliza o desempenho dos participantes em tarefas pré-determinadas pelos pesquisadores como parâmetro para avaliar habilidades que seriam importantes para o desenvolvimento pessoal nestes ambientes. Mayer et al. (2008) afirmam haver evidências de correlação entre a IE e o desempenho nessas tarefas. Um estudo com 203 participantes examinou a relação entre uma medida de desempenho e uma medida de autorrelato, e tarefas quentes ou frias que são tradicionalmente utilizadas para mensurar a memória de trabalho (Gutierrez-Cobo et al., 2017). Os pesquisadores encontraram evidências de que os participantes com alta IE medida pelo teste de desempenho, mas não os demais, tiveram escores significativamente melhores na tarefa quente.

Três meta-análises abordaram a predição de desempenho do IE nas organizações, com entre oito e dez estudos cada (Côté, 2014). Duas revelaram correlações médias de 0,16 (Joseph & Newton, 2010) e 0,21 (O'Boyle et al., 2011) entre alguma das quatro facetas de IE mensuradas pelo MSCEIT e variáveis de desempenho no trabalho. A outra, mais antiga, incluiu 57 estudos, encontrando uma correlação média com os escores do MEIS de 0,2 (Van Rooy & Viswesvaran, 2004).

Mais recentemente, um estudo realizado por Ashkanazy e Daus (2020) teve como objetivo fazer uma revisão dos estudos que abordaram a relação entre IE e o ambiente de trabalho. Os autores classificaram os estudos por tipo de instrumento, estabelecendo três

categorias. A primeira diz respeito ao uso dos testes de Mayer, Salovey e Caruso, ou seja, o MSCEIT ou o MEIS. A segunda diz respeito a apenas o uso de instrumentos de autorrelato ou heterorrelato que são baseados na teoria MSC. Nesta classificação, os únicos instrumentos utilizados foram o Teste de IE de Autorrelato de Schutte (*Schutte Self-Report Emotional Intelligence Test*, ou SSEIT), a Escala de IE de Wong e Law (*Wong and Law Emotional Intelligence Scale*, ou WLEIS), ou o Perfil de IE do Grupo de Trabalho (*Workgroup Emotional Intelligence Profile*, ou WEIP).

A terceira classificação inclui estudos que usaram instrumentos de autorrelato ou heterorrelato que não se conformam à teoria MSC. Os testes utilizados pelos estudos dessa categoria são o Inventário de Competência Emocional (*Emotional Competence Inventory*, ECI; Hay Group, 2005), o Teste de Quociente Emocional Informatizado (*Emotional Quotient Inventory*, EQ–i; Bar-On, 2000), o Questionário de Traços de IE (*Trait Emotional Intelligence Questionnaire*, TEIQue; Petrides & Furnham, 2001), e o *Genos*. Estes testes formam parte da tradição de pesquisa de IE como traço de personalidade.

Esta última classificação chama atenção para um fenômeno comum nas ciências humanas: a pluralidade de perspectivas e níveis de análise. Tendo isto em vista, é natural que haja considerável sobreposição sobre os temas de estudo. Para obter uma visão universal de um tema, é relevante levantar-se algumas destas perspectivas em uma breve discussão, realizada a seguir.

Outras Perspectivas de Inteligência Emocional

No nível biológico, os limites do conteúdo da IE sobrepõem-se, por exemplo, com temas estudados nos campos das neurociências da emoção e da neuropsicologia social. Nas neurociências da emoção há uma longa tradição de estudo que embasa os primeiros desenvolvimentos da literatura em emoção na psicologia moderna, que tiveram como pergunta de pesquisa se a resposta emocional fisiológica precedia ou ocorria paralelamente à resposta

emocional mental (Cannon, 1927; James, 1884). Já na neuropsicologia social, há temas de estudo como a empatia e teoria da mente (Ward, 2017). Nesta perspectiva são estudados os mecanismos neurais destes construtos.

Com respeito ao estudo da empatia, a neuropsicologia social embasa o entendimento de que a habilidade de agir de forma eficaz em uma situação emocional não significa a tomada de decisão de agir dessa forma (Mayer et al., 2016). Por exemplo, Weisz e Zaki (2018) mostraram que essa tomada de decisão é mediada pela motivação para agir de forma empática. Esta modificação, ainda, não é inteiramente intrínseca, ou seja, além da característica disposicional, há um componente situacional da tomada de decisão.

Já "teoria da mente" é o nome que se dá ao construto acerca da teorização individual sobre o estado mental dos demais indivíduos com quem se entra em contato. Entendido como uma habilidade, é um pré-requisito básico para inteligência emocional que subjaz os quatro fatores da teoria de Mayer et al. (2012).

A neuropsicologia como um todo aborda processos de regulação emocional em um nível de análise ainda mais básico, por meio da função do executivo chamada controle inibitório (Diamond, 2013). Esta função diz respeito à "capacidade de controlar a atenção, comportamento, pensamentos e emoções a despeito de uma forte predisposição ou atração externa" (Diamond, 2013, p.137). No entanto, esta perspectiva não é particularmente dedicada ao estudo das emoções. Uma perspectiva que aplica conceitos cognitivos similares à emoção especificamente é a teoria de autorregulação, que foca em como os indivíduos regulam seus pensamentos, emoções e comportamentos (Bandura, 1991). Esta perspectiva propõe a existência de um mecanismo de autorregulação com 3 funções principais: o monitoramento do próprio comportamento, das suas causas e consequências; o julgamento deste comportamento com base em critérios pessoais e relativamente às condições do ambiente físico e social; e a reação afetiva consequente ao julgamento.

Em uma outra proposta, Drigas e Papoutsi (2018) partem de uma perspectiva metacognitiva para teorizar acerca da IE por meio de uma pirâmide de habilidades complementares e cumulativas. Estas habilidades dizem respeito desde a captação dos estímulos sensoriais emocionais até a resposta emocional, mas, na perspectiva metacognitiva, o foco está na medida em que indivíduos têm conhecimento sobre ou são capazes de controlar suas emoções.

Além das perspectivas já discutidas, o estudo da IE envolve uma série de interfaces. Por exemplo, na psicopatologia, é estudado como parte do mecanismo do Transtorno do Espectro Autista (Almeida et al., 2021). Na psicodinâmica do trabalho, a dificuldade para gerenciar o próprio comportamento emocional por meio da mobilização subjetiva pode incorrer em decepção, adoecimento psíquico e desvalorização do trabalho (Amaral et al., 2019).

Conclusão

O tema de IE tem atraído considerável interesse no meio acadêmico pelo seu potencial de contribuição nas faces largamente consideradas mais importantes da vida humana: família, relações, estudo, trabalho e bem-estar. Há também o interesse de conhecer o construto por seu papel em trabalhos específicos—os estudos que sugerem o aumento do desempenho de indivíduos com maior inteligência emocional já apontam para possíveis melhorias na sociedade como um todo. A qualidade do serviço de diversos profissionais está irrevogavelmente associada ao desenvolvimento emocional da sociedade. Novos estudos podem se ocupar dos benefícios da IE em diversas áreas.

O modelo de IE como habilidade apresenta uma proposta de estudo de um tipo de raciocínio que, até poucas décadas atrás, era controverso. Estudos recentes têm mostrado que tanto o modelo de IE como traço, como o modelo de IE como habilidade, tem contribuições independentes para o estudo e melhoramento da sociedade em diversos níveis, mas ainda existe demanda para uma série de estudos, teóricos e metodológicos, para resolver questões que têm

desafiado as medidas disponíveis há mais de 20 anos.

As controvérsias iniciais do construto criaram um ambiente de falta de confiabilidade na teoria Mayer–Caruso–Salovey injustificada, uma vez que, tanto pela natureza ontológica do construto enquanto inteligência, quanto pela definição explícita dos autores, a IE é um construto que se limita à habilidade de resolução de problemas de conteúdo emocional, e não pela propensão a esta resolução. Da mesma forma que, dentro da teoria da inteligência espacial descreve-se a habilidade de resolver um cubo de Rubik, e não a propensão a ou a vontade de resolver um cubo de Rubik, dentro da teoria da IE está a habilidade de resolver problemas emocionais, e não a propensão ou vontade de resolver estes problemas. Aspectos de personalidade são essenciais para o entendimento do fenômeno como um todo, mas, no contexto da teoria proposta por Mayer et al. (2012), são externos aos limites do construto. É exatamente por este motivo que autores como Mayer et al. (2016), Schneider e McGrew (2018), Gonzaga e Monteiro (2011), e Grubb e McDaniel (2007) recomendam que a IE seja medida por testes de desempenho, e não de autorrelato.

Tendo superada esta discussão, há a necessidade de elucidar aspectos de dimensionalidade da IE habilidade. Apesar de ser possível extrair um modelo com as dimensões de um determinado instrumento, a maneira na qual essas dimensões de fato representam o construto estudado só pode ser confirmada quando outro instrumento, independente, baseado naquele modelo chega ao mesmo resultado. Esta é uma metodologia que ainda não é desenvolvida, e não reflete o estado da arte do estudo de confirmação de modelos, que é comumente realizado em uma etapa exploratória e uma etapa confirmatória (Bollen, 1989).

Além de elucidar questões de dimensionalidade no nível das análises estatística, é importante que as habilidades de raciocínio da IE sejam mais claramente explicitadas em termos de habilidades cognitivas específicas. Estas habilidades devem ser claramente distintas de uma para outra dimensão. Tendo em vista Mayer et al. (2016) admitirem que as habilidades que a

teoria adscreve às dimensões são definitivamente representativas da IE como um todo, mas não necessariamente servem para diferenciar a facilitação e o gerenciamento emocional, é importante que uma reorganização seja feita que, de fato, contraste as reais dimensões da IE. Este esforço foi iniciado por Elfenbein e MacCann (2017), e precisa ser continuado.

Na literatura brasileira, recentes contribuições têm sido cruciais para avaliação da IE, mas há demanda de construção ou adaptação de novos instrumentos de IE para o Brasil, tendo em vista que a avaliação especificamente do modelo de habilidades do IE depende de testes de desempenho (Gonzaga & Monteiro, 2011; Grubb & McDaniel, 2007; Mayer et al., 2016; Schneider & McGrew, 2018;). Esta forma de contribuição, com o objetivo de eventualmente elucidar essas e outras questões da dimensionalidade da IE, é uma das mais importantes para o desenvolvimento da pesquisa em IE, nacionalmente e internacionalmente.

Referências

Almeida, F. S., Giordani, J. P. Yates, D. B., & Trentini, C. M. (2021). Avaliação de aspectos emocionais e comportamentais de crianças com Transtorno do Espectro Autista. *Aletheia, 54*(1) 85–95. https://doi.org/10.29327/226091.54.1-9

Amaral, G. A., Mendes, A. M., & Facas, E. P. (2019). (Im)possibilidade de mobilização subjetiva na clínica das patologias do trabalho: o caso das professoras readaptadas. *Revista Subjetividades, 19*(2). https://doi.org/10.5020/23590777.rs.v19i2.e8987

Ashkanasy, N. M., & Daus, C. S. (2020). Emotional Intelligence in the Workplace. *The Wiley Encyclopedia of Personality and Individual Differences*, 485–490. https://doi.org/10.1002/9781119547181.ch345

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*(2), 248–287. https://doi.org/10.1016/0749-5978(91)90022-L

Bar-On, R. (2000). Emotional and social intelligence: Insights from the Emotional Quotient Inventory. Em R. Bar-On & J. D. A. Parker (Orgs.), *The handbook of emotional intelligence: Theory, development, assessment, and application at home, school, and in the workplace* (pp. 363–388). Jossey-Bass/Wiley.

Bar-On, R. (2004). The Bar-On Emotional Quotient Inventory (EQ-i): Rationale, description and summary of psychometric properties. Em G. Geher (Org.), *Measuring emotional intelligence: Common ground and controversy* (pp. 115–145). Nova Science Publishers.

Bar-On, R. (2006). The Bar-On Model of Emotional-Social Intelligence. *Psicothema*, *18*(Suppl.), 13–25. https://www.psicothema.com/pdf/3271.pdf

Barrett, L. F. (2017). *How Emotions are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.

Brackett, M. A., & Mayer, J. D. (2003). Convergent, Discriminant, and Incremental Validity of Competing Measures of Emotional Intelligence. *Personality and Social Psychology Bulletin, 29*(9), 1147–1158. https://doi.org/10.1177/0146167203254596

Brackett, M. A., & Salovey, P. (2006). Measuring emotional intelligence with the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT). *Psicothema*, *18*(Suppl.), 34–41. https://www.psicothema.com/pdf/3273.pdf

Brand, A. G. (1985). Hot cognition: Emotions and writing behavior. *Journal of Advanced Composition*, *6*(1985–1986), 5–15. JSTOR 20865583. https://www.jstor.org/stable/20865583

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley Series in Probability and Mathematical Statistics. Wiley/Interscience.

Bryan, V. M., & Mayer, J. D. (2020). A meta-analysis of the correlations among broad intelligences: Understanding their relations. *Intelligence*, *81*(1), Artigo 101469. https://doi.org/10.1016/j.intell.2020.101469

Cabello, R., Sorrel, M. A., Fernández-Pinto, I., Extremera, N., & Fernández-Berrocal, P. (2016). Age and gender differences in ability emotional intelligence in adults: A cross-sectional study. *Developmental Psychology, 52*(9), 1486–1492. https://doi.org/10.1037/dev0000191

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Ciarrochi, J. v., Chan, A. Y. C., & Caputi, P. (2000). A critical evaluation of the emotional intelligence construct. *Personality and Individual Differences*, *28*(3), 539–561. https://doi.org/10.1016/S0191-8869(99)00119-1

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin, 52*(1), 281–302. https://doi.org/10.1037/h0040957

Davies, M., Stankov, L., & Roberts, R. D. (1998). Emotional intelligence: in search of an elusive construct. *Journal of Personality and Social Psychology*, *75*(4), 989–1015. https://doi.org/10.1037/0022-3514.75.4.989

Deary, I. J. (2012). Intelligence. *Annual Review of Psychology, 63*(1), 453–482. https://doi.org/10.1146/annurev-psych-120710-100353

DeYoung, C. G. (2011). *Intelligence and personality*. Em R. J. Sternberg & S. B. Kaufman (Orgs.), *Cambridge handbooks in psychology. The Cambridge handbook of intelligence* (pp. 711–737). Cambridge University Press. https://doi.org/10.1017/CBO9780511977244.036

Denham, S. A., McKinley, M., Couchoud, E. A., & Holt, R. (1990). Emotional and behavioral predictors of preschool peer ratings. *Child Development*, *61*(4), 1145–1152. https://doi.org/10.2307/1130882

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology, 64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Drigas, A. S., & Papoutsi, C. (2018). A New Layered Model on Emotional Intelligence. *Behavioral Sciences, 8*(5), 45. https://doi.org/10.3390/bs8050045

Elfenbein, H. A., & MacCann, C. (2017). A closer look at ability emotional intelligence (EI): What are its component parts, and how do they relate to each other? *Social and Personality Psychology Compass, 11*(7), Artigo e12324. https://doi.org/10.1111/spc3.12324

Elfenbein, H. A., Jang, D., Sharma, S., & Sanchez-Burks, J. (2017). Validating emotional attention regulation as a component of emotional intelligence: A Stroop approach to

individual differences in tuning in to and out of nonverbal cues. *Emotion*, *17*(2), 348–358. https://doi.org/10.1037/emo0000145

Evans, T. R., Hughes, D. J., & Steptoe-Warren, G. (2019). A Conceptual Replication of Emotional Intelligence as a Second-Stratum Factor of Intelligence. *Emotion*, *20*(3), 507–512. https://doi.org/10.1037/emo0000569

Fan, H., Jackson, T., Yang, X., Tang, W., & Zhang, J. (2010). The factor structure of the Mayer-Salovey-Caruso Emotional Intelligence Test V 2.0 (MSCEIT): A meta-analytic structural equation modeling approach. *Personality and Individual Differences*, *48*(7), 781–785. https://doi.org/10.1016/j.paid.2010.02.004

Fernández-Berrocal, P., & Extremera, N. (2016). Ability Emotional Intelligence, Depression, and Well-Being. *Emotion Review*, *8*(4), 311–315. https://doi.org/10.1177/1754073916650494

Fernández-Berrocal, P., Cabello, R., & Gutiérrez-Cobo, M. J. (2017). Understanding the relationship between general intelligence and socio-cognitive abilities in humans. *Behavioral and Brain Science*, *40*, Artigo e202. https://doi.org/10.1017/S0140525X1600162X

Goleman, D. (1995). *Emotional Intelligence: Why It Matters More than the IQ?* Bantam Books.

Gonzaga, A. R., & Monteiro, J. K. (2011). Inteligência emocional no Brasil: um panorama da pesquisa científica. *Psicologia: Teoria e Pesquisa, 27*(2), 225–232. https://doi.org/10.1590/S0102-37722011000200013

Gross, J. J. (1998). The Emerging Field of Emotion Regulation: An Integrative Review. *Review of General Psychology, 2*(3), 271–299. http://doi.org/10.1037/1089-2680.2.3.271

Gutiérrez-Cobo, M., Fernández-Berrocal, P., & Cabello, R. (2017). Performance-based ability emotional intelligence benefits working memory capacity during performance on hot

tasks. *Scientific Reports*, *7*(1), Artigo 11700. https://doi.org/10.1038/s41598-017-12000-7

Hay Group. (2005). *Emotional Competence Inventory (ECI): Technical Manual*. McClelland Center for Research and Innovation. Retrieved from https://www.eiconsortium.org/pdf/ECI_2_0_Technical_Manual_v2.pdf

Hodzic, S., Zenasni, F., Scharfen, J., Holling, H., & Ripoll, P. (2018). How Efficient Are Emotional Intelligence Trainings: A Meta-Analysis. *Emotion Review*, *10*(2), 138–148. https://doi.org/10.1177/1754073917708613

Horn, J. L. (1988). Thinking about human abilities. Em J. R. Nesselroade & R. B. Cattell (Orgs.), *Handbook of multivariate experimental psychology* (pp. 645–685). Academic Press.

Husin, W. N. I. W., Santos, A., Ramos, H. M., & Nordin, M. S. (2013). The Place of Emotional Intelligence in the 'Intelligence' Taxonomy: Crystallized Intelligence or Fluid Intelligence Factor? *Procedia: Social and Behavioral Sciences*, *97*(17), 214–223. https://doi.org/10.1016/j.sbspro.2013.10.225

Izard, C., Fine, S., Schultz, D., Mostow, A., Ackerman, B., & Youngstrom, E. (2001). Emotion knowledge as a predictor of social behavior and academic competence in children at risk. *Psychological Science*, *12*(1), 18–23. https://doi.org/10.1111/1467-9280.00304

James, W. (1884). What is an Emotion? *Mind*, *9*(1), 188–205. https://doi.org/10.1093/mind/os-IX.34.188

Joseph, D. L., & Newman, D. A. (2010). Emotional Intelligence: An Integrative Meta-Analysis and Cascading Model. *Journal of Applied Psychology*, *95*(1), 54–78. https://doi.org/10.1037/a0017286

Joseph, D. L., Jin, J., Newman, D. A., & O'Boyle, E. H. (2015). Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. *Journal of Applied Psychology, 100*(2), 298–342. https://doi.org/10.1037/a0037681

Lopes, P. (2016). Emotional Intelligence in Organizations: Bridging Research and Practice. *Emotion Review*, *8*(4), 316–321. https://doi.org/10.1177/1754073916650496

MacCann, C., Jiang, Y., Brown, L. E. R., Double, K. S., Bucich, M., & Minbashian, A. (2020). Emotional intelligence predicts academic performance: A meta-analysis. *Psychological Bulletin, 146*(2), 150–186. https://doi.org/10.1037/bul0000219

MacCann, C., Joseph, D. L., Newman, D. A., & Roberts, R. D. (2014). Emotional intelligence is a second-stratum factor of intelligence: Evidence from hierarchical and bifactor models. *Emotion*, *14*(2), 358–374. https://doi.org/10.1037/a0034755

Marin, A. H., da Silva, C. T., Andrade, E. I. D., Bernardes, J., & Fava, D. C. (2017). Social-emotional competence: concepts and associated instruments. *Revista Brasileira de Terapias Cognitivas*, *13*(2), 92–103. https://doi.org/10.5935/1808-5687.20170014

Mayer, J. D., & Salovey, P. (1993). The intelligence of emotional intelligence. *Intelligence, 17*(4), 433–442. https://doi.org/10.1016/0160-2896(93)90010-3

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? Em P. Salovey & D. J. Sluyter (Orgs.), *Emotional development and emotional intelligence: Educational implications* (pp. 3–31). Basic Books.

Mayer, J. D., Caruso, D. R., & Salovey, P. (2000). Emotional Intelligence Meets Traditional Standards for an Intelligence. *Intelligence*, *27*(4), 267–298. https://doi.org/10.1016/S0160-2896(99)00016-1

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion, 1*(3), 232–242. https://doi.org/10.1037/1528-3542.1.3.232

Mayer, J. D., Salovey, P., & Caruso, D. (2002). *Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT© V2.0) User's Manual*. MHS Publishers.

Mayer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology*, *59*, 507–536. https://doi.org/10.1146/annurev.psych.59.103006.093646

Mayer, J. D., Salovey, P., & Caruso, D. R. (2012). The validity of the MSCEIT: Additional analyses and evidence. *Emotion Review*, *4*(4), 403–408. https://doi.org/10.1177/1754073912445815

Mayer, J. D., Caruso, D. R., & Salovey, P. (2016). The Ability Model of Emotional Intelligence: Principles and Updates. *Emotion Review*, *8*(4), 290–300. https://doi.org/10.1177/1754073916639667

Mattingly, V. P., & Kraiger, K. (2019). Can emotional intelligence be trained? A meta-analytical investigation. *Human Resource Management Review, 29*(2), 140–155. https://doi.org/10.1016/J.HRMR.2018.03.002

McGrew, K. S., & Evans, J. J. (2004). *Carroll Human Cognitive Abilities Project: Research Report No. 2*. Institute for Applied Psychometrics. http://www.iapsych.com/HCARR2.pdf

McGrew, K. S. (2017). *Cattell–Horn–Carroll (CHC) theory of cognitive abilities (v2.5) "official" broad and narrow definitions*. IQ's Corner. Publicado em 10 de julho de 2017. Acessado em 2 de maio de 2023. Recuperado de http://www.iqscorner.com/2017/07/cattell-horn-carroll-chc-theory-of.html

McRae, K. (2016). Cognitive Emotion Regulation: A Review of Theory and Scientific Findings. *Current Opinion in Behavioral Sciences, 10*, 119–124. https://doi.org/10.1016/j.cobeha.2016.06.004

Mestre, J. M., Guil, R., Lopes, P. N., Salovey, P., Gil-Olarte, P. (2006). Emotional intelligence and social and academic adaptation to school. *Psicothema, 18*(Suppl.), 112–117. PMID: 17295953.

Miguel, F. K., Zuanazzi, A. C., de Lima, R., Eurich, J. C., & Tavares, C. A. (2016a). Estudo da aplicação coletiva de um teste de percepção emocional em surdos. *Avaliação Psicológica*, *15*(2), 197–205. https://doi.org/10.15689/ap.2016.1502.08

Miguel, F. K., Giromini, L., Colombarolli, M. S., Zuanazzi, A. C., & Zennaro, A. (2016b). A Brazilian Investigation of the 36- and 16-Item Difficulties in Emotion Regulation Scales. *Journal of Clinical Psychology, 73*(9), 1146–1159. https://doi.org/10.1002/jclp.22404

Miguel, F. K., Amaro, M. C. P., Huss, E. Y., & Zuanazzi, A. C. (2017a). Emotional perception and distortion correlates with Rorschach cognitive and interpersonal variables. *Rorschachiana, 38*(2), 143–159. https://doi.org/10.1027/1192-5604/a000096

Miguel, F. K., Zuanazzi, A. C., & Villemor-Amaral, A. E. (2017b). Assessment of Emotional Intelligence Aspects in the Methods of Pfister's and Zulliger's. *Trends in Psychology, 25*(4), 1863–1872. https://doi.org/10.9788/TP2017.4-17Pt

O'Boyle, E. H., Humphrey, R. H., Pollack, J. M., Hawver, T. H., & Story, P. A. (2011). The relation between emotional intelligence and job performance: A meta-analysis. *Journal of Organizational Behavior*, *32*(5), 788–818. https://doi.org/10.1002/job.714

Peña-Sarrionandia, A., Mikolajczak, M., & Gross, J. J. (2015). Integrating emotion, regulation and emotional intelligence traditions: a meta-analysis. *Frontiers in psychology*, *6*, Artigo 160. https://doi.org/10.3389/fpsyg.2015.00160

Petrides, K. V., & Furnham, A. (2000). On the dimensional structure of emotional intelligence. *Personality and Individual Differences, 29*(2), 313–320. https://doi.org/10.1016/S0191-8869(99)00195-6

Petrides, K. V., & Furnham, A. (2001). Trait Emotional Intelligence: Psychometric Investigation with Reference to Established Trait Taxonomies. *European Journal of Personality, 15*(6), 425–448. https://doi.org/10.1002/per.416

Petrides, K. V., Pita, R., & Kokkinaki, F. (2007). The Location of Trait Emotional Intelligence in Personality Factor Space. *British Journal of Psychology*, *98*(2), 273–289. https://doi.org/10.1348/000712606X120618

Petrides, K. V., Siegling, A. B., & Saklofske, D. H. (2016). Theory and Measurement of Trait Emotional Intelligence. Em U. Kumar (Org.). *The Wiley Handbook of Personality Assessment*. Wiley Blackwell. https://doi.org/10.1002/9781119173489.ch7

Petrides, K. V., Sanchez-Ruiz, M.-J., Siegling, A. B., Saklofske, D. H., & Mavroveli, S. (2018). Emotional intelligence as personality: Measurement and role of trait emotional intelligence in educational contexts. In K. V. Keefer, J. D. A. Parker & D. H. Saklofske (Eds.), *Emotional intelligence in education: Integrating research with practice* (pp. 49–81). Springer. https://doi.org/10.1007/978-3-319-90633-1_3

Primi, R., Santos, D., John, O. P., & de Fruyt, F. (2021a). SENNA Inventory for the Assessment of Social and Emotional Skills in Public School Students in Brazil: Measuring Both Identity and Self-Efficacy. *Frontiers in Psychology, 12*(1), Artigo 716639. https://doi.org/10.3389/fpsyg.2021.716639

Primi, R., Santos, D., John, O. P., & de Fruyt, F. (2021b). *SENNA Inventory for the Assessment of Social and Emotional Skills: Technical Manual*. Instituto Ayrton Senna. https://doi.org/10.31234/osf.io/byvpr

Reddy, W. M. (2001). *The Navigation of Feeling: A Framework for the History of Emotion*s. Cambridge University Press.

Roberts, R. D., Zeidner, M., & Matthews, G. (2001). Does Emotional Intelligence Meet Traditional Standards for an Intelligence? Some New Data and Conclusions. *Emotion*, *1*(3), 196–231. https://doi.org/10.1037/1528-3542.1.3.196

Sánchez-Álvarez, N., Extremera, N., & Fernández-Berrocal, P. (2016). The relation between emotional intelligence and subjective well-being: A meta-analytic investigation. *Journal of Positive Psychology*, *11*(3), 276–285. https://doi.org/10.1080/17439760.2015.1058968

Sanchez-Garcia, M., Extremera, N., & Fernández-Berrocal, P. (2016). The factor structure and psychometric properties of the Spanish version of the Mayer-Salovey-Caruso Emotional Intelligence Test. *Psychological Assessment*, *28*(11), 276–285. https://doi.org/10.1037/pas0000269

Santos, M. V., Nakano, T. C., & Silva, T. F. (2015). *Competências socioemocionais: análise da produção científica nacional e internacional* [Apresentação de Trabalho, Psicologia USP, São Paulo]. ResearchGate 283777647. Recuperado de https://www.researchgate.net/publication/283777647.

Schneider, W. J, & McGrew, K. S. (2012). The Cattell-Horn-Carroll Theory of Cognitive Abilities. Em D. P. Flanagan, E. M. McDonough. (Orgs.). *Contemporary Intellectual Assessment: Theories, Tests and Issues* (3rd ed.). Guildford Publications.

Schneider, W., Mayer, J., & Newman, D. (2016). Integrating Hot and Cool Intelligences: Thinking Broadly about Broad Abilities. *Journal of Intelligence*, *4*(1), Artigo 1. https://doi.org/10.3390/jintelligence4010001

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. Em D. P. Flanagan, E. M. McDonough. (Orgs.). *Contemporary Intellectual Assessment: Theories, Tests and Issues* (4th ed.). Guildford Publications.

Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., & Dornheim, L. (1998). Development and validation of a measure of emotional intelligence. *Personality and Individual Differences, 25*(2), 167–177. https://doi.org/10.1016/S0191-8869(98)00001-4

Thurstone, L. L. (1938). *Primary Mental Abilities*. University of Chicago Press. http://www.uwpsychiatry.org/sls/publications/primary_mental_abilities.pdf.

Thorndike, E. L. (1920). Intelligence and its uses. *Harper's Magazine, 140*, 227–235.

van der Linden, D., Pekaar, K. A., Bakker, A. B., Schermer, J. A., Vernon, P. A., Dunkel, C. S., & Petrides, K. v. (2017). Overlap between the general factor of personality and emotional intelligence: A meta-analysis. *Psychological Bulletin*, *143*(1), 36–52. https://doi.org/10.1037/bul0000078

Van Rooy, D. L., & Viswesvaran, C. (2004). Emotional intelligence: A meta-analytic investigation of predictive validity and nomological net. *Journal of Vocational Behavior, 65*(1), 71–95. https://doi.org/10.1016/S0001-8791(03)00076-9

Vieira-Santos, J., Lima, D. C., Sartori, R. M., Schelini, P. W., & Muniz, M. (2018). Inteligência emocional: revisão internacional da literatura. *Estudos Interdisciplinares em Psicologia*, *9*(2), 78–99. https://doi.org/10.5433/2236-6407.2018v9n2p78

Weisz, E., & Zaki, J. (2018). Motivated empathy: a social neuroscience perspective. *Current Opinion in Psychology, 24*, 67–71. https://doi.org/10.1016/j.copsyc.2018.05.005

Zuanazzi, A. C., Ricci, D. S., & Miguel, F. K. (2015). Avaliação da Alexitimia e Percepção Emocional: Comparação entre Autorrelato e Desempenho. *Trends in Psychology, 23*(4), 831–842. https://doi.org/10.9788/TP2015.4-03

MANUSCRITO 2

Validity Evidence for the Brazilian Version of the Situational Tests of Emotional Intelligence

Evidências de Validade para a Versão Brasileira dos Testes Situacionais de Inteligência Emocional

Abstract

This study aimed to adapt the Situational Test of Emotional Understanding (STEU) and Situational Test of Emotional Management (STEM) to Brazilian culture, as well as to collect validity evidence based on test structure and based on relationship with external variables. The adaptation was authorized by the original author and followed a traditional backtranslation process. The adapted test version was then analyzed empirically. Data from 688 participants who answered the tests, as well as the Satisfaction with Life Scale (SWLS) and the Reduced Scale of the Big Five Personality Factors, were subjected to item response theory modeling. After removing items that did not reach the required cutoffs, the STEU had 32 items, and the STEM, 30. These final test versions reached adequate levels of both global (M^2*, NNFI, RMSEA) and per-item (*a*, *S*-χ^2, RMSEA) fit indices. As expected, the results show medium correlation among the EI tests, r = .5, and little to no correlation with personality factors. However, a small relationship between the STEM and the SWLS described by the original authors was not replicated, whereas a small correlation between the STEU and SWLS was found. In line with Mislevy (2007)'s concept of *validity by design*, the employment of a widely recognized adaptation procedure that led to the validity evidence being presented, coupled with the theory-based construction of the instruments in MacCann and Roberts's (2008) original study, as well as the empirical relationships observed, support an overall convincing validity argument. Future studies should examine the relationship between the adapted tests and other performance measures of EI. However, given the lack of practical significance of the correlations with the SWLS, studies should avoid replicating this finding and seek out other measures of life satisfaction, if that is the construct of interest.

Keywords: emotion understanding, emotion management, emotional intelligence, Mayer-Salovey–Caruso theory.

Resumo

Este estudo teve como objetivo adaptar o Teste Situacional de Compreensão Emocional (STEU) e de Gerenciamento Emocional (STEM) para a cultura brasileira, bem como coletar evidências de validade com base na estrutura do teste e com base no relacionamento com variáveis externas. A adaptação foi autorizada pelas autoras originais e seguiu um processo tradicional de retrotradução. A versão adaptada do teste foi então analisada empiricamente. Dados de 688 participantes que responderam aos testes, bem como à Escala de Satisfação com a Vida (SWLS) e à Escala Reduzida dos Fatores de Personalidade (ER5FP), foram submetidos à modelagem da teoria de resposta ao item. Após a remoção dos itens que não atingiram os pontos de corte exigidos, o teste STEU ficou com 32 itens e o STEM com 30. Essas versões finais do teste atingiram níveis adequados nas medidas de ajuste globais (M^2*, NNFI, RMSEA) e por item (*a*, S-χ^2, RMSEA). Conforme esperado, os resultados mostram correlação média entre os testes de EI, r = 0,5, e pouca ou nenhuma correlação com fatores de personalidade. No entanto, uma pequena correlação entre o STEM e o SWLS descrito pelos autores originais não foi replicada, enquanto uma pequena correlação entre o STEU e o SWLS encontrada não havia sido descrita no estudo original. Em consonância com o conceito de validade por design de Mislevy (2007), o emprego de um procedimento de adaptação amplamente reconhecido, juntamente com a construção baseada em teoria dos instrumentos no estudo original de MacCann e Roberts (2008), bem como as relações empíricas observadas, suportam um argumento de validade convincente. Estudos futuros devem examinar a relação entre os testes adaptados e outras medidas de desempenho de IE. No entanto, dada a falta de significância prática das correlações com o SWLS, novos estudos devem evitar a replicação desse achado e buscar outras medidas de satisfação com a vida, se esse for o construto de interesse.

Palavras-chave: compreensão emocional, gerenciamento emocional, regulação emocional, inteligência emocional, teoria Mayer–Salovey–Caruso.

Validity Evidence for the Brazilian Version of the Situational Tests of Emotional Intelligence

The Mayer–Salovey–Caruso (MSC) theory remains the most prominent systematization used in the ability emotional intelligence (EI) research tradition (Vieira-Santos et al., 2018). The theory comprises four factors, which the authors call "branches" due to their ontogenetic development process. These factors are emotion perception, emotion understanding, emotion management, and emotion facilitation (Mayer et al., 2012).

The theory was initially presented as the underlying model of the Multifactor Emotional Intelligence Scale (Mayer et al., 1997), which has since been superseded by the Mayer–Salovey–Caruso Emotional Intelligence Test, or MSCEIT (Mayer et al., 2002). For the authors, emotional intelligence is defined as "the mental processes involved in the recognition, use, understanding, and management of one's own and others' emotional states to solve problems and regulate behavior" (p. 34; Brackett & Salovey, 2006). The definitions of the dimensions of EI, however, were as important as that of the EI construct itself. The factors of the MSC theory model, along with their constitutive and operational definitions, are displayed in Table 1.

Table 1

The Four-Factor Model of Emotional Intelligence According to the Mayer–Salovey–Caruso Ability Emotional Intelligence Theory (Mayer et al., 2003; Mayer et al., 2016).

Branch and Definition	Skills (Operationalization)
First Branch: Emotion Perception The ability to perceive emotions in oneself and others, as well as in objects, art, stories, music, and other stimuli.	• Identify honest and dishonest emotional expressions. • Discriminate accurate from inaccurate (e.g., computer-generated) emotional expressions. • Understand how emotions are expressed depending on context and culture.

	• Express contextually relevant emotions accurately when desired. • Perceive emotional content in the environment, in visual arts, and in music. • Perceive others' emotions through clues in their voice, facial expression, language, and behavior. • Identify emotions in one's own mental states and the emotions associated with one's own feelings and thoughts.
Second Branch: *Emotion Facilitation* The ability to generate, use, and feel emotion as necessary to communicate feelings, or employ them in other cognitive processes.	• Identify emotions in one's own physical states, feelings, and thoughts. • Prioritize solving problems based on whether the ongoing emotional state can facilitate the requisite cognitive processes. • Take advantage of mood swings to bring about different cognitive perspectives. • Prioritize thinking by directing attention according to the present feeling. • Generate emotions as a means of relating to someone else's experiences. • Generate emotions as a means to aid judgment and memory.
Third Branch: *Emotion Understanding* The ability to understand emotional information, how emotions combine and progress through relationship transitions and to appreciate such emotional meanings.	• Recognize cultural differences in the evaluation of emotions. • Understand how a person can feel in the future or under certain conditions (affective forecast). • Recognize likely transitions between emotions, such as from anger to satisfaction. • Understand complex and mixed emotions. • Differentiate between humor and emotions.

	• Evaluate situations that are likely to evoke emotions. • Determine the background, meanings, and consequences of emotions. • Enumerate emotions and recognize the relationships between them.
Fourth Branch: Emotion Management The ability to be open to feelings, to modulate them in oneself and others so as to promote personal understanding and growth.	• Intentionally manage one's own emotions to achieve a desired result. • Evaluate strategies to maintain, reduce or intensify an emotional response. • Monitor emotional reactions to determine whether they are reasonable. • Engage emotions that are useful and disengage those that are not. • Accept pleasant and unpleasant feelings as necessary and stay open to the information they convey.

According to Brackett and Salovey (2006), their final factor structure was reached through analyses of statistical studies of emotion-related abilities. Then, Mayer et al. (2003) ran a confirmatory factor analysis wherein the eight tasks of the MSCEIT test were distributed among an increasing number of factors (Mayer et al., 2003). Although all models had acceptable fit indices that got increasingly good, the MSC theory's four-factor solution performed best.

The MSC theory has, since then, reached considerable acclaim (Vieira-Santos et al., 2018). For instance, after decades of arguing for the recognition of EI as a classical intelligence (MacCann et al., 2014; Mayer et al., 2000; Mayer et al., 2012), the Cattell–Horn–Carroll (CHC) theory has recently been updated to include emotional intelligence as a broad ability (Schneider & McGrew, 2018). Though the inclusion of the four-factor model under a broad ability, called emotional reasoning, was characterized as "tentative" (McGrew, 2018), it is nevertheless

recognition of the merits of the MSC theory, whose tetrafactorial structure it mirrors in its narrow abilities. Among these narrow abilities of the factor, called emotion reasoning, are emotion perception, emotion knowledge (emotion understanding, cf. MSC theory), emotion utilization (emotion facilitation, cf. MSC theory) and emotion management.

In Brazil, a substantial research tradition studies EI-related constructs under the name "socioemotional competencies", a construct which has considerable overlap with EI (Marin et al., 2017). This research tradition has generally focused on mixed EI models. These are models that also recognize the contribution of personality traits to the EI construct, in addition to the ability model (O'Connor et al., 2019).

Comprehensive Brazilian socioemotional competencies instruments have been developed, such as the BOLIE (Online Battery of Emotional Intelligence; Miguel, 2016) and the SENNA Inventory (Primi et al., 2021). The BOLIE measures aspects of ability EI such as emotional understanding, emotional speed, and emotion regulation. The SENNA Inventory measures, among a list of personality traits, the ability of emotion regulation, which features prominently in the MSC theory (Mayer et al., 2012).

Although these instruments have been successful at their aims, measuring the ability EI factors described under the MSC theory in Brazil has not been possible. The MSCEIT, which is the primary way that is usually accomplished (Vieira-Santos et al., 2019), is not approved for use in Brazilian samples. While there is a published analysis of its validity evidence and precision statistics (Junior & Noronha, 2008), it is not available in the commercial market Brazil, and even using it for research requires a partnership with the copyright holder.

Even under such a partnership, although Junior and Noronha (2008) followed a common validation protocol at the time, the validity evidence they produced would not be considered adequate today. For instance, the researchers used the visual screen plot "elbow" method alone to decide the number of factors, which has since been superseded by other, less subjective

methods, including parallel analysis, optimal coordinates, and acceleration factor (Raiche et al., 2013).

Additionally, the authors transformed the eigenvectors using the Varimax rotation, which is an orthogonal rotation method designed to maximize factor loadings when there is no correlation between the factors (Jackson, 2014). This means it is rarely useful in psychological instruments, for which oblique methods such as Direct Oblimin, or the computationally less intensive Promax, are adequate (Revelle, 2023; Watkins, 2018). The rotated factor solution may have suppressed results that would justify the removal of individual items, or even changes to the factor structure. Finally, reliability estimates used Cronbach's alpha, which underestimates reliability when there is collinearity. Present research indicates that Guttman (1945)'s *lambda-2* would be the adequate choice.

In sum, although the MSC theory has been dominant in the ability EI literature tradition, there are currently no instruments dedicated to measuring the aspects enumerated in that theory in Brazilian Portuguese. Therefore, the current study aimed to, first, adapt two emotional intelligence tests that do that. These are the Situational Test of Emotional Understanding (STEU) and the Situational Test of Emotional Management (STEM; MacCann & Roberts, 2008). Second, this study aimed at collecting and analyzing validity evidence for the adapted versions of the tests.

The Situational Tests of Emotional Understanding and the Situational Test of Emotional Management

The STEU and the STEM are two ability EI tests initially constructed by MacCann and Roberts (2008). The tests measure two dimensions of the MSC theory, and were designed in response to the fact that, although the MSC theory has achieved much success, almost all studies done on the ability EI paradigm have used the MSCEIT as the main EI measure (Vieira-Santos et al., 2013). While this has arguably been a source of benefits (for instance,

comparability), the authors of the tests suggest it is also a source of problems (MacCann & Roberts, 2008). They identify two major problems: first, that construct effects cannot be distinguished from test effects. Second, that the two scoring methods are contradictory.

To understand the first point, it is important to understand that the score of a test is composed by the true score in the construct plus an error factor (Hogan, 2007). The error is made up of construct-irrelevant variance in the test items, among other things. The way one could control for this error is by using more than one test as a measure of the construct. The true score of the construct will be correlated, but there is no reason for the error associated with the tests to be correlated. Therefore, the correlation between the scores of two tests will refer to the actual true score. The problem is that, by relying on that single test, the scores are contaminated with construct-irrelevant variance stemming from the test's development (MacCann & Roberts, 2008).

The second problem is that one of the scoring methods for the items in the MSCEIT is based on answers by a high EI group, and the other was based on an expert panel—so they were not based on theory (MacCann & Roberts, 2008; Mayer et al., 2003). The result of this was inconsistency between the two forms of scoring, which has been harshly criticized (Keele & Bell, 2009). To deal with that, MacCann and Roberts (2008) used a combination of Roseman (2001)'s structural emotions model and the situational judgment test (STJ) paradigm to design the STEU and the STEM (Motowidlo et al., 1990).

Roseman's Structural Theory of Emotions

Roseman's (2001) appraisal theory is a cognitive approach that seeks to describe the causal roots of emotions. It argues that emotions are caused as a reaction to people's appraisal of events, which are based on five "cognitive dimensions". By understanding a person's appraisal under those five different perspectives, the author argues it is possible to predict whether an emotion will occur, and which specific emotion it will be. The five different

dimensions of appraisal are called, simply: motivational state, situational state, probability, agency, and power (or legitimacy; Roseman, 1979, 1984, 2001).

The results of the appraisals can be one of two or three options, which are the following. Regarding the motivational state, a person may expect an event to be rewarding or punishing. Regarding the situation, a person may feel that an event is consistent or inconsistent with their motivational state. A person may also judge an event to be certain or uncertain, which characterizes the probability dimension. Regarding power, a person may feel their position in a given event is either weak or strong. Finally, the agent to which the event is attributed may be either unspecific circumstances, someone else specifically, or the person themselves.

The combinations between the appraisals result in one of 18 emotions, since many combinations result in the same emotion (MacCann & Roberts, 2008). The model was developed from reading 200 reports of emotional situations and assessed in a study in which 120 undergraduates participated.

Roseman (2001) calls these judgment aspects "dimensions" because they are independent, and each judgment in any one dimension is combined with each other judgment in all other dimensions. In this way, it is possible to plot these dimensions in five dimensions. However, since the assessment is categorical, visualization is more efficiently done through a five-way cross-table. The interaction between the values that are chosen in each dimension is one of 13 basic emotions spanning the affective spectrum.

The Situational Judgment Test Paradigm

The situational judgment test paradigm (SJT) is a method of psychological assessment that presents respondents with hypothetical situations and presents items that evaluate how they would manage those scenarios. It assesses individuals on responses to challenging situations that they might encounter in a simulated environment, encouraging them to find the most effective way of dealing with them. Since Motowidlo et al. (1990)'s seminal paper highlighting

its potential utility in predicting performance, it has become associated with selection processes in industrial psychology settings.

While the paradigm has drawn significant criticism over difficulties with factorability and reliability, recent research suggests that a significant problem was the use of traditional techniques to create SJT instruments (Lievens & Motowidlo, 2016), which require specific procedures. Guenole et al. (2017) have noted that the usual procedures lead to the construction of items associated by content, but not construct, leading to difficulties in establishing unidimensional factors.

The shortage of instruments to study ability EI is a worldwide problem (Bru-Luna et al., 2021; O'Connor et al., 2019). Reviews of EI testing methodology suggest that this could be due to a variety of reasons, such as the complexity of the construct, the difficulty in developing valid and reliable measures, or the lack of consensus on how to define and measure emotional intelligence. In Brazil, the situation is exacerbated by the fact that one of the most popular tests identified in these reviews (Bru-Luna et al., 2021), the MSCEIT, is not authorized for use.

A Brazilian validity study of the MSCEIT struggled to identify more than two factors (Junior & Noronha, 2008). This has been a point of debate and is reflected in the EI factor structure proposed in the CHC model. While the full four-factor model was introduced "tentatively", two of the factors, emotion knowledge, and emotion perception, are currently marked as "major abilities", while the other two, emotion management and emotion utilization, are marked "tentative abilities" (McGrew, 2017).

There is wide support for the idea that the content of those abilities marked tentative describes valid aspects of EI and that they are distinct from emotion knowledge and perception (Mayer et al., 2016). Still, it is not clear whether this factor structure is optimal. For instance, several studies have found that, while ability EI fits the bigger general intelligence model, a

three-factor solution, omitting the emotion utilization factor was statistically preferred (Gignac, 2005; Palmer et al., 2005; Rossen et al., 2008).

Despite limited use of the MSCEIT in Brazil, including efforts to collect validity evidence (Junior & Noronha, 2008) and a study associating EI and work performance (Cobêro et al., 2006), the MSCEIT has not been available for use, including in research. This has created a gap in the literature, and there is a clear demand for new instruments to measure EI under the ability model (Gonzaga & Monteiro, 2011) and to provide alternative means of measuring EI that do not rely on self-report measures. Such measures have been shown to be susceptible to faking high scores (Grubb & McDaniel, 2007), which compromises their ability to be used in high-stakes situations such as personnel selection. For this reason, many authors recommend the use of performance EI tests (e.g., Mayer et al., 2016; Schneider & McGrew, 2018; Grubb & McDaniel, 2007; Gonzaga & Monteiro, 2011). This study aimed to adapt to the Brazilian cultural context two ability EI performance tests, the STEU and the STEM, and to collect further validity evidence to support their use.

Methods

Study 1: Adaptation

Participants

The adaptation process involved two translators and four subject-matter experts (SMEs). Translators were recruited based on their fluency in both Brazilian Portuguese and English and experience in either translation or English teaching. SMEs were selected based on their graduate-level training in psychometrics, psychological test development (as suggested by Sireci et al., 2006), and intellectual assessment.

Instruments

Translation and retranslation instrument. The translation instrument consisted of a form with separate sections for each test. The first section requested the translation of the STEU instructions, followed by the original, English version instructions text, and an area in which the translation to be typed. Next, translation of the items was requested. The items were made available on a table, which had an empty column in which the translations could be typed. The second section was analogous to the first but pertained to the STEM. The retranslation instrument was similar to the translation instrument, but requested the opposite translation, and contained the translated, Brazilian Portuguese text, instead of the original.

Translation and retranslation validation instrument. The translation validation instrument contained the instructions and the constitutive and operational definitions of the constructs evaluated by the STEU and the STEM. Two sections followed, the first section containing STEU translations with Portuguese language instructions and an approval checkbox, then the Portuguese version items, followed by the approval checkbox. Next to each item, there was also a text area in which SMEs could type additional recommendations and comments. The retranslation validation instrument was similar to the translation validation instrument but contained the retranslated English language text instead of the Portuguese language translation.

Procedure

The adaptation procedure followed the steps recommended by Brislin (1970) while implementing recommendations by Sireci et al. (2006) and the International Test Commission (2017) regarding the educational and professional background of the SMEs. The steps were as follows.

Translation. Two experienced translators were invited to participate in the study and provided with the translation instrument. Following the International Test Commission (2017)'s recommendation, they were instructed to prioritize consistency in the translation of emotional

terminology across items, and to convey emotional situations accurately. The researchers collected the completed translation forms, compared them, and either selected the definitive version, or created one, incorporating elements from both translations.

Validation. The SMEs were provided with the translation validation form. They were given four weeks to study the construct definition and to evaluate the accuracy of the translation, the pertinence of the distractors, and the adherence of the item to the construct definitions. After receiving all the SME instruments, the researchers composed a final Brazilian Portuguese version incorporating feedback from the SME's evaluations. In cases where SMEs provided conflicting feedback, a consensus was reached through discussion and consultation with additional experts if necessary.

Backtranslation. A translator with test developing experience was selected to perform the backtranslation. A digital form containing only the final Brazilian Portuguese version instructions and items, as well as a text field to type in the English retranslation, was provided. The translator was also instructed to prioritize consistency across emotion translations. The completed instrument was returned the next day.

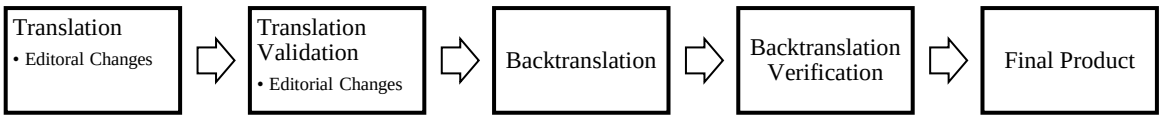
Backtranslation validation. A translator with experience in test development and emotional intelligence research was recruited. They received an instrument containing the original English version, the retranslated version, and approval checkboxes, which they were instructed to check when they verified that the retranslated English version was accurate.

Challenges encountered during the adaptation process included resolving disagreements between translators and SMEs regarding item translation accuracy and content specification adherence. These challenges were addressed through consultation with additional experts and discussion among study team members to reach a consensus.

The procedure is summarized in Figure 1.

Figure 1

Backtranslation Procedure for Cross-Cultural Test Adaptation (Brislin, 1970).



Study 2: Validation

Participants

A sample of 688 participants was recruited through social media platforms. Median age was 23 (MAD = 7.4), with a predominantly female (81.5%) sample. The largest occupation group was undergraduate students (21.7%), followed by psychologists (17.9%). Most participants were single (54.8%), while married people made up the second most common relationship status (34.01%).

Instruments

Data were collected through the Concerto Platform (University of Cambridge Psychometrics Center; Cambridge, UK), hosted on Amazon Web Services (Amazon, Inc.; Seattle, WA, USA), and Google Forms (Alphabet, Inc.; Mountain View, CA, USA). In addition to the Free Consent and demographic forms, participants completed four psychological tests in Brazilian Portuguese: the full versions of the adapted STEU and STEM (MacCann & Roberts, 2008), the Reduced Scale of the Big Five Personality Factors (ER5PF; Passos & Laros, 2015), and the Satisfaction with Life Scale (SWLS; Oliveira et al., 2009).

The STEU (MacCann & Roberts, 2008) consists of 42 multiple-choice items designed to evaluate test taker's ability to identify emotions in context. Each item describes an emotionally charged situation involving a fictitious character and requires the respondent to identify the emotion most likely felt by the character. An example is given on the instructions: "Clara receives a gift. Clara is most likely to feel?". The five response options are: a) happy; b) angry;

c) frightened; d) bored; or e) hungry. The test is scored according to a test key that specifies one correct answer.

The STEM (MacCann & Roberts, 2008) evaluates emotional management through 44 multiple-choice items. These items also develop an emotionally charged scenario but require the participant to choose the most effective action for managing the emotional situation. For instance, one item starts with "Lee's workmate fails to deliver an important piece of information on time, causing Lee to fall behind schedule also. What action would be the most effective for Lee?", and the participant must choose one of four options: a) Work harder to compensate; b) Get angry with the workmate; c) Explain the urgency of the situation to the workmate; and d) Never rely on that workmate again. The STEM allows for partial credit scoring.

The development of both tests was based on qualitative analyses of semi-structured interviews, from which were developed items within the framework of the situational judgment tests (Motowidlo et al., 1990) and with the underlying theoretical foundation of Roseman (1984)'s structural theory of emotions. The tests' reliability scores were α = .71 for the STEU, and α =.68 for the STEM. The tests were correlated with each other (r = .70) and with related constructs. They were positively correlated with constructs such as verbal reasoning (r =.49 and r = .41, respectively), academic success (r = .37 and r = .16, respectively), but negatively correlated with alexithymia (r = -.38 and r = -.43, respectively).

The ER5FP (Passos & Laros, 2015) is a 20-item scale consisting of four items for each of the Big Five personality dimensions: Openness, Conscientiousness, Agreeableness, Extroversion and Neuroticism. Participants rate their agreement on a Likert scale with whether stimulus words described themselves. Each item contained one word, for example, "Extroverted" or "Timid". The reliability of the scales varied between λ_2 = .71 and λ_2 = .85. A confirmatory model containing the five personality factors had acceptable fit measures, χ^2 (160) = 304.53, TLI = .94, CLI = .95, RMSEA = .05, SRMSR = .062.

The SWLS (originally by Diener et al., 1985), as adapted to Brazilian Portuguese (Gouveia et al., 2003), is a five-item instrument that also uses a Likert scale to assess attitudes toward sentences associated with life satisfaction. The English version uses items such as “I’m satisfied with my life”. A unidimensional confirmatory factor analysis (CFA) model was fit which revealed favorable fit measures, χ^2 (5) = 5.02, p > .05, SRMSR = .02, AGFI = .98, and reliability was also acceptable, α = .72 (Gouveia et al., 2003).

All data analysis was conducted using the R programming language (version 4.2.2). The packages utilized were, for CFA and item response theory (IRT) model fitting, *lavaan* 0.6-13 (Roseel, 2012), and *mirt* 1.37.1 (Chalmers, 2012), respectively.

Procedure

Data collection began on the Concerto Platform but later moved to Google Forms due to hosting issues. After data collection ended, the databases were extracted and merged. Exploratory data analysis revealed no abnormal response patterns and descriptive statistics were calculated from the sociodemographic variables. Because the data collection method only allowed submitting full responses, no missing data treatment was necessary.

The SWLS and ER5FP, which are Likert-type scales, did not require scoring. The STEU and the STEM scales were scored according to R-friendly adaptations of the IBM SPSS Statistics scripts provided by the authors (MacCann & Roberts, 2008). The scripts provided dichotomous (correct or incorrect) scoring for the STEU and polytomous scoring (allowing for partial credits) for the STEM.

Each instrument was submitted to confirmatory factor analysis using the diagonally weighted least squares method with robust standard errors, which is equivalent to the “WLSMV” option in Mplus (Muthen & Muthen; Los Angeles, CA). Model fit was assessed using absolute and relative fit indices AGFI, CFI and NNFI, as well as residual-related fit indices, the RMSEA and the SRMSR.

IRT modeling using the expectation maximization (EM; Bentler & Dijkstra, 1985) algorithm was then run for the four psychological tests. Modeling the STEU required the three-parameter logistic (3PL) model, due to its items being dichotomous. The STEM was modeled using the generalized partial credit model in classical parameterization (GPCM; Muraki & Carson, 1995). In contrast with the more commonly used graded response model (Samejima, 1969), the GPCM assumes each score to necessarily be more difficult to obtain than the last (e.g., the difficulty to get a B grade is necessarily larger than the difficulty to get a C grade, but necessarily smaller than the difficulty to get an A grade). No other specifications, such as priors or constraints were modeled—the default prior of N (0;1) theta distribution was used. The SWLS and ER5FP tests were both modeled under Samejima (1969)'s graded response model, which consists of sequential 2PL models.

Item fit was assessed according to the signed chi-squared test (S-χ^2; Orlando & Thissen, 2000; 2003), and its adaptation to polytomous models (Kang & Chen, 2008). The S-χ^2 is a goodness of fit index designed to overcome the limitations of traditional item-fit statistics which were only capable of assessing the fit of the entire model, which left the researcher at the risk of approving low-quality items. The study where the S-χ^2 index is originally described suggests that the measure may be a useful tool in detecting the misfit of any single item contained in an otherwise well-fitted test (Kang & Chen, 2008).

Items were dropped out of the test whenever they failed to reach the cutoff .7 discrimination level, if mean difficulty fell outside of the interval [-4, 4], or if they failed the S-χ^2 test. If any item was dropped out, IRT modeling was then repeated on the remaining items. Only the final item parameters were considered for further analysis. All items included in the calculation of final scores had discrimination levels of at least .7.

The fit of each score to each person ("person fit") was assessed by the infit and outfit statistics, as well as the *Zh* statistic, which is a polytomous version of the I_z statistic. Both were

proposed by Drasgow et al. (1985). The l_z refers to the standardized log-likelihood of each respondent's scored response vector and is considered the most effective person-fit measure for dichotomous IRT models (Armstrong et al., 2007).

Overall model fit was evaluated according to the M^2* or the hybrid C^2 statistic (Cai & Hansen, 2013; Cai & Monro, 2014). The M^2* statistic is the polytomous version of the M2 fit statistic used to measure the overall goodness of fit of IRT models. When the model had too few degrees of freedom to calculate the M^2* statistic, the hybrid C^2 statistic was used instead.

In either CFA or IRT modelling, data were considered to emerge from latent variables distributed as $N(0; 1)$. Scores were only calculated for the IRT models, using the same $N(0; 1)$ distribution as a prior for the expected a posteriori (EAP) method. Pearson correlations between all the calculated scores were estimated. The significance level for all analyses was .05.

Hypotheses

Both the STEU and the STEM were analyzed as unidimensional tests for two reasons. First, the theoretical model employed both in the construction and in the adaptation of the tests was the MSC theory (Mayer et al., 2016), which provided for homonymous unidimensional factors. That accounts for the factor structure, which is hierarchical with EI as the general factor, and the constitutive definition and operationalization of the constructs. Second, the development of the items was oriented by a substantive casual theory (Roseman, 1984), under which unidimensionality was also advised.

Given the hierarchical MSC model, significant, medium-sized correlations were found between STEM and STEU. Further hypotheses were based on MacCann and Roberts's (2008) original study: no correlations are expected with the major personality factors, except for Amiability, and, as further discussed by Mayer et al. (2008) and Schneider and McGrew (2018), EI is expected to be correlated with SWLS scores—which in MacCann and Roberts's (2008)

study was represented by a weak correlation between the SWLS and the STEM. Correlation p-values were one-tailed whenever one of these hypotheses were tested.

Results

For the translation validation procedure, the SMEs approved the instructions and all items, although some of them were approved with modifications. The researchers approved all the modifications, and the items were sent to be retranslated. For the retranslation validation procedure, the SME approved all the retranslated items, as well as the instruction texts. The final version of the STEU and the STEM are available as Supplementary Material. They are the versions utilized in the empirical validation study.

In that study's confirmatory factor analysis, the measures of each model fit for each test reached the required cutoff level for both absolute and relative fit indices. The χ^2-based fit indices all reached the literature standard of .9 (Baumgartner & Hombur, 1996) level, while the RMSEA and SRMSR measures reached the appropriate lower than .06 and lower than .08 levels, respectively (Hu & Bentler, 1999). The fit indices of the various CFA models can be seen in Table 2.

Table 2

Fit Indices of the Confirmatory Factor Analysis for the Situational Tests Initial and Final Versions, the Reduced Personality Factors Test, and the Satisfaction with Life Scale.

	STEU		STEM			
Measure	Initial	Final	Initial	Final	ER5FP	SWLS
χ^2	1822.77	777.96	813.97	290	131.3	1
DF	819	527	902	405	16	5
Scaled χ^2	1690.69	856.34	965.19	473.01	304.07	8.43
Scaled DF	819	527	902	405	16	5
χ^2 Scaling Factor	1.34	1.13	2.59	.90	.54	.12
Robust CFI	.886	.967	.851	.993	.992	1.000
Robust NNFI	.880	.965	.844	.993	.991	1.005
Robust RMSEA	.045	.032	.038	.015	.027	.011
CI 95 Lower	.042	.028	.000	.008	.022	.000
CI 95 Upper	.049	.036	.058	.020	.031	.024
SRMSR	.054	.044	.089	.037	.035	.014

Notes. ER5FP = Reduced Personality Factors Test. STEM = Situational Test of Emotional Management. STEU = Situational Test of Emotional Understanding. SWLS = Satisfaction with Life Scale. All analyses were run with the diagonally weighted least squares estimator with robust standard errors.

Situational Test of Emotional Understanding

Out of the original English version of the STEU's 42 items, ten items failed to reach the cutoff discrimination (*a*) or the difficulty (*b*) levels. The removal of those items resulted in mean parameters discrimination equal to 2.3 (SD 1.21) and difficulty equal to -.06 (1.04). Mean guessing levels (*g*) were .23 (.1), which is close to the .2 expected for items with five response categories. The parameters for each item can be viewed in Table 3, and the item fit parameters are displayed in Table 4.

Table 3

Item Response Theory Parameters for the Situational Test of Emotional Understanding.

Initial				Final				Initial				Final			
#	*a*	*b*	*g*	#	*a*	*b*	*g*	#	*a*	*b*	*g*	#	*a*	*b*	*g*
1	2.73	1.82	.16	1	2.99	1.83	.16	22	1.16	-2.29	.02	22	1.22	-1.92	.25
2	2.72	.65	.25	2	2.6	.62	.23	23	.16	1.04	.21				
3	1.03	-.71	.48	3	.87	-1.36	.3	24	1	-2.09	0	24	.95	-2.17	.01
4	3.2	.23	.28	4	2.92	.14	.24	25	2.92	-.15	.28	25	2.68	-.22	.25
5	2.56	.5	.2	5	2.37	.45	.17	26	.9	-1.58	0	26	.88	-1.6	0
6	.05	.52	.05					27	.8	-1.36	0	27	.74	-1.46	0
7	.71	-1.16	.03	7	.71	-1.07	.06	28	.64	-1.46	0				
8	2.55	-.28	.22	8	2.54	-.26	.24	29	1.18	-1.43	.09	29	1.17	-1.4	.11
9	.31	.34	0					30	.13	-4.69	.04				
10	1.11	-.96	.46	10	1.04	-1.32	.34	31	1.14	-.8	.11	31	1.13	-.77	.14
11	1.9	.84	.26	11	1.8	.78	.24	32	1.88	-.32	.21	32	1.88	-.31	.21
12	3.22	1.38	.25	12	3	1.38	.25	33	1.02	-.82	.37	33	.94	-1.18	.25
13	2.69	1.8	.58	13	3.53	1.88	.59	34	1.63	1.98	.64	34	1.92	2.1	.65
14	.19	-2.48	.16					35	.18	-.19	.01				
15	4.28	1.79	.24	15	4.84	1.81	.24	36	.4	-2.02	.01				
16	1.11	.52	.25	16	1.05	.43	.23	37	.77	-1.13	.01	37	.73	-1.18	0
17	1.92	.47	.21	17	1.91	.48	.21	38	6.75	1.7	.58	38	1.02	1.76	.58
18	2.36	.99	.27	18	2.27	.98	.26	39	.63	-1.92	.01				
19	2.8	.11	.26	19	2.67	.08	.24	40	.29	-4.01	.02				
20	5.68	-.23	.36	20	4.59	-.28	.33	41	3.49	.48	.26	41	3.29	.46	.25
21	3.74	.49	.26	21	3.59	.47	.25	42	.95	-1.04	.23	42	.88	-1.22	.19

Initial				Final				Initial				Final			
#	*a*	*b*	*g*	#	*a*	*b*	*g*	#	*a*	*b*	*g*	#	*a*	*b*	*g*

Note. *a* = discrimination, *b* = difficulty, *g* = guessing.

Table 4

Item Response Theory Item-Fit Measures for the Situational Test of Emotional Understanding.

Initial					Final				Initial					Final			
#	$S\chi^2$	df	*p*	RMSEA	$S\chi^2$	df	*p*	RMSEA	#	$S\chi^2$	df	*p*	RMSEA	$S\chi^2$	df	*p*	RMSEA
1	45.2	33	.08	.02	32.1	27	.23	.02	23	52.2	46	.25	.01	43.4	38	.25	.01
2	61.5	54	.23	.01	41.7	49	.76	.00	24	35.4	35	.45	.00	22.5	25	.60	.00
3	47.8	50	.56	.00	36.0	45	.83	.00	25	39.3	41	.55	.00	18.4	32	.97	.00
4	75.7	73	.39	.01	--	--	--	--	26	34.8	50	.95	.00	42.7	45	.57	.00
5	86.4	61	.02	.02	54.0	49	.29	.01	27	48.2	70	.98	.00	35.6	55	.98	.00
6	45.7	52	.72	.00	37.9	41	.61	.00	28	35.3	36	.50	.00	39.3	30	.12	.02
7	31.7	35	.63	.00	20.1	18	.32	.01	29	111.3	98	.17	.01	--	--	--	--
8	49.2	45	.31	.01	29.1	38	.85	.00	30	55.2	77	.97	.00	--	--	--	--
9	59.0	41	.03	.03	--	--	--	--	31	46.2	50	.63	.00	37.8	41	.62	.00
10	66.1	63	.37	.01	--	--	--	--	32	44.5	34	.11	.02	--	--	--	--
11	28.1	35	.79	.00	32.0	27	.23	.02	33	37.4	55	.97	.00	--	--	--	--
12	55.4	55	.46	.00	65.8	50	.07	.02	34	16.4	36	1.00	.00	29.8	32	.58	.00
13	61.9	63	.52	.00	--	--	--	--	35	62.9	80	.92	.00	--	--	--	--
14	70.9	71	.48	.00	--	--	--	--	36	19.6	24	.72	.00	22.5	19	.26	.02
15	29.6	36	.77	.00	44.9	30	.04	.03	37	27.8	40	.93	.00	34.1	33	.41	.01
16	52.6	54	.53	.00	--	--	--	--	38	28.2	39	.90	.00	42.9	30	.06	.02
17	58.0	55	.37	.01	44.6	46	.53	.00	39	13.2	10	.21	.02	18.9	10	.04	.04
18	34.4	40	.72	.00	--	--	--	--	40	74.0	60	.11	.02	--	--	--	--
19	70.2	53	.06	.02	48.3	48	.46	.00	41	36.2	38	.55	.00	49.2	53	.62	.00
20	49.5	48	.41	.01	63.2	45	.04	.02	42	45.1	50	.67	.00	25.5	31	.74	.00
21	98.7	70	.01	.02	--	--	--	--	43	35.1	33	.37	.01	51.3	46	.28	.01
22	82.0	76	.30	.01	67.2	64	.37	.01	44	22.8	24	.53	.00	33.5	20	.03	.03

Note. $S\chi^2$ = signed test chi-square.

Both models failed the M^{2*} test, as is common with χ^2-type measures, but the more important RMSEA level was adequate for the final model, and the final model reached the adequate absolute fit indices levels of .9 (Hu & Bentler, 1999). That model was also significantly better than the initial model, $\Delta\chi^2$ (345) = 1466.90, $p < .001$. The *Zh* measure was normally distributed and, in both models, reached acceptable levels of less than .5. These goodness-of-fit indices can be seen in Table 5.

Table 5

Item Response Theory Fit Indices for All Tests and Comparison Between the Models of the Situational Tests.

	STEU		STEM		ER5FP					
Model	Initial	Final	Initial	Final	C	O	A	N	E	SL
M^2*/C^2	2216.6	749.7	1453.8	408.5	.1	2.3	16.2	14.8	6.3	9.8
df	777	432	854	375	2	2	2	2	2	5
p (*M²**)	< .001	< .001	< .001	.112	.938	.321	< .001	.001	.043	.080
RMSEA	.052	.033	.032	.011	.000	.014	.102	.090	.056	.038
IC Lower	.049	.029	.029	.011	.000	.000	.060	.055	.009	.000
IC Higher	.054	.037	.035	.018	.018	.075	.150	.145	.106	.092
SRMSR	.061	.045	.061	.045	.010	.018	.032	.035	.020	.021
NNFI	.818	.948	.988	.998	1.003	.999	.976	.964	.990	.997
CFI	.836	.955	.989	.998	1.000	1.000	.992	.988	.997	.998
$\Delta\chi^2$		1466.9		1043.26						
Δdf		345		479						
p ($\Delta\chi^2$)		< .001		< .001						
Mean *Zh*	.20	.21	.08	.11	.24	.28	.22	.25	.30	.21
(SD)	(1.03)	(.95)	(1.40)	(1.09)	(.89)	(.94)	(.94)	(.90)	(.90)	(1.13)

Notes. STEU = Situational Test of Emotional Understanding, STEM = Situational Test of Emotional Management, ER5FP = Reduced Scale of Personality Factors, SWLS = Satisfaction with Life Scale, C = Conscientiousness, O = Openness, A = Amiability, N = Neuroticism, E = Extroversion. The 3-parameter model was employed for the STEU, the graded partial credit model for the STEM, and the graded response model for the ER5FP and SWLS.

[a] The C^2 statistic was used in place of the M^2* statistic for the tests ER5FP and SWLS because these tests had fewer degrees of freedom than required.

Situational Test of Emotional Management

Fourteen out of the initial test's 44 items had to be dropped for failing to reach the discrimination or difficulty cutoff values. The remaining 30 items had a mean discrimination of 1.14 (SD .34), and mean difficulty (*b*) of -1.27. All guessing parameters were estimated to be

close to zero. The item parameters can be viewed in Table 6 and the item fit indices are displayed in Table 7.

Table 6

Item Response Theory Parameters for the Situational Test of Emotional Management.

	Initial						Final				
#	*a*	b_1	b_2	b_3	Mean *b*	#	*a*	b_1	b_2	b_3	Mean *b*
1	1.09	.21			.21	1	1.07	.20			.20
2	.85	-2.68	-1.76	3.13	-.44	2	.83	-2.77	-1.83	3.20	-.47
3	.99	-2.03	-.93		-1.48	3	.96	-2.10	-.97		-1.54
4	.10	1.31	-1.80	-22.03	-7.51						
5	.77	-1.59	-.81		-1.20	5	.75	-1.64	-.85		-1.24
6	.79	-2.62	2.34		-.14	6	.78	-2.68	2.36		-.16
7	1.09	-3.13	1.18		-.97	7	1.06	-3.21	1.19		-1.01
8	1.00	-2.05	-1.53		-1.79	8	.97	-2.11	-1.58		-1.85
9	.34	-8.02	5.59		-1.22						
10	.13	14.01	-4.79		4.61						
11	1.47	-.79			-.79	11	1.40	-.84			-.84
12	.88	-1.25	-1.67		-1.46	12	.87	-1.29	-1.71		-1.50
13	.66	-1.90	-.23	-3.32	-1.81						
14	.20	-2.24	-4.39		-3.31						
15	1.01	-1.75	-2.48		-2.11	15	1.00	-1.82	-2.52		-2.17
16	.20	-1.18	5.80		-2.19						
17	1.08	-.81	-.99		-.90	17	1.07	-.85	-1.02		-.94
18	.29	-4.61			-4.61						
19	.71	.13	-2.83		-1.35	19	.70	.12	-2.92		-1.40
20	1.00	-2.61	.72		-.94	20	.98	-2.69	.73		-.98
21	.11	-9.04	4.23	-17.39	-7.40						
22	.73	-.94	-2.53	1.01	-.82	22	.71	-.99	-2.62	1.02	-.86
23	1.13	-2.61	.69		-.96	23	1.12	-2.66	.69		-.99
24	1.74	-2.22	2.06		-.08	24	1.73	-2.27	2.07		-.10
25	1.08	-2.17	-1.58		-1.87	25	1.07	-2.23	-1.62		-1.92
26	.98	-1.62	-1.79		-1.70	26	.96	-1.66	-1.84		-1.75
27	.84	-2.64	-.55	-.01	-1.07	27	.84	-2.69	-.58	-.02	-1.10
28	1.10	-1.98			-1.98	28	1.04	-2.10			-2.10
29	.13	2.96	-5.00	-3.85	-1.96						
30	.10	22.52	-3.98	1.08	6.54						
31	.76	-3.05	2.76		-.14	31	.74	-3.15	2.82		-.16
32	.67	3.33	-6.07		-1.37						
33	.18	-.22	-12.46		-6.34						
34	1.45	-1.43	-1.93		-1.68	34	1.45	-1.48	-1.96		-1.72
35	.08	25.17	.38	-5.95	6.53						
36	3.62	-1.36	-1.32		-1.34	36	3.61	-1.41	-1.36		-1.38
37	1.18	-2.37	-.92		-1.64	37	1.17	-2.43	-.95		-1.69

Initial						Final					
#	*a*	b_1	b_2	b_3	Mean *b*	#	*a*	b_1	b_2	b_3	Mean *b*
38	1.25	-2.24	-1.18		-1.71	38	1.23	-2.30	-1.21		-1.76
39	2.24	-2.44			-2.44	39	2.22	-2.50			-2.50
40	.81	-.20	-3.53	.96	-.92						
41	.89	-3.18	-.55		-1.86	40	.78	-.20	-3.67	.97	-.97
42	.86	-1.61	-2.11		-1.86	41	.86	-3.27	-.58		-1.93
43	.18	-1.87			-1.87	42	.83	-1.67	-2.19		-1.93
44	1.69	-.33	-2.62		-1.48	44	1.61	-.33	-2.75		-1.54

Notes. *a* = discrimination, *b* = difficulty. All guessing values estimated near zero. *Mean b* is the arithmetic mean of the *b* values.

Table 7

Item Response Theory Item-Fit Measures for the Situational Test of Emotional Management.

	Initial				Final					Initial				Final			
#	Sχ^2	*df*	*p*	RMSEA	Sχ^2	*df*	*p*	RMSEA	#	Sχ^2	*df*	*p*	RMSEA	Sχ^2	*df*	*p*	RMSEA
1	23.2	23	.45	.00	14.9	19	.73	.00	22	30.3	18	.03	.03	11.7	14	.63	.00
2	29.2	23	.17	.02	23.9	18	.16	.02	23	35.0	26	.11	.02				
3	20.3	22	.57	.00	16.2	19	.64	.00	24	21.3	21	.44	.00	16.5	17	.49	.00
4	22.6	21	.37	.01	9.8	15	.83	.00	25	26.0	19	.13	.02	18.0	16	.32	.01
5	22.9	22	.41	.01	18.7	18	.41	.01	26	20.3	23	.62	.00	15.2	19	.71	.00
6	29.5	26	.29	.01					27	31.1	24	.15	.02	18.5	20	.56	.00
7	27.7	25	.32	.01	23.5	20	.27	.02	28	30.7	25	.20	.02				
8	25.2	19	.15	.02	20.0	16	.22	.02	29	11.4	21	.95	.00	20.3	17	.26	.02
9	29.6	25	.24	.02					30	29.9	26	.27	.01				
10	23.7	22	.36	.01	20.0	17	.27	.02	31	17.6	22	.73	.00	12.2	19	.88	.00
11	32.7	24	.11	.02	15.1	20	.77	.00	32	22.3	21	.38	.01	15.3	17	.58	.00
12	24.6	24	.43	.01	28.2	19	.08	.03	33	16.3	23	.84	.00	13.6	19	.81	.00
13	39.6	26	.04	.03	14.7	20	.79	.00	34	28.9	26	.32	.01	17.1	21	.71	.00
14	33.2	26	.16	.02					35	41.2	26	.03	.03				
15	20.1	25	.74	.00	25.2	20	.19	.02	36	28.3	26	.35	.01				
16	31.0	24	.15	.02	24.3	20	.23	.02	37	30.6	24	.17	.02	23.5	20	.27	.02
17	18.3	22	.69	.00	22.5	18	.21	.02	38	26.9	25	.36	.01	20.0	20	.46	.00
18	44.7	24	.01	.04	29.2	20	.08	.03	39	19.6	25	.77	.00				
19	22.3	21	.38	.01	18.7	16	.28	.02	40	18.2	26	.87	.00				
20	12.0	16	.75	.00	8.9	13	.78	.00	41	15.9	21	.78	.00	13.1	18	.78	.00
21	21.8	21	.41	.01	15.4	18	.63	.00	42	18.7	23	.72	.00	18.7	19	.48	.00

Note. All *u*-parameter values equal to 1.

Only the initial model failed the M^2* test, which is supported by the low RMSEA levels of the final model. The final model also had marginally better fit indices across the board. The

Zh measure was normally distributed and, in both models, reached acceptable levels of less than .5. The model fit indices can be viewed on Table 5.

Reduced Personality Factors Test

Again, all parameters from the items of the ER5FP reached acceptable levels. Mean discrimination was 2.89 (SD .96). Each item had five difficulty parameters estimated. The mean difficulty of these means was .26 (.43). The mean of the standard deviation between the difficulties was 1.15 (.21). All item parameters for the ER5PF scale are displayed in Table 8.

Table 8

Item Response Theory Parameters for the Reduced Personality Factors Test (ER5FP).

#	*a*	b_1	b_2	b_3	b_4	b_5	Mean *b*
			Conscientiousness				
1	4.5	-2.3	-1.2	-0.4	1.1	1.5	-0.3
2	6.1	-1.4	-0.9	-0.1	1.1	1.7	0.1
3	2.2	-2.7	-1.3	-0.4	1.7	2.5	-0.1
4	4.1	-1.8	-1	-0.1	1.2	1.9	0.1
			Openness				
5	1.5	-2.3	-1.2	-0.2	2.1	3	0.3
6	2.6	-1.6	-0.5	0.4	2	2.8	0.6
7	2.4	-2	-1.2	0	1.8	2.5	0.2
8	1.3	-2.4	-0.8	0.5	2.6	3.8	0.8
			Amiability				
9	2.7	-2.7	-1	-0.1	1.4	2	-0.1
10	2.8	-2.9	-2.2	-1.1	1.3	2	-0.6
11	3.1	-2.8	-1.6	-0.6	1.5	2.2	-0.3
12	5.4	-2.3	-1.5	-0.5	1.4	2.1	-0.2
			Neuroticism				
13	2.9	-2.4	-1.6	-0.8	1.4	1.6	-0.4
14	3.1	-2.7	-1.1	-0.3	1.5	2.1	-0.1
15	1.6	-1.7	-0.5	0.4	2.4	3.3	0.8
16	1.4	-1.4	-0.1	1.1	3	3.7	1.2
			Extroversion				
17	2.6	-0.9	-0.1	1	2.5	3.5	1.2
18	3.5	-1.6	-0.6	0	1.6	2.1	0.3
19	2.9	-1.7	-0.7	0	1.7	2.4	0.3
20	1.2	-2.1	-0.9	0.9	3.2	4.7	1.2

Fit measures were calculated for each of the five IRT models that were run for the ER5FP. Because the models had few degrees of freedom, the C^2 statistic was calculated instead of the M^2*. Only the models for Amiability and Neuroticism failed the C^2 statistic test. These models also had residuals slightly over the level considered adequate by the literature when considering the root mean squared error of approximation, RMSEA > .06, but they did reach acceptable levels according to the standardized root mean squared residual, SRMSR < .08. Additionally, all models had adequate fit according to relative fit indices CFI and NNFI. These indices are displayed on Table 4.

Satisfaction with Life Scale

The parameters from the items of the SWLS all had satisfactory values. Mean discrimination was 3.45 (SD 1.13). Each item had five difficulty parameters estimated. The mean difficulty of these means was .22 (.32). The mean of the standard deviation between the difficulties was 1.5 (.16). The parameters of the SWLS are displayed in Table 7.

Table 9

Item Response Theory Parameters for the Satisfaction with Life Scale (SWLS).

#	*a*	b_1	b_2	b_3	b_4	b_5	Mean *b*
1	4.36	-1.36	-.92	-.11	1.53	2.69	.36
2	2.55	-2.20	-1.44	-.32	1.65	2.78	.09
3	5.38	-1.35	-.88	-.22	1.44	2.37	.27
4	2.83	-3.05	-1.65	-.96	1.29	2.04	-.47
5	2.15	-1.11	-.25	.41	2.10	2.99	.83

The SWLS model also had few degrees of freedom, the reason for which the C^2 statistic was employed in place of the M^2*. The model was approved, as the statistical test did not reject the null hypothesis. All goodness of fit indices were considered acceptable, C^2 (5) = 9.843, p = .08, RMSEA = .038 [CI95 0; .072], SRMSR = .021, TLI = .997, CFI = .998, Mean *Zh* = .21 (SD 1.13).

Finally, Pearson correlation between the factors varied greatly—only nine were significant at the $p < .05$ level. STEU scores represent seven of these correlations, as the test

was correlated with all other scores, but only the correlation with STEM was larger than a small correlation, $r = .501$, $p < .001$. The only other practically significant correlation is between Conscientiousness and Neuroticism, $r = -.228$, $p = .032$. The remaining correlation coefficients, significant and not, can be viewed in Table 8.

Table 10

Pearson Correlation Coefficients Between the Calculated Test Scores.

Factor	C	O	A	N	E	STEM	STEU	LS
C	–							
O	.123	–						
A	.214	.120	–					
N	-.228*	.028*	-.160	–				
E	-.118	-.006	-.132	.265	–			
STEM	.030	-.141	-.071	-.021	.048	–		
STEU	.005**	.091**	.029**	-.048**	.088**	.501**	–	
SWLS	.408	.018	.348	-.453*	-.274	.102	.058**	–

Note. C = Conscientiousness, O = Openness, A = Amiability, N = Neuroticism, E = Extroversion, STEU = Situational Test of Emotional Understanding, STEM = Situational Test of Emotional Management, SWLS = Satisfaction with Life Scale.

* $p < .05$ ** $p < .001$

Descriptive Statistics

The scores were calculated based on the expected a posteriori (EAP) method, configured to use a normally distributed prior, which is the default configuration (Rosseel, 2012). For this reason, the distributions of scores were expected to be similar. This was observed, and the standard errors were also similar, ranging from .29 to .44. The arithmetic mean of all distributions was approximately zero. These statistics are displayed in Table 9.

Table 11

Descriptive Statistics for the Distribution of Each Calculated Test Score.

Test	Mean	SD	P25	Med	P75	IQR	SE
STEU	.00	.93	-.68	-.07	.66	1.33	.37
STEM	.00	.93	-.58	.01	.57	1.15	.35
Conscientiousness	.00	.96	-.62	-.03	.46	1.08	.27
Openness	.00	.90	-.60	-.04	.54	1.14	.44
Amiability	.00	.94	-.60	.04	.39	.99	.34
Neuroticism	.00	.91	-.62	.01	.52	1.15	.42
Extraversion	.00	.93	-.67	.02	.59	1.26	.37
SWLS	.00	.96	-.63	-.03	.61	1.24	.29

Note. Med = median, P25 = percentile 25, P75 = percentile 75, Sample sizes = 688, SD = standard deviation, STEU = Situational Test of Emotional Understanding, STEM = Situational Test of Emotional Management.

Discussion

The aim of this study was to provide validity evidence for the Brazilian version of the STEU and the STEM. This task encompassed two studies: one adaptation study, and one validation study. Although often considered separately, the adaptation process ultimately serves as a content validation process.

In line with Mislevy (2007)'s concept of *validity by design*, the steps taken constitute validity evidence based on test content (AERA et al., 2014). The validation by SMEs of the original translation and of the retranslation (Sireci et al., 2006) mirrors the traditional method employed by test designers to validate originally constructed items. In addition to this type of validity evidence, the empirical portion of this study characterizes two other sources of validity evidence: evidence based on internal structure, and evidence based on relationships to other variables (traditionally known as "predictive" validity; AERA et al., 2014)

Internal structure evidence includes fit indices from latent variable modeling, such as through IRT, and structural equation modeling, such as through CFA. To that end, this study has provided internal structure evidence for all of the four tests that were administered. While several items underperformed, a definitive version of the STEU and the STEM, containing 32 and 30 items, respectively, reached the required fit indices, both globally and on a per item

basis. The other two tests, which were administered as criteria, also exhibited favorable fit indices, both globally and per item.

The second category of evidence, based on the relationships between the test scores and other variables, can also be demonstrated through structural equation modeling with the inclusion of variables external to the test scores and through other methods dedicated to the association of variables. Instead, in this study, relationships between the test and external variables were evaluated using correlation coefficients analyses, which is also a traditional method (Campbell & Fiske, 1959).

As hypothesized, there were significant correlations between the STEU and the STEM. However, while STEM test scores were significantly correlated with SWLS scores, no correlation was found between STEU scores and SWLS scores. This differs from MacCann and Robert's (2008) study, which found relationships between both STEM and STEU scores and Amiability, as well as between STEM and SWLS scores. However, all correlations were of low magnitude, and the absence of a strong relationship in the original study may account for the lack of a relationship in these data. While there have been studies that reported a correlation between EI measures and self-reported life satisfaction (Mayer et al., 2008; Schneider & McGrew, 2018), the fact that such correlations are only practically significant with trait EI measures suggests little value to the use of the SWLS as validity evidence for the STEU and the STEM. Future studies should address whether the deficiency in the relationship is specific to the SWLS, or with other life satisfaction measures.

Although this study failed to confirm the relationship between EI and the SWLS, this study did confirm the lack of practically relevant relationships between the Big Five personality factors and the STEM and STEU. This provides validity evidence based on the discrimination with other variables (Campbell & Fiske, 1959), which supports the notion that ability EI cannot be reduced to personality traits. The ongoing debate over the perspective to adopt in the study of

EI has led to arguments in favor of a trait perspective, with structural equation and confirmatory models associating EI and personality, notably by Petrides et al. (2016). However, such studies use trait measures of EI, as they acknowledge. The results of this study align more closely with those of MacCann et al. (2016), in which an EI measure based on the ability EI perspective was not substantially associated with personality factors.

While the results of this study primarily depend on discriminative validity evidence, MacCann and Roberts (2008)'s original study also provided extensive predictive validity evidence that the present study did not seek to replicate. For example, in their study, the authors displayed positive correlations between the STEU and measures such as the Story test (r = .40) and the Vocabulary test (r = .49), regarded to be the best general intelligence predictors. Both tests were also correlated with the grades of Psychology students (r = .42 and r = .34, respectively; MacCann & Roberts, 2008).

That study also showed that STEM scores were significantly correlated with the Vocabulary test (r = -.41; MacCann & Roberts, 2008). Furthermore, both the STEU and the STEM were negatively correlated with the externally oriented thinking style factor of the Toronto Alexithymia Scale's (Bagby et al., 1994), with correlation coefficients of r = -.38 and r = -.43, respectively. This provides evidence of the relationship between the test scores and constructs closely associated with ability EI.

Instead of replicating MacCann and Roberts's (2008) results, the present study aimed to provide novel validity evidence for the internal structure of the tests through the use of CFA and IRT, which the authors themselves had suggested. Although several items had to be removed to achieve acceptable levels of model fit, a substantial proportion of items for both the STEU (32 out of 42) and STEM (30 out of 44) tests exhibited favorable IRT item fit, discrimination and difficulty parameters, as well as favorable model fit for both the CFA and IRT.

At the same time, there are some limitations to this study that should be considered. Since no technique was employed to record the response process, it is not possible to determine the reasons for the deficient performance of the items that were removed. In the context of cross-cultural adaptation, it is possible that items designed within a specific social situation may have cultural aspects that do not translate reliably. In any case, further research is needed to draw definite conclusions. In addition to response process data, these studies could also draw on the extensive theoretical framework that underpinned the construction of the tests (Motowidlo et al., 1990; Roseman, 2001), as well as on the cross-cultural psychology literature's description of this versus the original sample's emotionality, namely, Brazilian and Australian emotionality.

Another possible path to investigate and possibly control the effects of external variables that may have influenced answers to those tests is multidimensional IRT (MIRT). For example, Primi et al. (2019) showed that is possible to use a MIRT model to statistically control for effects that may have impacted participants' answers, such as acquiescence bias. The authors also showed that validity evidence remained adequate in the sample's corrected scores.

Finally, one additional concern exists regarding whether the full amplitude contained within the limits of the construct's definition are explored by the EI measures available. For example, while MacCann and Roberts (2008) sought to provide a balanced sample of emotions, as described by Roseman (2001), emotion expression varies widely in intensity. While it can be inferred that emotion management under different emotion intensity would be related strictly to a difficulty, meaning it would simply require more emotional intelligent to manage more intense emotions, it is possible that intense emotions are qualitatively different. Future studies should investigate whether EI measures are underrepresenting the construct in their content.

In any case, the present study showed that the STEU and the STEM were successfully adapted to a Brazilian context using a traditional backtranslation method. Dimension and item analysis yielded acceptable fit indices on an individual item basis and for the test as a whole.

Correlation indices largely confirmed initial expectations, but the low correlation observed with life satisfaction was not found to be practically significant. While further research is necessary to situate the examined constructs within their nomological network (Cronbach & Meehl, 1955) and confirm their proximity to related constructs, the validity evidence obtained thus far suggests test adequacy.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing* (4th ed.). AERA Publications.

Armstrong, R. D., Stoumbos, Z. G., Mabel, T. K., &, Shi, M. (2007). On the Performance of the *Iz* Person-Fit Statistic, *Practical Assessment, Research, and Evaluation, 12*, Article 16. https://doi.org/10.7275/xz5d-7j62

Bar-On, R. (1997). *The emotional quotient inventory (EQ-i): A test of emotional intelligence*. Multi-Health Systems.

Baumgartner, H., & Hombur, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing, 13*(2), 139–161. https://doi.org/10.1016/0167-8116(95)00038-0

Brackett, M. A., & Salovey, P. (2006). Measuring emotional intelligence with the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT). *Psicothema, 18*(SUPPL.), 34–41. Retrieved from https://www.psicothema.com/pdf/3273.pdf.

Brislin, R. W. (1970). Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216. https://doi.org/10.1177/135910457000100301

Bru-Luna, L. M., Martí-Vilar, M., Merino-Soto, C., & Cervera-Santiago, J. L. (2021). Emotional Intelligence Measures: A Systematic Review. *Healthcare, 9*(12), Article 1696. https://doi.org/10.3390/healthcare9121696

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*(2), 245–276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Cai, L., & Monroe, S. (2014). *CRESST Report 839: A New Statistic for Evaluating Item Response Theory Models for Ordinal Data* (ERIC Document № ED555726). ERIC Database.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cobêro, C., Primi, R., Muniz, M. (2006). Inteligência emocional e desempenho no trabalho: um estudo com MSCEIT, BPR-5 e 16PF. *Paidéia (Ribeirão Preto), 16*(35), 337–348. https://doi.org/10.1590/S0103-863X2006000300005

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin, 52*, 281–302. Retrieved from https ://uopsych.github.io/psy611/readings/Cronbach_Meehl_1955.pdf.

Diener, E., Emmons, R. A., Larsen, R. J., Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment, 49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Gignac, G. E. (2005). Evaluating the MSCEIT V2.0 via CFA: Comment on Mayer et al. (2003). *Emotion, 5*(2), 233–235. https://doi.org/10.1037/1528-3542.5.2.233

Gonzaga, A. R., & Monteiro, J. K. (2011). Inteligência emocional no Brasil: um panorama da pesquisa científica. *Psicologia: Teoria e Pesquisa, 27*(2), 225–232. https://doi.org/10.1590/S0102-37722011000200013

Gouveia, V. V., Chaves, S. S. da S., de Oliveira, I. C. P., Dias, M. R., Gouveia, R. S. V., & de Andrade, P. R. (2003). A utilização do QSG-12 na população geral: estudo de sua validade de construto. *Psicologia: Teoria e Pesquisa, 19*(3), 241–248. https://doi.org/10.1590/S0102-37722003000300006

Grubb III, W., & McDaniel, M. (2007). The Fakability of Bar-On's Emotional Quotient Inventory Short Form: Catch Me if You Can. *Human Performance, 20*(1), 43–59. https://doi.org/10.1080/08959280709336928

Guenole, N., Chernyshenko O. S., & Weekly, J. (2017). On Designing Construct Driven Situational Judgment Tests: Some Preliminary Recommendations. *International Journal of Testing, 17*(3), 234–252. https://doi.org/10.1080/15305058.2017.1297817

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/BF02288892

Hogan, T. P. (2007). *Psychological Testing: A Practical Introduction* (2nd ed.). Wiley.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

International Test Commission. (2017). *ITC Guidelines for Translating and Adapting Tests* (2nd ed.). https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

Jackson, J. E. (2014). Varimax Rotation. *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat05679

Junior, A. G., & Noronha, A. P. (2008). Parâmetros psicométricos do Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT). *Psic: revista da Vetor Editora, 9*(2), 145–153.

http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1676-73142008000200003&lng=pt&tlng=pt.

Kang, T., & Chen, T. T. (2008). Performance of the generalized $S–\chi^2$ item-fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391–406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Keele, S. M., & Bell, R. C. (2009). Consensus scoring, correct responses, and reliability of the MSCEIT V2. *Personality and Individual Differences, 47*(7), 740–747. https://doi.org/10.1016/j.paid.2009.06.013

Lievens, F., & Motowidlo, S. J. (2016). Situational Judgment Tests: From Measures of Situational Judgment to Measures of General Domain Knowledge. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 9*(1), 3–22. https://doi.org/10.1017/iop.2015.71

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum.

MacCann, C., &, Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, *8*(4), 540–551. https://doi.org/10.1037/a0012746

MacCann, C., Joseph, D. L., Newman, D. A., & Roberts, R. D. (2014). Emotional intelligence is a second-stratum factor of intelligence: Evidence from hierarchical and bifactor models. *Emotion*, *14*(2), 358–374. https://doi.org/10.1037/a0034755

Marin, A. H., Silva, C. T., Andrade, E. I. D., Bernardes, J., & Fava, D. C. (2017). Competência socioemocional: conceitos e instrumentos associados. *Revista Brasileira de Terapias Cognitivas*, *13*(2), 92–103. https://doi.org/10.5935/1808-5687.20170014

Mayer, J. D., Salovey, P., & Caruso, D. R. (1997). *Emotional IQ test* [CD-ROM]. Virtual Knowledge.

Mayer, J. D., Caruso, D. R., & Salovey, P. (2000). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, *27*, 267–298. https://doi.org/10.1016/S0160-2896(99)00016-1

Mayer, J. D., Salovey, P., & Caruso, D. (2002). *Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT© V2.0) User's Manual*. MHS Publishers.

Mayer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology*, *59*, 507–536. https://doi.org/10.1146/annurev.psych.59.103006.093646

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion, 1*(3), 232–242. https://doi.org/10.1037/1528-3542.1.3.232

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion*, *3*(1), 97–105. https://doi.org/10.1037/1528-3542.3.1.97

McGrew, K. S. (2017). *Cattell–Horn–Carroll (CHC) theory of cognitive abilities (v2.5) "official" broad and narrow definitions*. IQ's Corner. Published in July 20th, 2017. Accessed on May 2nd, 2023. http://www.iqscorner.com/2017/07/cattell-horn-carroll-chc-theory-of.html

Miguel, F. K. (2016). *Bateria Online de Inteligência Emocional. Livro de Instruções*. Vetor.

Mislevy, R. J. (2007). Validity by Design. *Educational Researcher, 36*(8), 463–469. https://doi.org/10.3102/0013189X07311660

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*(6), 640–647. https://doi.org/10.1037/0021-9010.75.6.640

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Muraki, E., & Carlson, E. B. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*(1), 73–90. https://doi.org/10.1177/014662169501900109

O'Connor, P. J., Hill, A., Kaya, M., & Martin, B. (2019). The Measurement of Emotional Intelligence: A Critical Review of the Literature and Recommendations for Researchers and Practitioners, *Frontiers in Psychology, 10*, Article 1116. https://doi.org/10.3389/fpsyg.2019.01116

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64. https://doi.org/10.1177/01466216000241003

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of *S*-χ^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289–298. https://doi.org/10.1177/0146621603027004004

Palmer, B. R., Gignac, G., Manocha, R., & Stough, C. (2005). A psychometric evaluation of the Mayer-Salovey-Caruso Emotional Intelligence Test Version 2.0. *Intelligence, 33*(3), 285–305. https://doi.org/10.1016/j.intell.2004.11.003

Passos, M. F. D., & Laros, J. (2015). Construção de uma escala reduzida de Cinco Grandes Fatores de personalidade, *Avaliação Psicológica, 2015, 14*(1), 115–123. https://doi.org/10.15689/ap.2015.1401.13

Petrides, K. V., Siegling, A. B., & Saklofske, D. H. (2016). Theory and measurement of trait emotional intelligence. In U. Kumar (Ed.), *The Wiley Handbook of Personality Assessment* (pp. 90–103). Wiley Blackwell. https://doi.org/10.1002/9781119173489.ch7

Primi, R., Hauck-Filho, N., Valentini, F., Santos, D., Falk, C. F. (2019). Controlling Acquiescence Bias with Multidimensional IRT Modeling. In M. Wiberg, S. Culpepper, R. Janssen, J. González, D. Molenaar. (Eds.). *Quantitative Psychology: 83rd Annual Meeting of the Psychometric Society, New York, NY, 2018*. Springer. https://doi.org/10.1007/978-3-030-01310-3_4

Raiche, G., Walls, T. A., Magis, D., Riope, M., & Blais, J. G. (2013). Non-graphical solutions for Cattell's scree test, *Methodology, 9*(1), 23–29. https://doi.org/10.1027/1614-2241/a000051

Revelle, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research* [R package version 2.3.3]. Northwestern University. https://CRAN.R-project.org/package=psych.

Roseman, I. (1979). *Cognitive aspects of emotions and emotional behavior* [Paper presentation]. 87th Annual Convention of the American Psychological Association, New York, NY, United States of America. https://www.researchgate.net/publication/245683721_Cognitive_aspects_of_emotion_and_emotional_behavior.

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. In P. Shave (Ed.), *Review of Personality & Social Psychology* (Vol. 5, pp. 11–36). Sage Publications. https://www.researchgate.net/publication/245683603_Cognitive_Determinants_of_Emotion_A_Structural_Theory

Roseman, I. J. (2001). A model of appraisal in the emotion system: Integrating theory, research, and applications. In K. R. Scherer & A. Schorr (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 68–91). Oxford University Press.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rossen, E., Kranzler, J. H., & Algina, J. (2008). Confirmatory factor analysis of the Mayer–Salovey–Caruso Emotional Intelligence Test V 2.0 (MSCEIT). *Personality and Individ2023·06·232023·06·23ual Differences, 44*(5), 1258–1269. https://doi.org/10.1016/j.paid.2007.11.020

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, Cognition and Personality*, *9*(3), 185–211. https://doi.org/10.2190/DUGG-P24E-52WK-6CDG

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(SUPPL.1), 1–97. https://doi.org/10.1007/BF03372160

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In D. P. Flanagan, E. M. McDonough. (Eds.). *Contemporary Intellectual Assessment: Theories, Tests and Issues* (3rd ed.). Guildford Publications.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In D. P. Flanagan, E. M. McDonough. (Eds.). *Contemporary Intellectual Assessment: Theories, Tests and Issues* (4th ed.). Guildford Publications.

Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. J. (2006). Evaluating Guidelines for Test Adaptations: A Methodological Analysis of Translation Quality. *Journal of Cross-Cultural Psychology, 37*(5), 557–567. https://doi.org/10.1177/0022022106290478

Vieira-Santos, J., Lima, D. C., Sartori, R. M., Schelini, P. W., & Muniz, M. (2018). Inteligência emocional: revisão internacional da literatura. *Estudos Interdisciplinares em Psicologia*, *9*(2), 78–99. https://doi.org/10.5433/2236-6407.2018v9n2p78

Watkins, M. W. (2018). Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology, 44*(3), 219–246. https://doi.org/10.1177/0095798418771807

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

TESTE SITUACIONAL DE COMPREENSÃO EMOCIONAL
SITUATIONAL TEST OF EMOTIONAL UNDERSTANDING

Carolyn MacCann
Victor Vasconcelos
Cristiane Faiad

Questão 1: Uma experiência agradável termina inesperadamente e não há muito o que fazer a respeito. É mais provável que a pessoa envolvida sinta...?

Vergonha	Angústia	Raiva	Tristeza	Frustração

Questão 2: Xavier termina uma tarefa difícil dentro do prazo e abaixo do orçamento. É mais provável que Xavier sinta...?

Surpresa	Orgulho	Alívio	Esperança	Alegria

Questão 3: Um vizinho irritante de Eva se muda para outro estado. É mais provável que Eva sinta...?

Arrependimento	Esperança	Alívio	Tristeza	Alegria

Questão 4: O tempo está ótimo no dia em que Júlia vai fazer um piquenique ao ar livre. É mais provável que Júlia sinta...?

Orgulho	Alegria	Alívio	Culpa	Esperança

Questão 5: O arrependimento ocorre mais provavelmente quando...?

Os eventos ocorrem de forma inesperada.	Você causou algo que não queria que acontecesse e não pode mudá-lo.	As circunstâncias causaram algo que você não queria que acontecesse.	Você causou algo que não queria que acontecesse e está tentando mudá-lo.	Os eventos começam a ficar fora do seu controle.

Questão 6: A colega de trabalho de Edna organiza uma festa de despedida para Edna, que sairá de férias. É mais provável que Edna sinta...?

Surpresa	Gratidão	Orgulho	Esperança	Alívio

Questão 7: Uma coisa desagradável está acontecendo. Nem a pessoa envolvida, nem qualquer outra pessoa pode fazer com que isso pare. É mais provável que a pessoa envolvida sinta...?

Culpa	Angústia	Tristeza	Medo	Raiva

Questão 8: Se a situação atual continuar, o empregador de Denise provavelmente poderá mudar seu emprego para um local muito mais próximo de sua casa, o que ela quer muito. É mais provável que Denise sinta...?

Angústia	Alegria	Surpresa	Esperança	Medo

Questão 9: Sônia descobre que uma amiga dela pegou dinheiro emprestado de outras pessoas para pagar contas urgentes, mas na verdade usou o dinheiro para fins menos sérios. É mais provável que Sônia sinta...?

© Carolyn MacCann, 2008
& Victor Vasconcelos, 2023

Raiva	Entusiasmo	Desprezo	Vergonha	Terror

Questão 10: É mais provável que alguém se sinta surpreso depois de?

Algo inesperado acontecer.	Algo estranho acontecer.	Algo incomum acontecer.	Algo assustador acontecer.	Algo bobo acontecer.

Questão 11: Ieda trabalha como solucionadora de problemas. Apresentam a ela a um problema que inicialmente parece comum, mas Ieda não consegue descobrir como resolvê-lo. É mais provável que Ieda se sinta...?

Confusa	Frustrada	Surpresa	Aliviada	Angustiada

Questão 12: Carlos vai encontrar um amigo para ver um filme. O amigo está muito atrasado e eles não conseguirão chegar ao cinema a tempo. É mais provável que Carlos se sinta...?

Deprimido	Frustrado	Com raiva	Desprezado	Angustiado

Questão 13: Ricardo precisa cumprir uma cota antes de sua avaliação de desempenho. Há apenas uma pequena chance que ele será capaz de atingir a meta e não há muito que ele possa fazer para melhorar o resultado. É mais provável que Ricardo se sinta...?

Irritado	Assustado	Angustiado	Triste	Esperançoso

Questão 14: Maira acredita que outra pessoa a prejudicou de propósito. Não há muito o que fazer para melhorar as coisas. É mais provável que Maira sinta...?

Desgosto	Raiva	Ciúmes	Surpresa	Ansiedade

Questão 15: Bartolomeu pede ao seu colega de trabalho, Felipe, que minta sobre o dinheiro que Bartolomeu está roubando da empresa, para não implicar Bartolomeu. Felipe não concorda. É mais provável que Felipe sinta...?

Entusiasmo	Raiva	Terror	Desprezo	Vergonha

Questão 16: Tiago gosta de passar os sábados brincando com seus filhos no parque. Este ano eles têm atividades esportivas aos sábados e não podem mais ir ao parque com ele. É mais provável que Tiago se sinta...?

Irritado	Triste	Frustrado	Angustiado	Vergonha

Questão 17: Se tudo correr bem, é bem provável que a casa de Davi aumente de valor. É mais provável que Davi sinta...?

Angústia	Medo	Surpresa	Alegria	Esperança

Questão 18: O colega de trabalho de Sheila intencionalmente não dá a Sheila algumas informações importantes sobre como pedir um aumento. É mais provável que Sheila (se) sinta...?

© Carolyn MacCann, 2008
& Victor Vasconcelos, 2023

Deprimida	Desprezo	Frustrada	Enraivecida	Angustiada

Questão 19: Marisa está procurando uma nova casa. Algo aconteceu e ela se arrependeu. O que é mais provável que tenha acontecido?

Ela não fez uma oferta pela casa que queria e agora está tentando descobrir se é tarde demais.	Ela encontrou uma casa de que gostou e que não achava que encontraria.	Ela não pôde fazer uma oferta por uma casa da qual gostou porque o banco não lhe deu o dinheiro a tempo.	Ela não fez uma oferta por uma casa que ela gostou e agora outra pessoa comprou.	Ela fez uma oferta de uma casa e está esperando para ver se ela é aceita.

Questão 20: Maria estava trabalhando em sua mesa. Algo aconteceu que a fez se sentir surpresa. O que é mais provável que tenha acontecido?

Seu colega de trabalho contou uma piada boba.	Ela estava trabalhando em uma nova tarefa com a qual não havia lidado antes.	Ela encontrou alguns resultados que eram diferentes do que ela pensava que seriam.	Ela percebeu que não seria capaz de completar seu trabalho.	Ela teve que fazer uma tarefa que normalmente não fazia no trabalho.

Questão 21: A pequena empresa de Gabriel está atraindo cada vez menos clientes e ele não sabe dizer por quê. Não parece haver nada que ele possa fazer para ajudar. É mais provável que Gabriel se sinta...?

Assustado	Irritado	Triste	Culpado	Angustiado

Questão 22: Leonardo pensa que outra pessoa intencionalmente fez com que algo bom acontecesse com ela. É mais provável que Leonardo sinta...?

Esperança	Orgulho	Gratidão	Surpresa	Alívio

Questão 23: Kevin trabalha em seu emprego atual há alguns anos. De repente, ele descobre que receberá uma promoção. É mais provável que Kevin sinta...?

Orgulho	Alívio	Alegria	Esperança	Culpa

Questão 24: Por suas próprias ações, uma pessoa atinge um objetivo que queria alcançar. É mais provável que a pessoa sinta...?

Alegria	Esperança	Alívio	Orgulho	Surpresa

Questão 25: Uma situação indesejada para uma pessoa torna-se menos provável ou acaba completamente. É mais provável que essa pessoa sinta...?

Arrependimento	Esperança	Alegria	Tristeza	Alívio

Questão 26: Hélder tenta usar seu novo celular. Ele sempre foi capaz de descobrir como usar aparelhos diferentes, mas não consegue fazer o telefone funcionar. É mais provável que Hélder se sinta...?

Angustiado	Confuso	Surpreso	Aliviado	Frustrado

© Carolyn MacCann, 2008
& Victor Vasconcelos, 2023

Questão 27: O amigo de Douglas está doente e tosse em cima de Douglas sem se preocupar em virar as costas ou cobrir a boca. É mais provável que Douglas sinta...?

Ansiedade	Aversão	Surpresa	Ciúmes	Raiva

Questão 28: Embora tenha tido o cuidado de evitar todos os fatores de risco, Tina contraiu câncer. Há apenas uma pequena chance de que o câncer seja benigno e nada que Tina faça agora pode fazer a diferença. É mais provável que Tina se sinta...?

Assustada	Angustiada	Irritada	Triste	Esperançosa

Questão 29: Ruan e sua esposa estão conversando sobre o que aconteceu com eles naquele dia. Algo aconteceu que fez Ruan se sentir surpreso. O que é mais provável de ter acontecido?

Sua esposa estava falando muito, o que não costuma acontecer.	Sua esposa falou sobre coisas que eram diferentes do que eles costumavam discutir.	Sua esposa disse que poderia ter más notícias.	Sua esposa contou a Ruan algumas notícias que não eram o que ele pensava que seria.	Sua esposa contou uma história engraçada.

Questão 30: Um evento futuro pode ter consequências ruins. Nada pode ser feito para alterar isso. É mais provável que a pessoa envolvida se sinta...?

Triste	Irritada	Angustiada	Assustada	Esperançosa

Questão 31: É certo que Alexandre vai conseguir o que quer. É mais provável que Alexandre sinta...?

Orgulho	Alívio	Alegria	Esperança	Culpa

Questão 32: Por acaso, surge uma situação em que existe a possibilidade de Ronaldo conseguir o que deseja. É mais provável que Ronaldo sinta...?

Angústia	Esperança	Surpresa	Alegria	Medo

Questão 33: Um supervisor que é desagradável deixa de trabalhar na empresa de Alfonso. É mais provável que Alfonso sinta...?

Alegria	Esperança	Arrependimento	Alívio	Tristeza

Questão 34: A natureza do trabalho de Sara muda devido a fatores imprevisíveis e ela não consegue mais fazer as partes de seu trabalho que ela mais gostava. É mais provável que Sara sinta...?

Vergonha	Tristeza	Raiva	Angústia	Frustração

Questão 35: Leila não tem conseguido dormir bem ultimamente e não há mudanças em sua vida que possam indicar o motivo. É mais provável que Leila se sinta...?

Irritada	Assustada	Triste	Angustiada	Culpada

© Carolyn MacCann, 2008
& Victor Vasconcelos, 2023

Questão 36: Uma pessoa sente que tem controle sobre uma situação. A situação acaba mal por nenhuma razão particular. É mais provável que a pessoa sinta?

Confusão	Alívio	Surpresa	Frustração	Angústia

Questão 37: Evandro acredita que outra pessoa intencionalmente fez com que algo bom parasse de acontecer com ele. No entanto, ele sente que pode fazer algo sobre isso. É mais provável que Evandro se sinta...?

Irritado	Desprezado	Angustiado	Deprimido	Frustrado

Questão 38: O novo gerente do trabalho da Eunice muda o horário de todos os funcionários para um padrão menos flexível, sem deixar espaço para discussão. É mais provável que Eunice sinta...?

Apatia	Raiva	Ciúmes	Surpresa	Ansiedade

Questão 39: Jonas acredita que outra pessoa lhe causou danos, devido ao mau caráter dessa pessoa. Ele pensa que provavelmente pode lidar com a situação. É mais provável que Jonas sinta...?

Desprezo	Raiva	Terror	Entusiasmo	Vergonha

Questão 40: Pedro chegou em casa tarde e perdeu seu programa de TV favorito acabou. O parceiro de Pedro gravou o show para ele. É mais provável que Pedro sinta...?

Surpresa	Esperança	Orgulho	Alívio	Gratidão

Questão 41: Mateus está no emprego atual há seis meses. Algo aconteceu que o fez sentir arrependimento. O que é aconteceu, mais provavelmente?

Ele não se candidatou ao cargo que queria e descobriu que outra pessoa menos qualificada conseguiu o emprego.	Ele não se candidatou a um cargo que desejava e começou a procurar um cargo semelhante.	Ele descobriu que as oportunidades de promoção se esgotaram.	Ele descobriu que não conseguiu uma posição que pensava que conseguiria.	Ele não ouviu falar de uma posição que poderia ter se candidatado e agora é tarde demais.

Questão 42: O time de vôlei de Pâmela treinou muito e ganhou o campeonato. É mais provável que Pâmela sinta...?

Esperança	Orgulho	Alívio	Alegria	Surpresa

© Carolyn MacCann, 2008
& Victor Vasconcelos, 2023

Supplementary Material

TESTE SITUACIONAL DE GERENCIAMENTO EMOCIONAL

SITUATIONAL TEST OF EMOTIONAL MANAGEMENT

Carolyn MacCann
Victor Vasconcelos
Cristiane Faiad

Questão 1: O colega de trabalho de Leandro atrasou a entrega de uma informação importante, fazendo com que Lee também se atrase. Qual ação seria a mais eficaz para Leandro?

Trabalhar mais para compensar.	Ficar com raiva do colega de trabalho.	Explicar a urgência da situação ao colega de trabalho.	Nunca mais depender desse colega de trabalho.

Questão 2: Renata deixou seu emprego para ser mãe em tempo integral, o que ela ama, mas ela sente falta da companhia e do companheirismo de seus colegas de trabalho. Qual ação seria a mais eficaz para Renata?

Gostar de ser mãe em tempo integral, algo que ela ama fazer.	Tentar ver seus antigos colegas de trabalho socialmente, convidando-os para sair.	Juntar-se a um grupo social de novas mães.	Ver se consegue encontrar trabalho de meio-período.

Questão 3: Pedro tem habilidades específicas que seus colegas de trabalho não têm e sente que sua carga de trabalho é maior por causa disso. Qual seria a ação mais eficaz para Pedro?

Falar com seu chefe sobre isso.	Começar a procurar um novo emprego.	Ter muito orgulho de suas habilidades únicas.	Falar com seus colegas de trabalho sobre isso.

Questão 4: Mário está mostrando a Minerva, uma nova funcionária, como o sistema funciona. O chefe deles passa e chama atenção ao fato de que Mário está errado sobre várias coisas, pois foram feitas mudanças. Mário tem uma boa relação com o chefe, apesar de não terem muito em comum. Qual ação seria a mais eficaz para Mário?

Fazer uma piada com Minerva, explicando que ele não sabia das mudanças.	Não se preocupar com isso, apenas ignorar a interrupção.	Procurar aprender as novas mudanças.	Dizer ao chefe que a crítica foi inadequada.

Questão 5: Wanda e Carla são amigas e dividem um escritório há anos, mas Wanda consegue um novo emprego e Carla perde contato com ela. Que ação seria a mais eficaz para Connie?

Só aceitar que ela se foi e que a amizade acabou.	Ligar para Wanda e convidá-la para almoçar ou tomar um café para conversar sobre as novidades.	Entrar em contato com Wanda e combinar de conversar, mas também fazer amizade com sua substituta.	Passar algum tempo conhecendo as outras pessoas no escritório e fazer novas amizades.

Questão 6: Martina é selecionada para um trabalho muito procurado, mas tem que viajar de avião para o local. Martina tem fobia de voar. Qual seria a ação mais eficaz para Martina?

Consultar um médico sobre isso.	Não ir ao local.	Encarar o medo de uma vez.	Encontrar formas alternativas de transporte.

Questão 7: Manuel está a poucos anos de se aposentar quando descobre que seu cargo não existirá mais—apesar de manter o emprego, ele terá um cargo de menor prestígio. Qual ação seria a mais eficaz para o Manuel?

Considerar cuidadosamente suas opções e discuti-las com sua família.	Conversar com seu chefe ou com a gestão responsável sobre isso.	Aceitar a situação, apesar de ainda sentir amargura sobre ela.	Sair de vez desse trabalho.

Questão 8: Alan ajuda Taís, uma colega com quem trabalha ocasionalmente, em uma tarefa difícil. Taís reclama que o trabalho de Alan não é muito bom, e Alan responde que Taís deveria estar grata por ele estar fazendo um favor a ela. Eles discutem. Qual ação seria a mais eficaz para Alan?

© Carolyn MacCann, 2008
& Victor Vasconcelos, 2023

Parar imediatamente de ajudar Taís e não a ajudar novamente.	Se esforçar mais para pensar em formas de ajudá-la da forma que ela precise.	Pedir desculpas a Taís por ter trabalhado de forma errada.	Acabar com a discussão pedindo que Taís o informe como ajudá-la.

Questão 9: Solange começa um novo emprego onde não conhece ninguém e descobre que ninguém é muito amigável. Qual ação seria a mais eficaz para Solange?

Se divertir com seus amigos fora do horário de trabalho.	Se concentrar em fazer bem o seu trabalho no novo emprego.	Se esforçar para falar com as pessoas e ser amigável.	Deixar o emprego e encontrar um com um ambiente melhor.

Questão 10: Daniela está nervosa em apresentar seu trabalho para um grupo de idosos que podem não entender, pois não conhecem muito sobre sua área. Qual seria a ação mais eficaz para Daniela?

Ser positiva e confiante, sabendo que tudo correrá bem.	Só dar a apresentação.	Trabalhar em sua apresentação, simplificando as explicações.	Praticar a apresentação com leigos, como amigos ou familiares.

Questão 11: André se muda da cidade onde seus amigos e familiares estão. Ele descobre que seus amigos fazem menos esforço para manter contato do que ele pensava que fariam. Qual seria a ação mais eficaz para André?

Tentar ajustar-se à vida na nova cidade juntando-se a clubes e atividades de lá.	Ele deve se esforçar para contatá-los, mas também tentar conhecer pessoas em sua nova cidade.	Deixar de lado seus velhos amigos, que se mostraram pouco confiáveis.	Dizer a seus amigos que está decepcionado por eles não terem entrado em contato.

Questão 12: A equipe de Helena vem apresentando um desempenho muito bom. Eles acabaram de receber trabalho de baixa qualidade de outra equipe que devem incorporar em seu próprio projeto. Qual seria a ação mais eficaz para Helena?

Não se preocupar com isso.	Dizer à outra equipe que eles devem refazer seu trabalho.	Informar o gerente do projeto sobre a situação.	Corrigir o trabalho da outra equipe.

Questão 13: Clayton está no exterior há muito tempo e volta para visitar sua família. Tanta coisa mudou que Clayton se sente deixado de lado.

Nada – isto se resolverá em breve.	Dizer à família que ele se sente deixado de lado.	Passar algum tempo ouvindo as novidades e se envolvendo novamente.	Refletir que os relacionamentos podem mudar com o tempo.

Questão 14: Catarina demora muito para ajustar o timer do DVD. Com a família assistindo, sua irmã diz “Sua idiota, você está fazendo tudo errado, você não consegue fazer o vídeo funcionar?” Catarina é bastante próxima de sua irmã e família. Qual seria a ação mais eficaz para Catarina?

Ignorar sua irmã e continuar na tarefa.	Fazer com que a irmã dela ajude ou ajuste o vídeo sozinha.	Dizer a sua irmã que ela está sendo chata.	Não mexer mais com eletrodomésticos na frente de sua irmã ou família novamente.

Questão 15: Os pais de Bruno estão com quase 80 anos e moram sozinhos em uma casa que fica após a divisa do estado. Ele está preocupado que eles precisem de ajuda, mas eles negam, com raiva, sempre que ele toca no assunto. Qual seria a ação mais eficaz para Bruno?

Visitar com frequência e pedir a outras pessoas que vejam como seus pais estão.	Acreditar na fala de seus pais de que estão bem.	Continuar contando a seus pais suas preocupações, enfatizando sua importância.	Forçar seus pais a se mudarem para um asilo.

Questão 16: Marcos se orgulha de seu trabalho ser da mais alta qualidade. Em um projeto conjunto, outras pessoas fazem um péssimo trabalho, supondo que Max corrigirá seus erros. Qual ação seria a mais eficaz para Marcos?

© Carolyn MacCann, 2008
& Victor Vasconcelos, 2023

Esquecer isso.	Confrontar os outros e dizer-lhes que devem corrigir seus erros.	Informar o gerente do projeto sobre a situação.	Corrigir os erros.

Questão 17: Daniel foi aceito para um cargo de prestígio em um país diferente de sua família, de quem é próximo. Ele e sua esposa decidem que vale a pena se mudar. Qual ação seria a mais eficaz para Daniel?

Perceber que ele não deveria ter se candidatado ao emprego se não quisesse sair.	Montar um esquema para manter contato, como telefonemas semanais ou videochamadas.	Pensar nas grandes oportunidades que essa mudança oferece.	Não aceitar a posição.

Questão 18: Um funcionário júnior fazendo ajustes de rotina em alguns equipamentos de Téo o acusa de causar o mau funcionamento do equipamento. Qual ação seria a mais eficaz para Téo?

Repreender o funcionário por fazer tais acusações.	Ignorar a acusação, ela não é importante.	Explicar que o mau funcionamento não foi culpa dele.	Aprender mais sobre como usar o equipamento para que ele não quebre.

Questão 19: Miranda atende o telefone e ouve que parentes próximos estão no hospital gravemente doentes. Qual ação seria a mais eficaz para Miranda?

Se permitir chorar e expressar emoção pelo tempo que ela quiser.	Falar com outros parentes para se acalmar e descobrir o que está acontecendo, e então ir ao hospital.	Não há nada que ela possa fazer.	Visitar o hospital e perguntar aos funcionários sobre a situação deles.

Questão 20: A mulher que substitui Célia no final de seu turno chegou vinte minutos atrasada, sem ter uma justificativa, nem pedir desculpas. Qual seria a ação mais eficaz para Célia?

Esquecer isso, a menos que aconteça novamente.	Contar ao chefe sobre isso.	Pedir uma explicação sobre o atraso dela.	Dizer a ela que isso é inaceitável.

Questão 21: Ao ingressar no estudo em tempo integral, Vicente não pode arcar com o tempo ou dinheiro que costumava gastar no treinamento de natação, no qual era muito bom. Apesar de gostar de estudar em tempo integral, ele sente falta dos treinos. Qual seria a ação mais eficaz para Vicente?

Concentrar-se em estudar muito, para ser aprovado no curso.	Ver se existe uma liga local ou um esporte mais barato e que exija menos tempo.	Considerar profundamente se o esporte ou o estudo são mais importantes para ele.	Informar-se sobre bolsas de estudo para atletas.

Questão 22: O colega de casa de Evandro cozinhou comida tarde da noite e deixou uma bagunça enorme na cozinha, que Evandro descobriu no café da manhã. Qual ação seria mais eficaz para Evan?

Dizer ao seu colega de casa para limpar a bagunça.	Pedir ao seu companheiro de casa que isso não aconteça novamente.	Limpar a bagunça.	Supor que o companheiro de casa irá limpá-la mais tarde.

Questão 23: Gregório acaba de voltar para a universidade após uma pausa de vários anos. Ele está cercado por alunos mais jovens que parecem muito confiantes em suas habilidades e não tem certeza se está no nível deles. Qual ação seria mais eficaz para Gregório?

Concentrar-se em sua vida fora da universidade.	Estudar bastante e assistir a todas as aulas.	Conversar com outras pessoas na situação dele.	Perceber que ele é melhor do que os alunos mais jovens, pois tem mais experiência de vida.

© Carolyn MacCann, 2008
& Victor Vasconcelos, 2023

Questão 24: Os colegas de casa de Glória nunca compram produtos de higiene ou de limpeza quando eles estão acabando, contando que Glória os comprarão, o que ela não gosta. Eles se conhecem razoavelmente bem, mas ainda não discutiram questões financeiras. Qual seria a ação mais eficaz para Glória?

Não comprar os itens.	Introduzir um novo sistema para compras de supermercado e compartilhamento de custos.	Falar com seus colegas de casa que ela tem um problema com isso.	Esconder os itens que ela compra dos outros.

Questão 25: Sheila não fala com seu sobrinho há meses, mas quando ele era mais jovem eles eram muito próximos. Quando ela liga para ele, ele só pode falar por cinco minutos. Qual ação seria a mais eficaz para Sheila?

Perceber que ele está crescendo e talvez não queira mais passar tanto tempo com a família.	Fazer planos para passar na casa dele e ter uma conversa agradável.	Entender que os relacionamentos mudam, mas continuar ligando para ele de vez em quando.	Ficar chateada com isso, mas perceber que não há nada que ela possa fazer.

Questão 26: Márcio descobre que alguns membros de sua equipe de vôlei têm dito que ele não é um jogador muito bom. Que ação seria a mais eficaz para Márcio?

Embora ele possa ser ruim em esportes, lembrar que ele é bom em outras coisas.	Esquecer disso.	Fazer algum treinamento extra para tentar melhorar.	Sair do time.

Questão 27: Joel sempre lidou com um determinado cliente, mas, quando este cliente solicitou um trabalho muito complexo, seu chefe deu a tarefa para seu colega de trabalho, em vez dele. Joel se pergunta se seu chefe acha que ele não pode lidar com os trabalhos mais difíceis. Qual seria a ação mais eficaz para Joel?

Acreditar que ele está tendo um bom desempenho e que receberá o próximo trabalho complexo.	Trabalhar bem para que ele receba tarefas complexas no futuro.	Perguntar ao seu chefe por que o colega de trabalho que recebeu o trabalho.	Não se preocupar com isso, a menos que aconteça novamente.

Questão 28: Heloísa está no exterior quando descobre que seu pai faleceu de uma doença que ele tinha há anos. Qual seria a ação mais eficaz para Heloísa?

Entrar em contato com seus parentes próximos para obter informações e apoio.	Tentar não pensar nisso, e continuar com sua vida diária da melhor maneira possível.	Se sentir terrível por ela ter deixado o país em um momento tão importante.	Refletir sobre o significado mais profundo dessa perda.

Questão 29: Marcela e sua cunhada normalmente se dão muito bem, e a cunhada regularmente cuida do filho de Marcela por uma pequena quantia. Ultimamente ela também tem limpado teias de aranha na casa e comentando sobre a bagunça, o que Marcela acha um insulto. Qual ação seria a mais eficaz para Marcela?

Dizer à cunhada que esses comentários a incomodam.	Conseguir uma nova babá.	Ficar grata que sua casa está sendo limpa de graça.	Dizer para a cunhada não limpar a casa, apenas tomar conta do filho.

Questão 30: Guilherme está nervoso sobre atuar em uma cena quando há muitos atores muito experientes na plateia. Qual ação seria a mais eficaz para Guilherme?

Colocar as coisas em perspectiva – não é o fim do mundo.	Usar técnicas de atuação para se acalmar.	Acreditar em si mesmo e entender que tudo ficará bem.	Praticar mais suas cenas para que ele atue bem.

Questão 31: Jéssica tem quase certeza de que a empresa em que trabalha está falindo e seu emprego está ameaçado. É uma grande empresa e nada oficial foi dito. Qual ação seria a mais eficaz para Jéssica?

© Carolyn MacCann, 2008
& Victor Vasconcelos, 2023

Descobrir o que está acontecendo e discutir suas preocupações com sua família.	Tentar manter a empresa a salvo trabalhando mais.	Começar a se candidatar a outros empregos.	Pensar nesses eventos como uma oportunidade para um novo começo.

Questão 32: Maria Luiza muda de uma empresa pequena para uma muito grande, onde há pouco contato pessoal, o que ela sente falta. Qual ação seria a mais eficaz para Maria Luiza?

Conversar com seus colegas de trabalho, tentar criar contatos e fazer amigos.	Começar a procurar um novo emprego para que ela possa sair daquele ambiente.	Apenas dar um tempo, e as coisas ficarão bem.	Focar em seus amigos de trabalho e colegas de empregos anteriores.

Questão 33: Um cliente exigente toma muito tempo de Geni e depois pede para falar com o chefe de Geni sobre seu desempenho. Embora o chefe de Geni assegure a ela que seu desempenho está bom, Geni se sente chateada. Que ação seria a mais eficaz para Geni?

Conversar com seus amigos ou colegas de trabalho sobre isso.	Ignorar o incidente e passar para sua próxima tarefa.	Acalmar-se respirando fundo ou fazendo uma caminhada curta.	Pensar que ela foi bem-sucedida no passado e que não é culpa dela este cliente seja difícil.

Questão 34: Bruno e Flávio costumam ir a um café depois de uma semana de trabalho e conversar sobre o que está acontecendo na empresa. Depois que Bianco é transferido para uma seção diferente da empresa, ele para de ir ao café. Flávio sente falta dessas conversas de sexta-feira. Qual ação seria mais eficaz para Flávio?

Ir ao café ou socializar com outros trabalhadores.	Não se preocupar com isso, ignorar as mudanças e deixar Bruno em paz.	Não falar com Bruno novamente.	Convidar Bruno novamente, talvez remarcando para outra ocasião.

Questão 35: Jefferson teve vários empregos temporários em um mesmo setor, mas está animado para começar um emprego em um setor diferente. Seu pai, casualmente, faz a observação de que ele provavelmente durará seis meses. Qual ação seria mais eficaz para Jefferson?

Dizer ao pai que ele está completamente errado.	Provar que ele está errado trabalhando duro para ter sucesso no novo emprego.	Pensar nos aspectos positivos do novo emprego.	Ignorar os comentários de seu pai.

Questão 36: Daniela, amiga de Michele, está se mudando para o exterior para morar com seu parceiro. Elas são boas amigas há muitos anos e é improvável que Daniela volte. Qual seria a ação mais eficaz para Michele?

Esquecer Daniela.	Passar tempo com outros amigos, mantendo-se ocupada.	Pensar que Dara e seu parceiro retornarão em breve.	Certificar-se de que elas mantenham contato por e-mail, telefone ou carta.

Questão 37: Douglas precisa fazer uma cirurgia de próstata e está bastante assustado com o processo. Ele ouviu que é bastante doloroso. Qual ação seria a mais eficaz para Douglas?

Descobrir o máximo que puder sobre o procedimento e concentrar-se em se acalmar.	Manter-se ocupado para que ele não pense na cirurgia iminente.	Conversar com sua família sobre suas preocupações.	Conversar com seu médico sobre o que vai acontecer.

Questão 38: O acesso de Hannah a recursos essenciais para seu trabalho foi postergado e seu trabalho está bem atrasado. O relatório de progresso dela não menciona a falta de recursos. Que ação seria mais eficaz para Hannah?

Explicar a falta de recursos para seu chefe ou para a gerência.	Aprender que ela deve planejar com antecedência para a próxima vez.	Documentar a falta de recursos em seu relatório de progresso.	Não se preocupar com isso.

© Carolyn MacCann, 2008
& Victor Vasconcelos, 2023

Questão 39: Jane recebe uma advertência oficial por entrar em uma área restrita. Ela nunca foi informada de que a área era restrita e perderá o emprego se receber mais duas advertências, o que ela considera injusto. Que ação seria a mais eficaz para Jane?

Pensar na injustiça da situação.	Aceitar o aviso e tomar cuidado para não entrar em áreas restritas de agora em diante.	Explicar que ela não sabia que era restrito.	Respirar fundo algumas vezes e acalmar-se.

Questão 40: Alana atua há vários meses em um cargo de alto escalão. Uma decisão foi feita que apenas funcionários de longa data podem agora atuar nessas funções, e Alana não está na empresa há tempo suficiente para fazê-lo. Qual seria a ação mais eficaz para o Alana?

Se demitir.	Usar essa experiência para ser promovida quando ela for funcionária de longa data.	Aceitar esta nova regra, apesar de sentir-se injustiçada.	Perguntar à administração se uma exceção pode ser feita.

Questão 41: A amiga de Rebeca comenta que seus filhos pequenos parecem estar se desenvolvendo mais rapidamente do que os de Rebeca. Rebeca vê que isso é verdade. Qual seria a ação mais eficaz para Rebeca?

Discutir a questão com outra amiga.	Brigar com a amiga por ter falado isso.	Perceber que as crianças se desenvolvem em ritmos diferentes.	Conversar com um médico sobre as taxas normais de desenvolvimento.

Questão 42: Júlio está trabalhando meio período em um novo emprego enquanto faz curso superior. Seus horários de plantão da semana foram alterados no último minuto, sem consultá-lo. Qual ação seria a mais eficaz para Júlio?

Recusar-se a trabalhar nos novos plantões.	Descobrir se há alguma explicação razoável para as mudanças de plantão.	Dizer ao gerente responsável pelos plantões que ele não está feliz com isso.	Aceitar as mudanças de mal gosto e fazer os plantões.

Questão 43: Jacó está tendo uma grande festa de família para comemorar sua mudança para sua nova casa. Ele quer que o dia corra bem e está um pouco nervoso com isso. Que ação seria a mais eficaz para Jacó?

Conversar com amigos ou parentes para aliviar suas preocupações.	Tentar se acalmar, talvez fazer uma caminhada curta ou meditar.	Preparar-se com antecedência para que ele tenha disponível tudo o que precisa.	Aceitar que as coisas não vão ser perfeitas, mas que sua família vai entender.

Questão 44: Júlia não via Kátia há séculos e esperava ansiosamente pela viagem de fim de semana. No entanto, Ka mudou muito e Julie descobre que ela não é mais uma companhia interessante. Qual ação seria a mais eficaz para Julie?

Cancelar a viagem e ir para casa.	Se dar conta de que é hora de desistir da amizade e seguir em frente.	Entender que as pessoas mudam, então seguir em frente, mas lembrar-se dos bons tempos.	Focar em suas outras amizades, mais gratificantes.

© Carolyn MacCann, 2008
& Victor Vasconcelos, 2023

MANUSCRIPT 3

Advantages and Challenges of Computer-Adaptive Testing

Vantagens e Desafios da Testagem Adaptativa Computadorizada

Abstract

Computer-adaptive testing (CAT) is a powerful tool that leverages computational power to enhance measurement efficiency, validity, and reliability. An algorithm selects items based on the test taker's ability, using criteria that maximize the information gained at their estimated ability level. The test begins with a medium-difficulty item and adjusts the difficulty based on the test taker's responses. This study examines the advantages and disadvantages of CAT and reviews essential aspects for prospective test developers. CAT offers several benefits that make it a superior choice over traditional testing methods. These include reduced measurement error at all ability levels, shorter test times, increased security against fraud, detection of low-quality responses, and improved validity. These advantages translate into a more accurate and efficient assessment process, saving time and resources for both test takers and administrators. However, there are also potential drawbacks to consider, such as test-related construct-irrelevant variance, increased infrastructure costs, and the need for a large item pool. To construct an optimal CAT and mitigate its disadvantages, test developers must understand concepts such as pool utilization, item selection criteria, stopping rule, and test information function. This study provides an in-depth explanation of these concepts and demonstrates how careful planning and implementation can overcome potential challenges. In conclusion, CAT is a powerful tool that offers significant advantages over traditional testing methods. With careful planning and implementation, it can provide a more accurate and efficient assessment process that benefits both test takers and administrators.

Keywords: computer-adaptive testing, item selection criteria, test development.

Resumo

O teste adaptativo computadorizado (CAT) é uma ferramenta poderosa que aproveita o poder computacional para aumentar a eficiência, validade e confiabilidade da medição. Um algoritmo seleciona itens com base na habilidade do candidato, usando critérios que maximizam as informações obtidas em seu nível de habilidade estimado. O teste começa com um item de dificuldade média e ajusta a dificuldade com base nas respostas do candidato. Este estudo examina as vantagens e desvantagens do CAT e revisa aspectos essenciais para desenvolvedores de testes em potencial. O CAT oferece vários benefícios que o tornam uma escolha superior em relação aos métodos de teste tradicionais. Isso inclui erros de medição reduzidos em todos os níveis de habilidade, tempos de teste mais curtos, maior segurança contra fraudes, detecção de respostas de baixa qualidade e validade aprimorada. Essas vantagens se traduzem em um processo de avaliação mais preciso e eficiente, economizando tempo e recursos tanto para os candidatos quanto para os aplicadores. No entanto, também há possíveis desvantagens a serem consideradas, como variância irrelevante ao construto relacionada ao teste, aumento dos custos de infraestrutura e a necessidade de um grande banco de itens. Para construir um CAT ideal e mitigar suas desvantagens, os desenvolvedores de testes devem entender conceitos como utilização do banco de itens, critérios de seleção de itens, regra de parada e função de informação do teste. Este estudo fornece uma explicação detalhada desses conceitos e demonstra como o planejamento e a implementação cuidadosos podem superar possíveis desafios. Em conclusão, o CAT é uma ferramenta poderosa que oferece vantagens significativas sobre os métodos de testagem tradicionais. Com planejamento e implementação cuidadosos, ele pode fornecer um processo de avaliação mais preciso e eficiente que beneficia tanto os candidatos quanto os administradores.

Palavras-chave: teste adaptativo computadorizado, critérios de seleção de itens, desenvolvimento de testes.

Advantages and Challenges of Computer-Adaptive Testing

Computer-based testing is an umbrella term for a series of techniques associated with the use of computers to develop and administer psychological and educational tests. This method of testing has been studied since the 1970's (Lushene et al., 1974), but has received increased attention in the last decade (Ghosh & Lan, 2021). Typically, this term refers to two things. First, it refers to testing that is administered through a computer (computer-delivered testing [CDT]), including testing that is simply translated from a paper-and-pencil (P&P) form. Alternatively, it may refer to testing that exploits advanced computational techniques in user interaction (e.g., Heesacker et al., 2020), test scoring (e.g., Kurisu et al., 2022), or any other testing facet. One major technique that uses such advanced features is called computer-adaptive testing.

Computer-adaptive testing (CAT) is a form of CDT that uses an algorithm, typically based on item response theory (IRT) parameters, to administer a test in which the selection and presentation of items are tailored to the test taker's ability level. While CDT is contrasted with P&P, CAT is contrasted with linear fixed-form testing, a method in which test items are presented in a predetermined order, without any recourse for adaptability (Becker & Bergstrom, 2013; Luecht & Sireci, 2011).

The basic principle of CAT is straightforward: at the beginning of the test, an item selection algorithm (ISA) provides the examinee with an item of average difficulty. If the examinee answers that item correctly, the ISA selects a more challenging item for the next question. If the examinees answers incorrectly, the ISA selects an item with a lower difficulty parameter for the next question. This procedure is repeated with each subsequent item until a stopping rule is reached. As an examinee answers more items close to his true ability level, the measurement error decreases (Luecht & Sireci, 2011). Eventually, the difficulty of the items will converge with the examinee's ability level, resulting in a sharp decline in measurement

error. At this point, answering additional questions will not significantly improve the standard error of measurement. This stagnation in standard error change is frequently used as the stopping rule (Stafford et al., 2019).

The literature on CAT highlights numerous advantages, while also cautioning against significant challenges that may impede a researcher's ability to address their research questions. This work aims to review these advantages and challenges, as well as to provide guidance to researchers and test developers who wish to design a CAT test, so that they may meet these challenges.

Advantages

The advantages of CAT, when compared to fixed-form, P&P tests, include:

Low Estimation Error for Test Takers in All Aptitude Levels

In developing fixed format P&P tests, test developers aim to construct items that provide the highest amount of information for a given aptitude interval. Though this may vary according to a test's objectives, the information that a test provides is typically distributed as a Gaussian curve, centered on the arithmetic mean of the population ability (Baker & Kim, 2017). For test takers whose ability level is not in that interval, measurement error may be so high that test scores may contain more error than information. CAT offers a solution to this problem by allowing the choice of items administered to be based on the expected information that the examinee's answers provide. Given a sufficient item pool, this leads to reduced standard error levels across ability levels (Wainer et al., 2000; Wang et al., 2019).

Shorter Test Times

If item selection is made considering information about past score, and the examinee misses a group of questions of a certain difficulty level, it can logically be concluded that no additional information will be gained from answers to even harder questions (Luecht, 2016). For example, in an algebra test composed of items with all four basic algebraic operations, and

then combinations of these algebraic operations, if a test taker starts getting multiplication items wrong, it becomes increasingly unlikely that they will correctly answer items that involve multiplication and division, for example.

Therefore, administering items with increased complexity is unnecessary when the algorithm has already determined, within a pre-established error margin, that the test taker's ability level is lower than the difficulty of those items. Therefore, test takers will see fewer items that are outside their difficulty level, and the algorithm may stop the test when it is most efficient. Research in educational testing has shown that the implementation of CAT could reduce test administration times by up to 40% (Luecht & Sireci, 2011). The consequences of reduced test times will be discussed later in this paper.

Increased Security Against Fraud

Ensuring fairness in high-stakes testing is a primary concern of test developers and stakeholders (American Educational Research Association [AERA] et al., 2014). For this reason, fraudulent behavior used to obtain better scores has become a growing concern in recent years (van der Linden, 2009). One effective way to prevent fraud is by employing CAT with a large item pool. This combination allows examinees to receive completely different test forms, but with the same content requirements and difficulty level. If examinees have access to only one dynamically generated test forms out of a myriad of possible test forms using different item combinations, their ability to commit fraud is reduced sharply (Luecht & Sireci, 2011). CAT may also speed up the process of including novel items by optimizing the calibration of item parameters via incomplete block designs (Ariel et al., 2006; van der Linden et al., 2004).

Several techniques have been developed to identify fraudulent behavior in computer-delivered testing based on information that can only be practically collected in this format. For example, Choe et al. (2018) report methods that include monitoring responses in real-time for significant changes in response time patterns. Evidence of items leakage is produced if items are

answered correctly under a minimum time that would be required, for example, to make calculations necessary to answer the question posed. Furthermore, several authors have proposed test statistics for detecting response patterns that may be associated with individuals who have previous knowledge of items being answered, both based on response patterns and ability estimates alone (Sinharay, 2017), and based on response times as well (van der Linden, 2022).

Immediate and Precise Feedback

Almost all traditional item types can be scored automatically, and scores can be calculated using both classical test theory and item response theory, if there is sufficient computational power. This enables the immediate calculation and returning of examinee scores at the conclusion of testing sessions. More sophisticated algorithms may provide detailed feedback, including an analysis of the examinee's strengths and suggestions for areas of improvement (Kyllonen & Christal, 1991).

Although most CAT models are intended for one-time evaluation, Yang et al. (2022) explored the use of a CAT-based adaptive formative assessment system to support personalized learning in a university programming course. The results of their experiment demonstrated that, after a period of just 7 weeks, students who utilized the proposed assessment system outperformed their peers who used a traditional non-adaptive assessment system.

Detection of Low-Quality Responses

In low-stakes assessment contexts, where participants lack an incentive to provide accurate responses, distinguishing between high and low-quality responses presents a significant challenge. Such data can negatively impact analyses aimed at producing validity evidence for a test (Gören et al., 2022). However, in addition to helping detect fraudulent behavior, there is research that shows that response time data may provide a valuable tool for

detecting both fraudulent behavior and low-quality responses (Linden & Krimpten-Stoop, 2003).

Sinharay (2016) suggested a person-fit statistic capable of detecting abrupt changes in test performance during a CAT, indicating that a participant may be answering items randomly. Van der Linden (2022) further suggests that responses given in less time than it takes to read the item may be considered suspicious. Additional research has examined the continuity of disengaged responding behavior across sections in low-stakes assessments and the impact of filtering student data based on their response behaviors (Bulut, 2021). These findings highlight the potential utility of response time data and advanced statistical methods in improving the validity of low-stakes assessments.

Increased Ecological Validity Through the Use of Innovative Items

The use of digital tools enables the development of tasks that are more fun and, therefore, more engaging to test-takers (Sireci & Zenisky, 2011). Dynamic, coordinated use of a mouse, keyboard, or touch screen allows for the navigation of 3d environments in simulation-based assessments. The ability to interact with objects through actions such as dragging-and-dropping enables examinees to manipulate their location, shape, size, and the physical environment around them. This development allows for more accurate assessment of skills in real situations, providing a significant source of ecological validity evidence (van der Ham et al., 2015), and addressing ecological validity concerns by Bjorkman (1969), who suggested that typical testing conditions did not accurately simulate the conditions where the use of the tested skills would be required.

The potential impact of computer simulations on validity was demonstrated in a study by Sarmet et al. (2013), who developed a measure of prosocial behavior utilizing a virtual game running on the Unity engine. While the measurement itself was conducted using a separate instrument, the game served as the stimuli for which players recorded their answers. Other

research has evaluated the direct use of measures collected during gaming, such as intelligence scores (Queiroga et al., 2016), or measures employed in the study of interpersonal conflict (Schlenker & Bonoma, 1978). Gundry (2022) describes a general model for building games to collect data from any psychological field. Gundry and Deterding (2019) also provide guidance on what should be avoided in game-based testing.

While not all research questions necessitate the use of a simulation, innovative techniques can be employed to test skills in various fields of study without the need for simulations. Drasgow and Olson-Buchanan (1998) narrate experiences with using innovative techniques assess skills such as group problem-solving through participant interaction, dermatological diagnosis using visual stimuli, and music aptitude using dynamically activated auditory stimuli. The use of innovative items has also helped constructs such as working memory (Oberauer et al., 2005), processing speed (Hunt et al., 1988), or even preconceptions and propensity to lie (Rooth, 2010) be measured in group, enabling efficient use of resources that has contributed to the execution of research projects that wouldn't have been possible without them. Conventional items can also be used in innovative ways by combining diverse sources of information, such as response time (Wise & Kong, 2005) or the number of times the examinee changed his answer, to make inferences about response process (Araneda et al., 2022).

Although there have been attempts to implement some of the mechanisms behind these advantages in computer-delivered linear tests (Ebel, 1953), CAT has been shown to make that implementation easier, more practical, or more effective (Rudner, 1998). For example, while a linear test can reach estimation errors as low as those of CAT, it requires the application of a greater number of items (Rudner, 1998), which may be more costly or otherwise more demanding, might not even be practically possible, and could negatively impact examinee's performance and test engagement (Wise, 2018).

Increased Overall Validity and Reliability

Many of these advantages contribute in some way to increase the validity of testing in CAT (Huff & Sireci, 2001). For instance, there is evidence that participants' performance may be impaired from long test times (Ackerman & Kanfer, 2009), which may harm test validity. Even worse, performance decline is not consistent across males and females (Balart & Oosterveen, 2019). Considering the Standards' guidelines (AERA et al., 2014), that could mean that a particular group of scores may not be valid depending on the difference in sex makeup between the sample of interest and the sample used in the tests' validity study.

Another element that may impair people's ability to reach their optimal performance levels during tests is boredom (Diehl & Wyrick, 2015). The data suggest that tests that are more engaging increase motivation and, therefore, performance, in at least some types of tests. Indeed, boredom has been associated with worse academic outcomes both for teenagers and for adults (Tze et al., 2016). The researchers found that boredom may account for as much as 25% of the variance in student achievement.

Given these findings, it is suggested that CAT can decrease the impact of these outside factors in test takers' scores because they can motivate test takers in two ways: by keeping them engaged with items at the adequate level of difficulty, and by reducing test times (Wise, 2018). These factors are related to reducing the estimation error for test takers in more aptitude levels. One other consequence of reducing the estimation error is increased reliability (Martin & Lazendic, 2018). This has been empirically confirmed in several studies (Rice et al., 2022).

Finally, other techniques may also help reduce the impact of boredom, demotivation, and other sources of construct-irrelevant variance in test scores (Martin & Lazendic, 2018). For example, research has shown that immediate feedback helps keep students engaged (Ling et al., 2017), whose possibility is one of the typically cited benefits of CAT (Huff & Sireci, 2011).

Disadvantages

Construct-Irrelevant Variance

Several validity concerns have been associated with CDT (Huff & Sireci, 2011). Despite Kapoor and Welch (2011) not finding significant differences in the classification of students in the United States through CDT tests compared to P&P tests, one major concern remains the potential for construct-irrelevant variance related to the use of input devices such as keyboards, mice, and touch screens.

While some studies have sought to assess the impact of having computational technology mediate testing on older individuals (Zygouris & Tsolaki, 2014), researchers suggest that care should be taken to choose adequate tests and settings to counter inadvertent discrimination against disadvantaged groups with limited computer access (de Beer, 2013). Additionally, most studies that did show that no significant differences were found between the performance of test takers in CAT and traditional testing, such as Kapoor and Welch's (2011) study, did not include innovative-format items. If test developers include such items, or require the use of different input devices, they may disadvantage test takers that are not familiar with such devices.

For instance, in a review of studies that utilized informatized cognitive assessment batteries available on older individuals, Zygouris and Tsolaki (2014) found that, although these studies reported accurate measurement of many variables, many of them noted that caregivers or technicians provided help to the patients by controlling the test's input device. Each test battery required the use of one or more out of a variety of input devices, including proprietary hardware. The authors also reported that psychometric measures were inadequate for some measures but did not attempt to hypothesize whether familiarity with the input devices played any role in these results.

While there have been no studies that called attention to a specific group being disadvantaged due to the use of novel input devices, there is evidence that different input devices affect participant performance, at least for specific tasks (Moreno & Segall, 1997). In this study, 3036 participants were asked to answer two tests in a time-limited context, a numerical operations (NO) test where they had to perform arithmetic operations, and a coding speed (CS) test where they had to assign code numbers to words. Depending on whether participants had a full keyboard, or an altered keyboard with only the relevant keys—their scores were significantly different on the NO task, $F (2, 3035) = 7.71$, $p = .001$, but not on the CS task, $F (2, 3035) = 1.17$, $p = .173$. The computers' form factor—whether a notebook or a desktop—also both the NO score, $F (2, 3035), = 4.41$, $p = .001$, and the CS score, $F (2, 3035) = 4.4$, $p = .013$ (Moreno & Segall, 1997).

Infrastructure and Expense

Another important consideration is the requisite infrastructure. In high-stakes testing, like job selection or professional habilitation tests, it may not be adequate to download the entire test into the computer of each examinee (Luecht, 2016). Even if the test is encrypted, at least one person would need to have the decryption key in each testing location, which presents a clear source of concern. It may be necessary to develop a network with specialized software to distribute the test items according to demand. This network may be quite simple for local selection but may become complex for state and national tests.

Item selection and ability estimation algorithms in IRT are computationally demanding, and unless a capable service is properly set up, examinees may face long loading times (Hetter et al., 1997). Regardless of the size of the testing session, information security professionals and expensive equipment are required. Therefore, like any CDT method, CAT has a starting price that is more elevated compared to P&P testing.

Additionally, the continued development of an item bank is an expensive prospect. This is complicated by the fact that many of the advantages of CAT can only be exploited with an item pool of substantial size (Bjorner et al., 2009).

Item Pool Size

While there are a lot of advantages associated with CAT that are not predicated on having an extensive item pool, some of its main advantages are. For instance, to keep measurement error low across ability levels, it is necessary to have items that provide information throughout all ability levels. Even worse, to ensure that test takers are not presented with the same test forms, it is necessary to have multiple items that provide information at each ability level.

Reduced Test Taker Control

On the one hand, CATs allow test developers more control over the testing session. On the other hand, this comes at the cost of reducing the test takers' control (Wise, 2018). One of the biggest problems reported by test takers consequently is that, since the items that the test taker is presented with are associated with this pattern of correct and wrong responses, the algorithm does not allow changes to be made to a test taker's answer history (Vispoel, 1998). Test takers have expressed that this hinders their ability to reach their optimal performance, especially for examinees that are used to employing a non-linear approach to answering tests.

Different solutions have been proposed for this problem. The most effective solution involves utilizing testlets instead of items. Instead of serving one item at a time, the ISA serves a grouping of items each time, known as a testlet, which may or may not be prepared in advance. So long as test takers are answering items that are part of the same testlet, they may go back and correct their answer. After submitting their answers, a new testlet is administered based on their performance, where they may, again, navigate back and forth as long as they

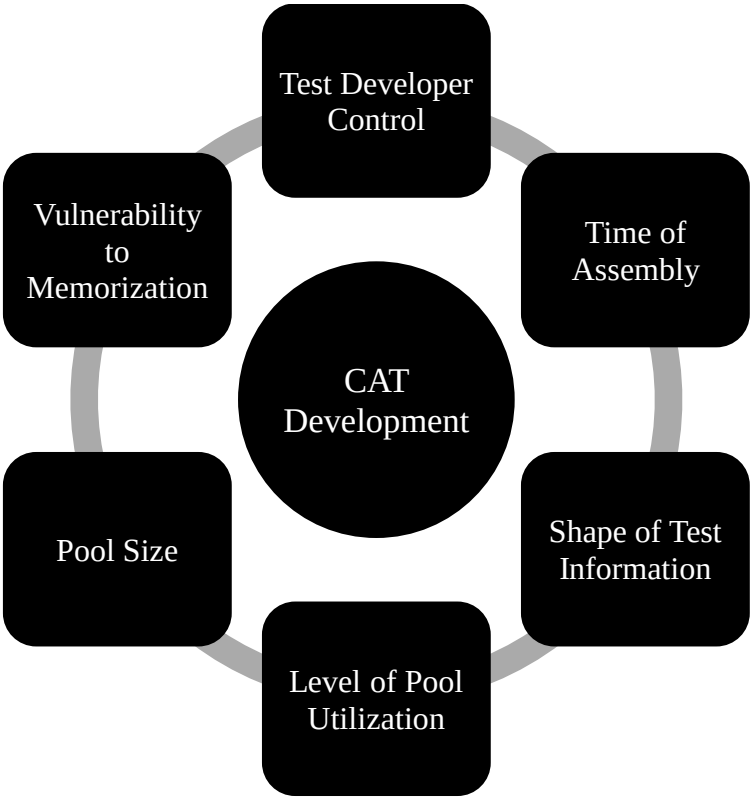
remain in the new testlet. One prominent testing system that uses this methodology is Multistage Testing (or MST; Zenisky et al., 2009).

CAT Properties

Becker and Bergstrom (2013) list several aspects that characterize a CAT model. These aspects are displayed in Figure 1.

Figure 1

Aspects of Computerized Adaptive Testing According to Becker and Bergstrom (2013).



The aspects described are: test developer control, meaning how much (or how little) control the test developers have over item selection; time of assembly, whether the test is assembled prior to, immediately before, or during administration; shape of test information, that is, what level of precision is desired for discriminating between ability levels (e.g., if an examinee fails a test, it might not matter if they failed by how many points; Becker & Bergstrom, 2013). Furthermore, the level of pool utilization and pool size required are two aspects associated with how many items, how precise the estimation, and how much of the

content domain the test covers. The last aspect the authors describe is vulnerability to memorization, a measure of how much the chosen model assists in the forensics necessary to catch cheating (Becker & Bergstrom, 2013).

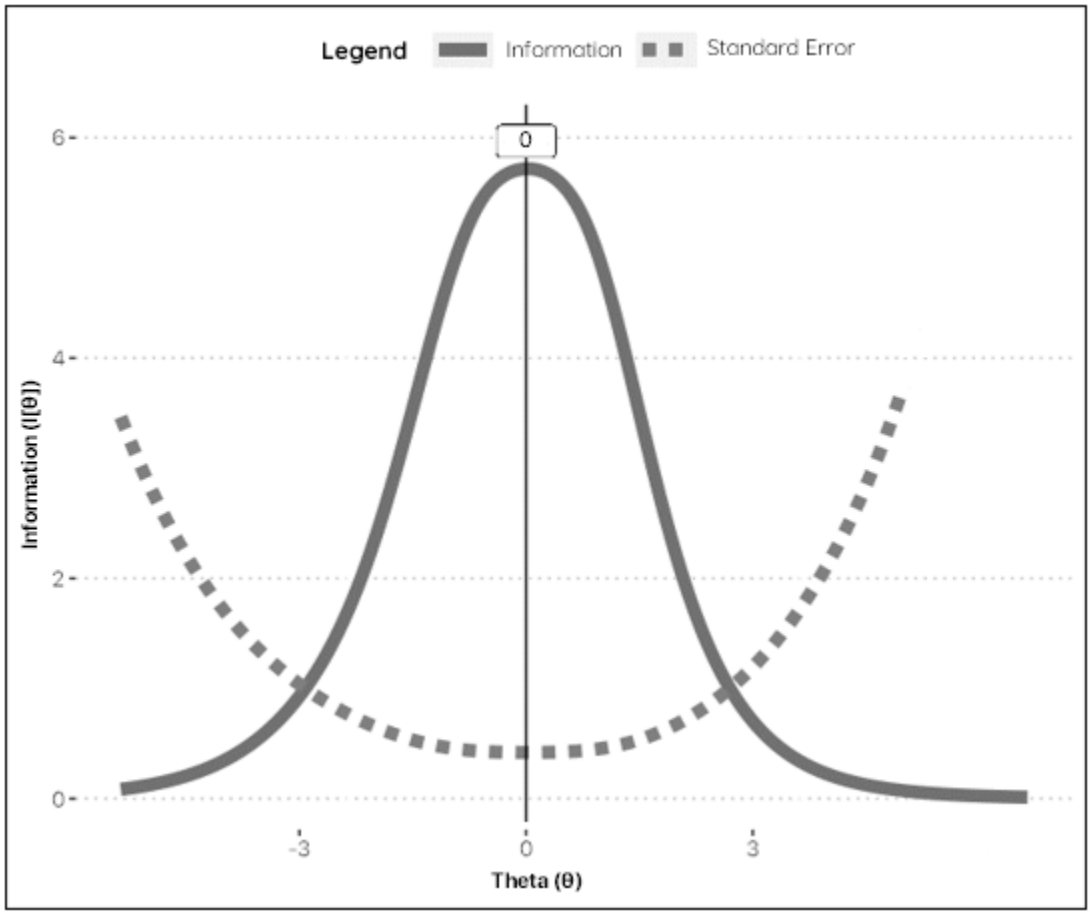
Many of these aspects seem to be interrelated. For instance, with regards to testing developer control, there are various parameters that may influence whether developers have a lot or no control over item selection (Stocking & Lewis, 2000). At one end of the spectrum, test developers may simply feed the item selection algorithm an item bank containing the items and their parameters, and the algorithm will choose the next item that contributes with the most information over the test taker's ability level, using the standard error as the sole stopping rule. At the other end, it is possible to choose how many total items from each specified content are shown, how many items overall are shown (Stocking & Lewis, 2000). This also seems to control item pool utilization, which refers to the percentage of items from an item bank a given model tends to use.

Meanwhile, the time of test assembly refers to one of three options (Becker & Bergstrom, 2013). First, the test may be composed of items that are selected immediately after the test taker submits an answer, with one new item being administered after each previous item is answered. Second, the ISA may select a group of items to form a testlet, of predetermined size, after the previous testlet has been answered completely. Finally, test developers may combine items into testlets before the test is administered and limit the ISA to select one of the predefined testlets (Zenisky et al., 2009). Again, the interrelation between these aspects is visible. These options can also be understood in terms of test developer control, wherein the third option, where developers choose exactly what items go with one another, have the most amount of control.

The shape of test information refers to the shape of the curve plotted by the test information function (TIF). An example of that shape is displayed in Figure 1.

Figure 2

Example Test Information Function Plotted with Information and Standard Error of Measurement Distributed Normally.



The information that a test provides through the TIF is given by the sum of the information that each item of the test provides, that is, each item information function (Lord, 1980). The TIF informs the relative amount of information a test provides for one ability level, versus for other ability levels. Its shape can be used to help guide the selection of items for tests used for a specific purpose. For example, for general-purpose tests, test developers typically want to have more information around the average ability level, so that the test can discriminate between the ability levels of most of the population (Boone & Staver, 2020). On the other hand, a test designed for a certification exam may call for a different shape. While the test may be intended to discriminate between various levels of certification, they may not be interested in

contrasting various levels of failing the test (Baker & Kim, 2017). At the other end of the spectrum, a clinical test designed to diagnose a given condition and discriminate between various stages of a disorder may not be interested in discerning between various levels of health.

Pool utilization is an equally consequential factor. A reduction of items answered in a test such as a math test, in which the skill being assessed is easily reproduced in multiple items, might impair reliability or not even that, but does not impair validity. The same cannot be said by tests associated with wide knowledge areas, such as history, or philosophy. For those tests, content validity and the appropriate representation of the content domain are of great concern, and using a stopping rule only predicated on the standard error may lead to a content underrepresentation (Luecht & Sireci, 2011; Sireci, 1998).

In addition, many of the benefits associated with CAT may be lost if ISAs consistently select the same items, which is guaranteed to happen if the algorithm is set to simply select the one item with the highest information content. In such cases, people who answer the same way will consistently see the same items. This has profound consequences to test security, for example. If people know they will see a particular set of items, they can memorize only those items to cheat the test; in these situations, CAT will have made cheating easier, not harder (Luecht, 2016). For that reason, ISAs that allow the test developer to set a level of randomization to item selection have been developed. This degree of randomization ensures that the probability of an algorithm selecting the item with the highest information content is distributed among the items with the second or third highest information content, or among any number of items specified by the test developer (Chalmers, 2012; Yoo & Chang, 2020).

The aspects enumerated by Becker and Bergstrom (2013) are associated with the item make-up of the test. Two additional aspects of statistical nature require consideration. These are the stopping rule and the ISA.

Stopping Rule

One of the most consequential characteristics of an adaptive test is the stopping rule, sometimes called termination criterion (Thompson, 2011). Typically, a testing session ends for one of three reasons. First, the test may end after the test taker answers a predetermined minimum number of items for each of the test's content domain. Second, the ISA may stop serving items after any further items are judged to offer too little information to the test taker's ability level (Luecht & Sireci, 2011).

Two prevalent methods for determining whether the administration of additional items would yield significant information involve calculating the variation in ability estimates and standard errors (SE), with the latter being more commonly employed (Wainer et al., 2000). Every time a test taker answers an item, a new estimate of their ability and its associated SE are calculated. For the variation in SE rule (ΔSE), the previous SE estimate is subtracted from the current SE estimate, and if the resulting value is lower than an arbitrary amount—usually 10^{-3}—, the testing sessions ends. For the variation in ability rule ($\Delta\theta$), the same calculation applies, but the ability estimate is used instead of the SE estimate.

These are the rules which are commonly selected as the stopping rule. They may be combined to form a stopping rule which requires, for example, a minimum number of items for each facet of the content domain, and then a ΔSE rule (van der Linden & Glas, 2010).

Item Selection Algorithm Criteria

Finally, CATs differ in how they select the items that are presented to the examinee. ISAs employ a criterion to informs this selection. The two sources of information used most often are the expected posterior variance (EPV; van der Linden, 1998), which informs the minimum EPV criterion (MEPV), or the Fisher information (FI), which informs the maximum Fisher information criterion (often called maximum Fisher information [MFI]) and other criteria that employ the Fisher information along with other metrics (Lord, 1980; Owen, 1975).

The EPV is the variance of the posterior distribution of the estimate of ability that will result from answering a given item (van der Linden, 1998). It is the error of estimation—hence, the MEPV criterion seeks to minimize it. The FI is the amount of information that a given item extracts for a given theta interval (Lord, 1980). Since the information minimizes the error, the criterion seeks to maximize it. These criteria are mathematically related, as the EPV can be calculated as the inverse of the MFI.

Several other ISAs make use of the FI. One of them, the maximum likelihood-weighted information (MLWI), consists of the FI weighted by the likelihood of the test taker being in a certain ability level given their previous answers (Veerkamp & Berger, 1997). A Bayesian approach suggests the use of the posterior distribution instead of the likelihood, which gives rise to the maximum posterior-weighted information (MPWI). The authors who proposed both criteria found either of them to be better than the MFI alone at reducing the theta estimate's standard error (Veerkamp & Berger, 1997).

One other prominent ISA that utilizes the FI is the Maximum Expected Information. The EI measure is typically calculated by multiplying the FI by the probability of an examinee that has a given ability given a prior distribution (Han, 2018). Van der Linden and Pashley (2000) compared these criteria and found that the MFI criterion had the worst, and the MEI and MEPV had the best performance. However, further studies found these criteria to be roughly equal in performance (Penfield, 2006; Reeve, 2006).

To clear the confusion, Choi and Swartz (2009) ran a study that compared the effectiveness of MFI, MLWI, MPWI, MEPV, MEI using Fischer's information and MEI using observed information, and random selection methods. The authors used real data and item bank from the Health–Related Quality of Life Scale (HRQOL), simulated data with the real item bank, as well as simulated data and item bank, to simulate participants answering an adaptive test with 5, 10, and 20 items. They calculated the squared standard error of measurement, root

mean squared difference, and correlation between the theta estimates and the previously calculated ability levels, for each item selection method, in each data set. Remarkably, in all three situations, except for the random selection methods, all methods were equally effective—though the small, not significant difference got smaller the more items that were used. Additionally, the MEPWI method, which had been previously found to be superior to the others (van der Linden, 1997; van der Linden & Pashley, 2000), was found to be mathematically identical to the MPWI (Choi & Swartz, 2009).

Newer criteria have been developed for ISAs using different sources of information, such as the Kullback-Leibler information (KL; Cover & Thomas, 1991), and Shannon entropy (SHE; Tatsuoka, 2002; Xu et al., 2003). Variations of these criteria that mirror early development with the Fisher information are also being created, such as the Posterior-Weighted KL (Cheng, 2009). Though they have been shown to be better than random selection (Luzardo, 2019), no comparisons have been made with older criteria.

Importantly, one problem has arisen from the use of these criteria, old and new. Both are used to find the item that provides the largest amount of information for a given theta interval. Since only one item can realistically provide the most information, and most items have, at most, a handful of outcomes with regards to estimating the ability level, the ISA would be incentivized to consistently serve the same items. This presents two challenges.

The first challenge is exposure control: the ISA should select the largest proportion of items it has in the item bank while maximizing efficiency in estimating ability levels. The second challenge is content balancing: the ISA should select items that, in conjunction, represent all aspects of the content of the test. Finally, the ISA is the most computationally intensive operation of CAT (Magis & Barrada, 2017); for the CAT delivery platform to serve many users at once, ISAs should work while minimizing use of computer resources. These challenges can be overcome by using smarter selection criteria.

To control the exposure of items, there have been developed criteria such as randomesque (Kingsbury & Zara, 1989), and the Sympson-Hetter (Hetter & Sympson, 1997), the fade-away (Han, 2012) methods. These methods work not in place of, but in addition to other criteria—all of them require, in some way, information about the best items that is provided by another criterion.

The randomesque method allows the test developer to specify a parameter *d* that controls the level of randomness that is applied. Instead of selecting the best item, it randomly selects one among the most informative *d* items (Kingsbury & Zara, 1989). The Sympson-Hetter method starts by organizing the most informative items in a list. It then proceeds along the list, determining through random calculation whether each item should, in fact, be selected. This adds a level of randomization which increases the level utilization of the item pool (Han, 2018).

The fade-away method uses an empirical approach that weighs the probability of an item being chosen by the inverse of its likelihood of appearing (Han, 2012). In other words, through logging items that have previously been displayed to test takers, it is possible to flag items that have had poor utilization, and the less they are used, the more likely they are to be selected by the ISA.

The a-stratified multistage CAT model can also be considered a method to increase item bank utilization (Chang & Ying, 1999). A further development, the α-stratified multistage CAT with *b* blocking, improved item exposure rates and mean square errors by making sure that each stratum had a balanced distribution of *b* parameters so that, in each stage there are items close to the test taker's theta (Chang et al., 2001).

Regardless of these attempts, item exposure remains low in variable-length CATs when the actual test length ends up shorter than the test length that is expected (Wen et al., 2000). The Efficiency Balanced Information Criterion attempts to address this issue, as well as to include

other sources of information (such as the *c* parameter, when available) in item selection. Like the a-stratified multistage CAT, it is not a method for exposure control, but for item selection. Instead of Fisher's information it uses a metric called item efficiency (Han, 2012). This metric controls for a problem with the use of Fisher information whereby the use of items with high discrimination parameters hinders efficiency at the start of the test. At that point, the estimation error is still large, meaning there is no certainty that the theta falls on the estimated theta. Since high *a* parameter items work best discriminating among a narrower range of thetas, and the probable range of thetas is not narrow at that point, such items as wasted. The item efficiency metric works so that higher errors lead to more chance that items with varied discrimination parameters are chosen (Han, 2012).

CAT Software

The technical barrier of entry for developing a CAT test is significantly higher than that for a P&P. Although no programming knowledge is required, it is highly advisable. Still, tools like the Concerto Platform (University of Cambridge Psychometrics Centre, 2022), a free and open-source online adaptive testing platform developed by the University of Cambridge, allow anyone to execute CAT of any complexity.

The Concerto Platform is a customizable web front-end that includes a testing module and a control panel module. The testing module supports multiple-choice and open-ended questions and can be easily extended through the control panel by anyone with knowledge in HTML and JavaScript. The platform itself is a front-end to an SQL Server, where the database is kept, and an R server, which runs the testing logic.

While the Concerto Platform starts from a web application with an underlying R console running the computational features, it is possible to run a CAT program starting from R which uses its own web application. That is what R packages such as catR (Magis & Barrada, 2017)—the package which is run behind Concerto—and mirtCAT (Chalmers, 2017) do. These packages

also provide a testing module that can be run in an Internet browser, but in this approach, administration must be done through the R console, and requires, at least, beginner programing skills.

Both catR and mirtCAT feature various ISAs for dichotomous and polytomous items, customizable stopping rules and several theta estimators (Chalmers, 2017; Magic & Barrada, 2017). The main difference between the two packages is that mirtCAT also allows multidimensional IRT models—that is, IRT models that measure more than one latent variable in each item. The packages vary in which ISAs and IRT models are available specifically, but both feature the most widely used options.

Conclusion

CAT has been one of the most successful testing formats to come out of the ongoing informational revolution in the field of psychological and educational assessment. However, the technical barrier of entry, the elevated initial investment, and the sheer amount of considerations that CAT requires when compared to linear fixed-form tests has meant that even as tests have moved from a P&P format to be delivered digitally, they have retained their traditional format.

One reason may be that although initial studies have reported that no deleterious consequences emerge from taking tests in these digital environments (Kapoor & Welch, 2011), there remains a concern that the implementation of innovative items or the requirement that test takers use different input devices could harm the validity of scores from certain groups of people (de Beer, 2013). Future studies should address whether the observed stability in scores between P&P and CDT tests is limited according to the demographics of the group studied, and whether it is limited to the use of input devices with which test takers are familiar.

Another benefit of this research agenda is elucidating for which item types the use of different input devices may influence examinee scores. The growing dominance of mobile devices in Internet use has led to the development of mobile-friendly versions of CAT modules,

but it is not yet clear whether having some participants use computers and others use cell phones may introduce bias in research data. For now, it's important that every testing program implements its own statistical methodology to ensure fairness in testing across all supported devices.

With regards to the criteria used to adaptatively select the items, there have been new tools created in response to perceived problems with the methods that depend on the Fisher information metric (viz., Kullback-Leibler divergence-based criteria; Cover & Thomas, 1991). Though these criteria have fared better than random item selection in simulation studies (Luzardo, 2019), they have not been compared with the tools they aim to improve upon. New research should assess whether there are actual practical advantages to their use.

Finally, though CAT offers many potential advantages, there are challenges associated with making the best use of it. At the same time, even a simple adaptation of a P&P test has the potential to decrease testing times and increase test accuracy, which may lead to better outcomes and research participation rates. Therefore, it is important that researchers and test developers consider CAT for new testing projects.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied, 15*(2), 163–181. https://doi.org/10.1037/a0015719

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. AERA Publications.

Araneda, S., Lee, D., Lewis, J, Sireci, S., Moon, J. A., Lehman, B., Arslan, B., & Keehner, M. (2022). Exploring Relationships among Test Takers' Behaviors and Performance Using Response Process Data. *Education Sciences, 12*(2), 104–124. https://doi.org/10.3390/educsci12020104

Ariel, A., van der Linden, W. J., & Veldkamp, B. P. (2006). A Strategy for Optimizing Item-Pool Management. *Journal of Educational Measurement, 43*(2), 85–96. https://doi.org/10.1111/j.1745-3984.2006.00006.x

Baker, F. B., & Kim, S.-H. (2017). The Information Function. In *The Basics of Item Response Theory Using R* (pp. 89–104). Statistics for Social and Behavioral Sciences. Springer. https://doi.org/10.1007/978-3-319-54205-8_6

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications, 10*(1), Article 3798. https://doi.org/10.1038/s41467-019-11691-y

Becker, K. A., & Bergstrom, B. A. (2013). Test administration models. *Practical Assessment, Research & Evaluation, 18*(14). ERIC Document № EJ1032767. ERIC Database.

Boone, W. J., & Staver, J. R. (2020). Test Information Function (TIF). In *Advances in Rasch Analyses in the Human Sciences* (pp. 39–55). Springer. https://doi.org/10.1007/978-3-030-43420-5_4

Bradlow, E. T., Weiss R. E., & Cho, M. (1998). Bayesian Identification of Outliers in Computerized Adaptive Tests. *Journal of the American Statistical Association, 93*(443), 910–919. https://doi.org/10.1080/01621459.1998.10473747

Bradlow, E. T., & Weiss, R. E. (2001). Outlier Measures and Norming Methods for Computerized Adaptive Tests. *Journal of Educational and Behavioral Statistics, 26*(1), 85–104. https://doi.org/10.3102/10769986026001085

Bjorkman, M. (1969). On the Ecological Relevance of Psychological Research. *Scandinavian Journal of Psychology, 10*(1), 145–157. https://doi.org/10.1111/j.1467-9450.1969.tb00022.x

Bjorner, J. B., Chang, C.-H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research: An International Journal of Quality-of-Life Aspects of Treatment, Care & Rehabilitation, 16*(Suppl.), 95–108. https://doi.org/10.1007/s11136-007-9168-6

Bulut, H. C. (2021). The Continuity of Students' Disengaged Responding in Low-stakes Assessments: Evidence from Response Times. *International Journal of Assessment Tools in Education, 8*(3), 527–541. https://doi.org/10.21449/ijate.789212

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software, 71*(5), 1–39. https://doi.org/10.18637/jss.v071.i05

Chang, H.-H., & Ying, Z. (1999). a-Stratified Multistage Computerized Adaptive Testing. *Applied Psychological Measurement, 23*(3), 211–222. https://doi.org/10.1177/01466219922031338

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 619–632. https://doi.org/10.1007/s11336-009-9123-2

Choe, E. M., Zhang, J., & Chang, H.-H. (2018). Sequential detection of compromised items using response times in computerized adaptive testing. *Psychometrika, 83*(3), 650–673. https://doi.org/10.1007/s11336-017-9596-3

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT Item Selection Criteria for Polytomous Items. *Applied Psychological Measurement*, *33*(6), 419–440. https://doi.org/10.1177/0146621608327801

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Wiley.

de Beer, M. (2013). The Learning Potential Computerised Adaptive Test in South Africa. In S. Laher & K. Cockcroft (Eds.), *Psychological Assessment in South Africa: Research and applications* (pp. 137–157). Wits University Press. https://doi.org/10.18772/22013015782.15

Diehl, V. A., & Wyrick, M. (2015). The Relationships Between Need for Cognition, Boredom Proneness, Task Engagement, and Test Performance. *SAGE Open*, *5*(2), 1–10. https://doi.org/10.1177/2158244015585606

Drasgow, F., & Olson-Buchanan, J. B. (Eds.). (1999). *Innovations in computerized assessment.* Lawrence Erlbaum Associates Publishers.

Ebel, R. L. (1953). The Use of Item Response Time Measurements in the Construction of Educational Achievement Tests. *Educational and Psychological Measurement*, *13*(3), 391–401. https://doi.org/10.1177/001316445301300303

El-Masri, M. M., Mowbray, F. I., Fox-Wasylyshyn, S. M., & Kanters, D. (2021). Multivariate outliers: A conceptual and practical overview for the nurse and health researcher. *Revue Canadienne de Recherche En Sciences Infirmieres [The Canadian Journal of Nursing Research]*, *53*(3), 316–321. https://doi.org/10.1177/0844562120932054

Ghosh, A., & Lan, A. S. (2021). *BOBCAT: Bilevel Optimization-Based Computerized Adaptive Testing*. arXiv preprint. arXiv: 2108.07386 https://doi.org/10.48550/arXiv.2108.07386

Gören, S., Kara, H., Kara, B. E., & Kelecioğlu, H. (2022). The Effect of Aberrant Responses on Ability Estimation in Computer Adaptive Tests. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi [Journal of Measurement and Evaluation in Education and Psychology]*, *13*(3), 256–268. https://doi.org/10.21031/epod.1067307

Gundry, D., & Deterding, S. (2019). Validity Threats in Quantitative Data Collection with Games: A Narrative Survey. *Simulation and Gaming*, *50*(3), 302–328. https://doi.org/10.1177/1046878118805515

Gundry, D. (2022). *Designing Games to Collect Human–Subject Data* [Doctoral Dissertation, University of York]. Whiterose e-Thesis. https://etheses.whiterose.ac.uk/31655/.

Han, K. T. (2012). An Efficiency Balanced Information Criterion for Item Selection in Computerized Adaptive Testing. *Journal of Educational Measurement*, *49*(3), 225–246. https://doi.org/10.1111/j.1745-3984.2012.00173.x

Han, K. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal Of Educational Evaluation for Health Professions*, *15*(1), Article 7. https://doi.org/10.3352/jeehp.2018.15.7

Heesacker, M., Perez, C., Quinn, M. S., & Benton, S. (2020). Computer-assisted psychological assessment and psychotherapy for collegians. *Journal of Clinical Psychology*, *76*(6), 952–972. https://doi.org/10.1002/jclp.22854

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation*. American Psychological Association. https://doi.org/10.1037/10244-000

Hetter, R. D., Segall, D. O., Bloxom, B. M. (1994). A Comparison of Item Calibration Media in Computerized Adaptive Testing. *Applied Psychological Measurement, 18*(3), 197–204. https://doi.org/10.1177/014662169401800301

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16–25. https://doi.org/10.1111/j.1745-3992.2001.tb00066.x

Hunt, E., Pellegrino, J. W, Frick, R. W., Farr, S. A., Alderton, D. (1988). The ability to reason about movement in the visual field. *Intelligence, 12*(1), 77–100. https://doi.org/10.1016/0160-2896(88)90024-4

Kapoor, S., & Welch, C. (2011). *Comparability of paper and computer administrations in terms of proficiency interpretations* [Paper presentation]. Annual meeting of the National Council on Measurement in Education. New Orleans, LA, United States of America. https://t.ly/_J0YZ.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359–375. https://doi.org/10.1207/s15324818ame0204_6

Kurisu, K., Hashimoto, M., Ishizawa, T., Shibayama, O., Inada, S., Fujisawa, D., Inoguchi, H., Shimoda, H., Inoue, S., Ogawa, A., Akechi, T., Shimizu, K., Uchitomi, Y., Matsuyama, Y., & Yoshiuchi, K. (2022). Development of computer adaptive testing for measuring depression in patients with cancer. *Scientific Reports, 12*(1), Article 8247. https://doi.org/10.1038/s41598-022-12318-x

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*(4), 389–433. https://doi.org/10.1016/S0160-2896(05)80012-1

Linden, W., & Krimpen-Stoop, E. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika, 68*(2), 251–265. https://doi.org/10.1007/BF02294800

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied psychological measurement, 41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203056615

Luecht, R., & Sireci, S. (2011). *A Review of Models for Computer-Based Testing* (ERIC Document № ED465763). ERIC Database.

Luecht, R. M. (2016). Computer-Adaptive Testing. *Wiley StatsRef*. https://doi.org/10.1002/9781118445112.stat06405.pub2

Lushene, R. E., O'Neil, H. F., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment, 38*(4), 353–361. https://doi.org/10.1080/00223891.1974.10119985

Luzardo, M. (2019). Item Selection Algorithms in Computerized Adaptive Test Comparison Using Items Modeled with Nonparametric Isotonic Model. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.). *Quantitative Psychology: 83rd Annual Meeting of the Psychometric Society, New York, NY, 2018*. Springer Proceedings in Mathematics & Statistics. Springer. https://doi.org/10.1007/978-3-030-01310-3_9

Magis D., & Barrada, J. R. (2017). Computerized Adaptive Testing with R: Recent Updates of the Package catR, *Journal of Statistical Software, 76*(1), 1–19. https://doi.org/10.18637/jss.v076.c01

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology, 110*(1), 27–45. https://doi.org/10.1037/edu0000205

Moreno, K. E., & Segall, D. O. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 169–180). American Psychological Association. https://doi.org/10.1037/10244-018

Oberauer, K., Schulze, R., Wilhelm, O., & Süss, H. M. (2005). Working memory and intelligence--their correlation and their relation: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 61–75. https://doi.org/10.1037/0033-2909.131.1.61

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive testing. *Journal of the American Statistical Association, 70,* 351–356. https://doi.org/10.2307/2285821

Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*(1), 1–20. https://doi.org/10.1207/s15324818ame1901_1

Rice, N., Pêgo, J. M., Collares, C. F., Kisielewska, J., & Gale, T. (2022). The development and implementation of a computer adaptive progress test across European countries. *Computers and Education: Artificial Intelligence, 3*, Article 100083. https://doi.org/10.1016/j.caeai.2022.100083

Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence, *Labour Economics, 17*(3), 523–534. https://doi.org/10.1016/j.labeco.2009.04.005

Rudner, L. M. (1998). Item banking. *Practical Assessment, Research, and Evaluation, 6,* Article 4. https://doi.org/10.7275/29ek-tj23

Quiroga, M., Román, F., de la Fuente, J., Privado, J., & Colom, R. (2016). The Measurement of Intelligence in the XXI Century using Video Games. *The Spanish Journal of Psychology, 19,* Article E89. https://doi.org/10.1017/sjp.2016.84

Sarmet, M. M., de Sousa, I. R., de Carvalho, V. T. F., & Castanho, C. D. (2013). *Desenvolvimento e teste de um jogo para estudo do impacto de jogos eletrônicos no comportamento social* [Paper presentation]. Proceedings of the XII SBGames. São Paulo, SP, Brazil. https://www.sbgames.org/sbgames2013/proceedings/cultura/Culture-25_full.pdf.

Schlenker, B. R., & Bonoma, T. V. (1978). Fun and Games. *Journal of Conflict Resolution, 22*(1), 7–38. https://doi.org/10.1177/002200277802200102

Sinharay, S. (2016). Person fit analysis in computerized adaptive testing using tests for a change point. *Journal of Educational and Behavioral Statistics, 41*(5), 521–549. https://doi.org/10.3102/1076998616658331

Sinharay, S. (2017). Detection of item preknowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics, 42*(1), 46–68. https://doi.org/10.3102/1076998616673872

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*(1), 83–117. https://doi.org/10.1023/A:1006985528729

Sireci, S. G., & Zenisky, A. L. (2006). Innovative Item Formats in Computer-Based Testing: In Pursuit of Improved Construct Representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Lawrence Erlbaum Associates Publishers.

Stocking, M. L., Lewis, C. (2000). Methods of Controlling the Exposure of Items in CAT. In W. J. van der Linden, G. A. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice*. Springer. https://doi.org/10.1007/0-306-47531-6_9

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 51*(3), 337–350. https://doi.org/10.1111/1467-9876.00272

Thompson, N. A. (2011). Termination Criteria for Computerized Classification Testing. *Practical Assessment, Research, and Evaluation, 16,* Article 4. https://doi.org/10.7275/wq8m-zk25

Tze, V. M. C., Daniels, L. M., & Klassen, R. M. (2016). Evaluating the Relationship Between Boredom and Academic Outcomes: A Meta-Analysis. *Educational Psychology Review, 28,* 119–144. https://doi.org/10.1007/s10648-015-9301-y

University of Cambridge Psychometrics Center. (2022). *Concerto Platform (Version 5.0)* [Web application]. https://concertoplatform.com/

van der Ham, I. J. M., Faber, A. M. E., Venselaar, M., van Kreveld, M. J., & Löffler, M. (2015). Ecological validity of virtual environments to assess human navigation ability. *Frontiers in Psychology, 6,* Article 637. https://doi.org/10.3389/fpsyg.2015.00637

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63,* 201–216. https://doi.org/10.1007/BF02294775

van der Linden, W. J., Pashley, P. J. (2000). Item Selection and Ability Estimation in Adaptive Testing. In W. J. van der Linden, G. A. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice*. Springer. https://doi.org/10.1007/0-306-47531-6_1

van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement, 28*(5), 317–331. https://doi.org/10.1177/0146621604264870

van der Linden W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics,* 34(3), 378–394. https://doi.org/10.3102/1076998609332107

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of Adaptive Testing*. Springer.

van der Linden, W. J. (2022). Two Statistical Tests for the Detection of Item Compromise. *Journal of Educational and Behavioral Statistics, 47*(4), 485–504. https://doi.org/10.3102/10769986221094789

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some New Item Selection Criteria for Adaptive Testing. *Journal of Educational and Behavioral Statistics*, *22*(2), 203–226. https://doi.org/10.2307/1165378

Vispoel, W. P. (2005). Psychometric Characteristics of Computer-Adaptive and Self-Adaptive Vocabulary Tests: The Role of Answer Feedback and Test Anxiety. *Journal of Educational Measurement, 35*(2), 155–167. https://doi.org/10.1111/j.1745-3984.1998.tb00532.x

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized Adaptive Testing: A Primer* (2nd ed.). Routledge. https://doi.org/10.4324/9781410605931

Wang, C., Weiss, D. J., & Shang, Z. (2019). Variable-Length Stopping Rules for Multidimensional Computerized Adaptive Testing. *Psychometrika*, *84*(3), 749–771. https://doi.org/10.1007/s11336-018-9644-7

Wen, J.-B., Chang, H. H., & Hau, K.-T. (2000). *Adaptation of the a-stratified method in variable length computerized adaptive testing* (ERIC Document No. ED465763). ERIC Database.

Wise, S. L., & Kong, X. (2005). Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L. (2018). Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating. *Education Inquiry, 10*(1), 1–13. https://doi.org/10.1080/20004508.2018.1490127

Xu, X., Chang, H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis* [Paper presentation]. Annual Meeting of the American Educational Research Association, Chicago. https://t.ly/_J0YZ.

Yang. A. C. M., Flanagan. B., & Ogata, H. (2022). Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning. *Computers and Education: Artificial Intelligence, 3,* Article 100104. https://doi.org/10.1016/j.caeai.2022.100104

Zenisky, A., Hambleton, R. K., Luecht, R. M. (2009). Multistage Testing: Issues, Designs, and Research. In W. J. van der Linden, C. A. W. Glas (Eds.), *Elements of Adaptive Testing.* Statistics for Social and Behavioral Sciences. Springer. https://doi.org/10.1007/978-0-387-85461-8_18

Zygouris S., & Tsolaki, M. (2015). Computerized Cognitive Testing for Older Adults: A Review. *American Journal of Alzheimer's Disease & Other Dementias*, *30*(1), 13–28. https://doi.org/10.1177/1533317514522852

MANUSCRITO 4

The Situational Tests of Emotional Intelligence as Computer-Adaptive Tests

Os Testes Situacionais de Inteligência Emocional como um Teste Adaptativo Computadorizado

Abstract

The present study aimed to assess whether an emotional understanding test (STEU) and an emotional management test (STEM) could benefit from being administered as computer-adaptive tests (CATs) without impacting the validity of the test scores. To this end, 11 item selection algorithms (ISAs) were benchmarked for their bias and efficiency. Two simulation studies were run using the same response patterns from the 688 participants used in the original validation study, the same 11 ISAs, but differed in their stopping rules (SRs). For the first study, one simulation was run for each ISA with the SR being standard error lower than 10^{-3} ($\Delta SE < 10^{-3}$), the most commonly used stopping rule criterion. For the second study, *k* simulations were run for each ISA, for each test, with the SR being a fixed number of items between 1 and *k*, where *k* was the total number of items of the relevant test (32 for the STEU and 30 for the STEM). Results of the first simulation showed that testing with all ISAs resulted in accurate ability estimates, all of them having $r > .98$ between their estimates and the estimates calculated with the entire test. The first simulation also showed that, using the best performing ISA, 368 (53.5%) participants needed to answer at least one fewer item without loss of validity. The second study showed that the STEU stood to benefit the most, with the mean standard error (MSE) being minimized six items before the end of the test, though ISAs based on the Kullback–Leibler information performed worse. However, these ISAs also displayed slightly less bias, $r \approx .99$, than the Fisher information-based ones $r \approx .98$. For the STEM, no ISA minimized MSE levels before the end of the test, but up to six fewer items for the STEM and 15 fewer items for the STEU could be administered with a slightly higher tolerance of $\Delta SE < 10^{-2}$. These results indicate that the use of CAT methodology to administer these tests is viable, and EI testing stands to gain from using CAT tests. Future studies should test ISA performance with additional testing constraints.

Keywords: computer-adaptive testing, situational tests of emotional intelligence, emotional intelligence, item selection algorithms, item response theory.

The Situational Tests of Emotional Intelligence as Computer-Adaptive Tests

The Situational Test of Emotion Understanding (STEU) and the Situational Test of Emotion Management (STEM) are two ability EI tests initially described by MacCann and Roberts (2008). The tests measure two dimensions of the Mayer–Salovey–Caruso (MSC) EI theory (Mayer et al., 2012) and were designed partly in response to the fact that, despite the success of the MSC theory, almost all studies conducted within the ability EI paradigm have utilized the Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT; Mayer et al., 2003; Vieira-Santos et al., 2018). While this has granted benefits such as comparability across studies and cultures, it has also introduced several problems, which MacCann and Roberts (2008) sought to address. Namely, the difficulty in discriminating construct variance from test variance, and the employment of two scoring systems which are mutually contradictory.

A recent study sought to make both the STEU and the STEM available for use in a Brazilian setting, and after the cross-cultural adaptation procedure, the final test forms yielded favorable validity evidence (Manuscript 2). Given the novel advantages of computer-adaptive testing (CAT) applied to psychological tests, such as increased validity and precision, immediate feedback and reduction of the exposure effect, the present study aimed to evaluate whether the Brazilian version of the STEU and the STEM can be efficiently administered using CAT algorithms, and, if so, to find out which item selection algorithms (ISAs) most benefit such EI testing.

Computer-Adaptive Testing

CAT is a form of computer-delivered testing that employs an ISA to administer a test in a way that is adaptive to the user's responses (Luecht, 2016). It is contrasted with fixed-form testing, the traditional type of testing in which test items are presented in a predetermined order, which is known as the test form (Luecht & Sireci, 2011). The same test may have different item orders, and therefore different test forms, but in fixed-form testing these test forms are prepared

in advance and are not administered adaptatively to the examinee's performance. Typically, fixed-form testing gradually increases the difficulty of items. When this is done, it may also be called linear testing.

The fundamental principle of CAT is straightforward: the testing platform initially presents the examinee with a question of average difficulty. If the examinee answers correctly, the ISA selects an item with greater difficulty. Conversely, if the examinee answers incorrectly, the algorithm selects an easier item. As the examinee answers more items close to their true aptitude, measurement error decreases (Luecht & Sireci, 2011). Typically, the testing session ends when measurement error reaches a predetermined low value, though additional constraints, such as content constraints, may be specified.

One of the most important properties of an adaptive test is the algorithm employed to select the items that are presented to the examinee, the ISA. ISAs employ a criterion to select the item that should be presented to the examinee after they answer their current item. The two most common sources of information for the measures that are typically used as criteria by ISAs are the Fisher information (FI; Lord, 1980) and the expected posterior variance (EPV; van der Linden, 1998).

The FI is the amount of information that a given item extracts for a given theta interval, represented by $\mathrm{I_j}(\theta)$, which is the information function at item *j* for ability θ. It is given by the probability function multiplied by the square of the logarithm of the probability function of answer *x* given ability θ (Frieden & Gatenby, 2013). An ISA employing the maximum Fisher information (MFI) criterion calculates the FI for each item and selects the items with the largest value. The formula for calculating the FI can be viewed in Equation 1.

$$I_j(\theta) = \int_{\mathbb{R}} (log\, p(x|\theta))^2 \cdot p(x|\theta) dx \qquad (1)$$

The EPV (Bock & Mislevy, 1982) measures the opposite of the FI of the item, that is, it measures the uncertainty associated with the estimation that can be made after that item is

administered. This is why it is mathematically equivalent to the inverse of the FI, as it can be seen in Equation 3. Alternatively, the EPV measure, at item *j*, can be calculated by the argument of the minimum of the difference between the expected value of the ability θ, given the responses to the items that have been administered, and the expected value of the ability θ, given the responses to the items that have been administered plus the next item. This can be seen in Equation 2.

$$EPV = \frac{1}{I_j(\theta)} \tag{2}$$

$$EPV = argmin\left(E(\theta|x_1, \dots, x_k) - E\left(\theta\middle|x_1, \dots, x_j\right)\right) \tag{3}$$

It is, therefore, a measure of how much uncertainty will have shrunk after administering item *j*. The minimum EPV criterion (van der Linden, 1998) selects items that minimize it. However, it is the only criterion that uses the EPV measure, which is contrasted with the fact that many other criteria use the FI (Han, 2018; Veerkamp & Berger, 1997).

The likelihood-weighted information (LWI) is a source of information derived from the FI with an additional source of data: the likelihood function (Veerkamp & Berger, 1997). In a study devised by its creators, the maximum LWI criterion (MLWI) was shown to be better than the MFI at choosing an item that will optimize ability estimates. However, further research has struggled to reproduce these findings (Choi & Schwarz, 2009; Penfield, 2007; Reeve, 2007). The LWI at item *j* for ability θ is the Fisher information multiplied by the likelihood of θ given responses $x_1, \dots, x_k$, where k is the item that has just been answered (Veerkamp & Berger, 1997). It is displayed in Equation 4.

$$LWI_j(\theta) = \int_{-\infty}^{\infty} I_j(\theta) \cdot L(\theta \mid x_1, \dots, x_k) d\theta \tag{4}$$

That same logic can be employed using a Bayesian approach that was also presented by Veerkamp and Berger (1997). The posterior-weighted information (PWI) weighs the information function by the posterior, meaning that, in addition to the likelihood function, it

also takes the prior into account. So, the PWI at item *j* for ability θ is also the FI multiplied by the likelihood function, but it includes a prior function π at current ability estimate θ in the calculation (van der Linden, 1998). This results in the formula displayed in Equation 5.

$$PWI_j(\theta) = \int_{-\infty}^{\infty} I_j(\theta) \cdot \pi(\theta) \cdot L(\theta \mid x_1, \ldots, x_k) d\theta \quad (5)$$

The criterion that employs the PWI, the maximum PWI (MPWI), has also been shown to be superior to the MFI at choosing an item that will optimize ability estimates (van der Linden & Pashley, 2000). But once again, other researchers struggled to reproduce these findings (Choi & Schwarz, 2009; Penfield, 2007; Reeve, 2006).

The expected information (EI) is a newer measure that reproduces the technique of weighing the FI, but using the theta estimation function as weight (Han, 2018). The EI measure for ability θ at item *j* is the FI multiplied by the probability of ability θ given a prior with mean μ and standard deviation σ^2 (Han, 2018). This can be seen in Equation 6.

$$EI_j(\theta) = \int_{-\infty}^{\infty} I_j(\theta) \cdot p(\theta \mid \mu, \sigma^2) d\theta \quad (6)$$

The performance of these criteria have been compared in multiple studies. Choi and Swartz (2009) ran a study which compared the effectiveness of MFI, MLWI, MPWI, MEPV, MEI, and random selection methods. The authors also examined the maximum expected posterior-weighted information (MEPWI), which they found to be mathematically identical to the MPWI. The results of van der Linden and Pashley (2000) were put into question, as they had suggested that the MEPWI had been statistically superior to the MPWI.

In any case, the authors found that the performance of all methods were similar, except for the random algorithm, which was used as a negative control. This also comes into conflict with results from Veerkamp and Berger (1997) that had found the MFI to be inferior to the MLWI and MPWI (Choi & Swartz, 2009). The reason for this conflict may be related to the characteristics of the tests. While Verkamp and Berger (1997) used educational test data, Choi

and Swartz (2009) used a quality of life scale. No studies have compared these different criteria in psychological performance tests.

Psychological Test CATs

Few psychological tests based on CAT report the criterion they use to select items adaptively. When they do report, comparisons are not typically made. Chang (2009) reported the development of a cognitive diagnosis CAT along with two metrics to be used as criteria for item selection. The author compared these criteria with algorithms based on Kullback-Leibler divergence (Cover & Thomas, 1991; Kullback & Leibler, 1951), which, applied to CAT, became known as the Kullback-Leibler information (KL), and with algorithms based on Shannon entropy (Shannon, 1948). They found that the two new criteria had improved performance in some, but not all, situations. However, these algorithms were compared to a negative control, the random algorithm. The author did not compare these new criteria with any of the criteria which uses the FI, such as the MFI, MLWI, MPWI and MEI.

In Brazil, a systematic review by Peres (2019) revealed that few CAT experiences have been reported. Most experiences were dissertations that employed CAT in educational assessment. Only two studies report the use of CAT in psychology: one dedicated to the screening of dyslexia (Santos, 2017), and other to create an item bank for assessing the Big Five factors of personality (Oliveira, 2017). However, the study by Santos (2017) did not actually employ any CAT methodology. Meanwhile, Oliveira (2017) used the Concerto platform to test 525 items through an incomplete block design. The final item bank was composed of 317 items. Since this study aimed to create an item bank for a CAT, there were no indications that an ISA was used, since using incomplete block designs with the specific goal of calibrating all items would require specific items to be administered, making it incompatible with the use of an ISA.

The criteria used by ISA have had different performance in psychological and educational tests, but this has not been studied further. In fact, no psychological tests were

found to use the CAT format at all in Brazil. The present study sought to examine whether two EI tests originally developed by MacCann and Roberts (2018) and then adapted to a Brazilian Portuguese audience by (Manuscript 2) could be efficiently administered as a CAT without impact to score validity. In this study, we aimed to assess the different criteria employed by the ISA to elucidate whether a computerized, adaptive administration of these tests has advantages compared to the traditional form. We also aimed to compare the performance of the criteria to find out the ideal settings for a CAT.

Methods

Participants

The study utilized the same dataset from the EI tests' validation study. The sample comprised 688 participants, overwhelmingly female (81.5%), undergraduate students (21.7%) and single (54.8%). The median age was 23 (MAD = 7.4; Revelle, 2023). Participants were recruited by means of a social media campaign.

Instruments

The Concerto Platform (University of Cambridge Psychometrics Center) and Google Forms (Alphabet Inc.) were utilized for data collection. A Free Consent and a demographic form were also utilized. In total, four psychological tests were utilized. These included the Brazilian Portuguese adaptations of both the Situational Test of Emotional Understanding (STEU) and the Situation Test of Emotional Management (STEM; Manuscript 2), as well as two scales that were used in a different study (Manuscript 2): the Reduced Scale of the Big Five Personality Factors (ER5PF; Passos & Laros, 2015) and the Satisfaction with Life Scale (SWLS; Oliveira et al., 2009).

The Brazilian version of the STEU is a 32-item (originally 42-item, cf. MacCann & Roberts, 2008) multiple-choice assessment that measures an individual's ability to identify emotions in context (Manuscript 2). Each item presents a description of an emotionally charged

situation involving a fictitious character, and the respondent must identify the emotion that the character is most likely to experience in that scenario. For each item, only one alternative is correct. Outcomes are correct or wrong. Item response theory (2-parameter model; Lord & Novick, 1968) fit indices were favorable, M^{2*} (432) = 749.62, p < .001, RMSEA = .033 [CI 95 .029; .037], SRMSR = .045, NNFI = .0948, CFI = .955.

The Brazilian adaptation of the STEM test uses 30 multiple-choice items (originally 44, cf. MacCann & Roberts, 2008) to measure whether individuals can identify the most effective response to an emotionally charged situation among the presented options (Manuscript 2). For the STEM, answer outcomes may be correct or wrong, but multiple answers can also have different scores. Fit indices for item response theory modeling, using the generalized partial credit model (Muraki, 1992), were also favorable, M^{2*} (375) = 408.517, $p < .001$, RMSEA = .011 [CI 95 0; .018], SRMSR = .045, NNFI = .998, CFI = .998.

Correlation between the EAP-calculated scores from the STEU and the STEM are .501. The scores were calculated with the prior set for a Gaussian distribution, $N(0; 1)$.

The R programming language (version 4.2.2) was utilized for data analysis. CAT simulations were performed using the *mirtCAT* package (Chalmers, 2016), and visualizations were created with the *ggplot2* package (Wickham, 2016). All code used in this study was original and is available as supplementary material.

Procedure

Recruitment for the study was conducted through social media platforms, in which participants were provided with a link to the testing platforms. Data collection initially took place on the Concerto Platform but was later switched to Google Forms due to hosting issues. The forms and tests were presented on separate pages and only on the Google Forms platform could participants resume previous sessions. Participants could only submit the form when it had been fully filled out.

Analysis

CAT Simulations

Two different simulation studies were run. Their details are displayed in Table 1.

Table 1

Input Parameters of Computer-Adaptive Testing Simulation Studies 1 and 2.

Characteristics	Study 1	Study 2
Item parameters	The item parameters for both the STEU and the STEM were calibrated in Manuscript 2 (2023).	
Algorithms	MFI, MLWI, MPWI, MEI, IKL, IKLn, IKLP, and IKLPn	
Stopping Rule	$\Delta SE < .001$	Fixed number of items
Repetition	One simulation per algorithm per test.	*k* simulations per algorithm per test, where *k* = number of items per test.
Number of simulations	1 simulation per 11 algorithms per each of the two tests, $11 \times 2 = 22$ simulations	32 simulations for the STEU and 30 simulations for the STEM per 11 algorithms, $32 \times 11 + 30 \times 11 = 682$ simulations
Analysis	Correlation between theta estimate per algorithm and the theta estimate with the entire test. Frequency of number of items required by each algorithm to reach the stopping rule for each participant.	Mean standard error levels at each number of items administered.
Purpose	Bias Test length reduction	Precision estimate Test length reduction

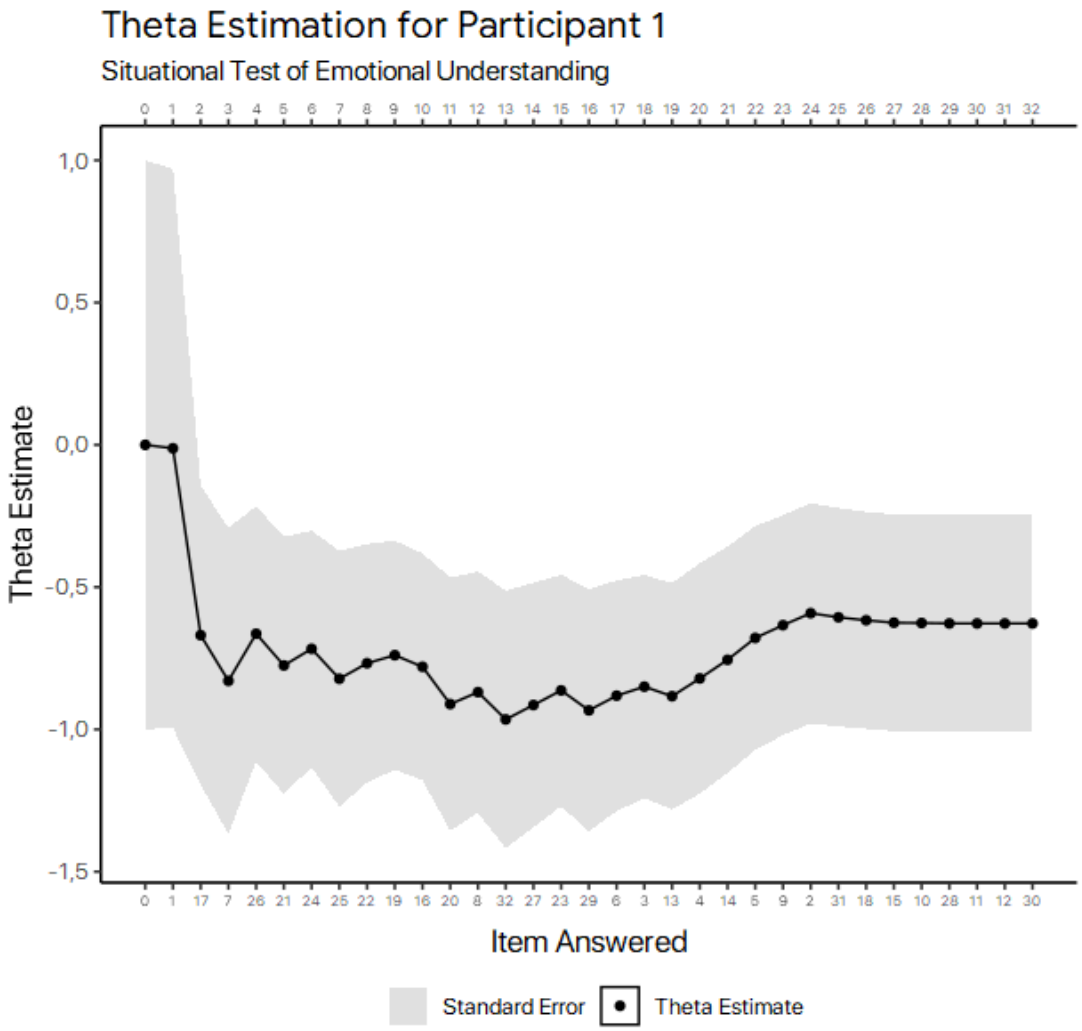
For both studies, the mirtCAT package was used to generate answers based on the empirical data response pattern collected during the validation study, simulating new data collection using an ISA for each participant. The IRT parameters were also calibrated in the validation study. In the first study, the simulation procedure was repeated once for each ISA for each of the two tests.

The algorithms that were tested were the maximum Fisher information (MFI), the minimum expected posterior variance (MEPV), the maximum likelihood-weighted information (MLWI), the maximum posterior-weighted information (MPWI), the maximum expected information (MEI), and the integration-based Kullback–Leibler criteria with or without prior density, and with or without root-n Weight (corresponding to algorithms IKLPn, IKLP, IKLn and IKLn, where the "P" denotes the prior density weight and, the "n", the root-n weight).

Each simulation study yielded one or more databases containing data for each of the participants as if they had answered a CAT test, a database containing the items that they answered until they reached the stopping rule, the estimated theta values after each item answered, the standard error (SE) of each estimate, and the final theta estimate. An example of simulated test administration can be found in Figure 1.

Figure 1

Theta Estimates and Standard Errors After Each Response from One Simulated Participant.



For each ISA, the correlation between the final theta found and the estimated theta was calculated. The same was done between the estimated SE and the number of items administered.

For the second study, each algorithm was simulated *k* times, where *k* was the test's maximum number of items (32 for the STEU and 30 times for the STEM). This was necessary because this study employed a different stopping rule: the test would stop when it administered a fixed number of items, at every number between one and the number of items of the test. This made it possible to calculate the mean SEs for each ISA at each fixed number of items

administered, simulating alternative testing conditions, such as content constraints. The standard deviations of the SEs were also calculated.

Results

All ISAs reached a correlation greater than $r = .98$ between the theta estimates they produced and the theta estimates calculated using the entire test. The correlation between the number of items required for the test to end and the SE was also calculated, and was significant for all ISAs, for both tests, ranging between .412 (medium-sized) and .71 (high). These statistics are shown in Table 2.

Table 2

Correlation Between Abilities Estimated Utilizing Item Selection Algorithms and Abilities Estimated Using All Test Items.

	STEU		STEM	
	Correlation		Correlation	
Algorithm	Theta	N-SE	Theta	N-SE
MFI	.982	.601	.996	.624
MEPV	.989	.594	.994	.614
MLWI	.985	.580	.994	.606
MPWI	.988	.595	.993	.611
MEI	.983	.599	.994	.607
IKL, IKLn, IKLp, IKLPn	.994	.423	.998	.710
Random	.992	.435	.996	.573
Sequential	.994	.421	.998	.710

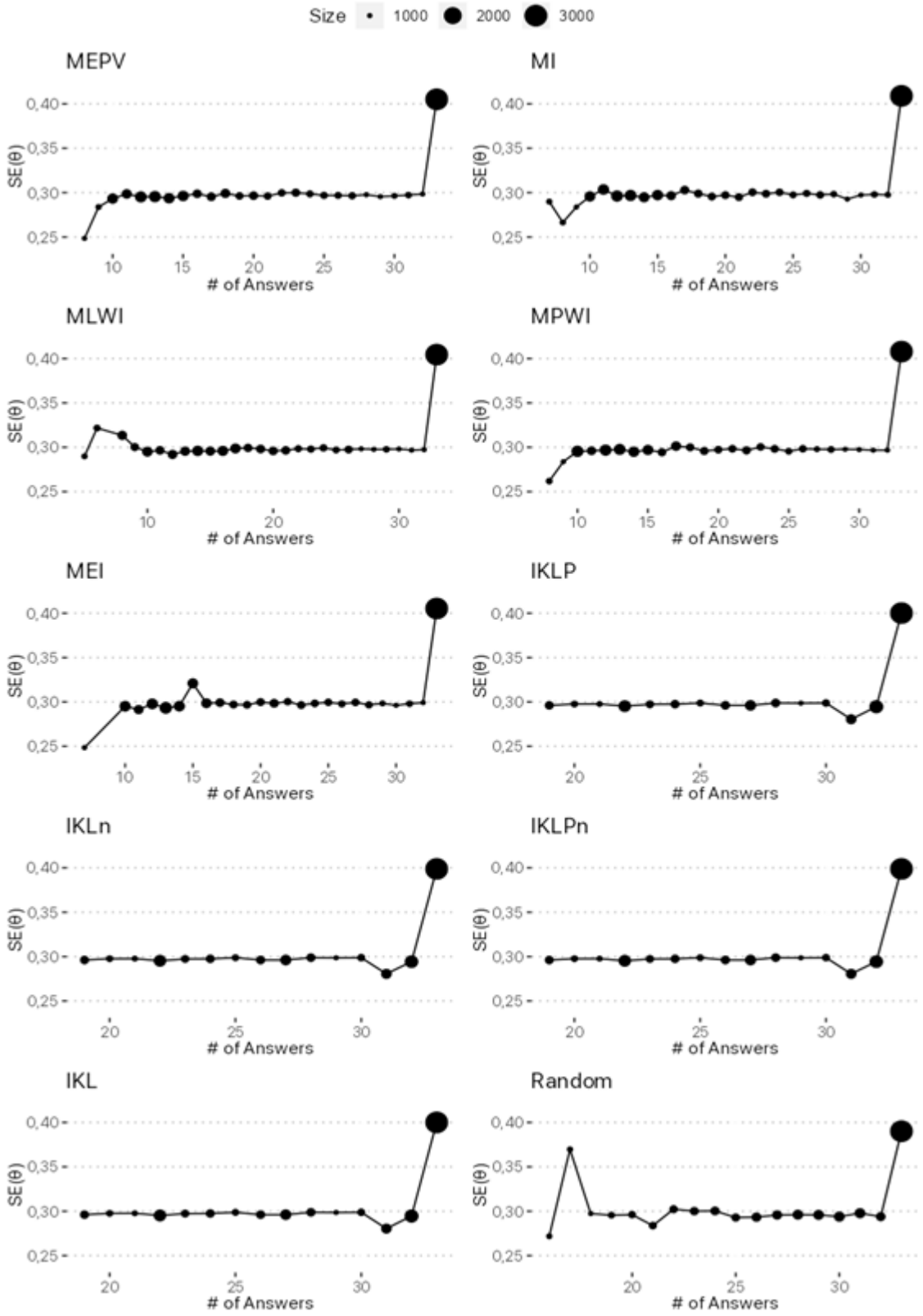
Notes. MFI = maximum Fisher information, MEPV = minimum expected posterior variance, MLWI = maximum likelihood-weighed information, MPWI = maximum posterior-weighted information, MEI = maximum expected information, IKL = integration-based Kullback–Leibler criteria, IKLn = IKLn with root-n weight, IKLP = IKLP with prior density weight, IKLPn = IKLP with root-n weight and prior density weight. All KL-based algorithms yielded the same correlation sizes.

As shown in Figure 2, the overall SE outcomes of the ISAs were close between the FI-based algorithms MEPV, MFI, MLWI and MPWI, and the KL-based algorithms. This is

expected, as the stopping rule is designed for the SE to be minimized—however, the KL-based algorithms administered more items to reach these low SE levels. In general, the pattern in the plot indicates that when item information matches the participants' ability, the test ends quickly with low SE. Until the ISA had to serve all items, information about the participants' proficiency was able to reach the required $\Delta SE < 10^{-1}$ criterion. When all items are served, that means the criterion was not met, but this does not limit the SE; it can vary freely. Indeed, different ISAs have different maximum SEs.

Figure 2

Association of the Mean Standard Error and Number of Responses of the Item Selection Algorithms.



Frequency statistics for the number of items required to reach the stopping rule are shown in Tables 3 and 4 for the STEU and STEM tests, respectively. For the STEU, the KL-based algorithms did not reach the required $\Delta SE < .001$ rule to end the test for any simulated

participant before item 19. This contrasts with the MFI, MEPV, MLWI, MPWI and MEI methods, which reached that requirement for between 37.64% and 40.12% simulated participants before item 19, including between 5.32% and 10.17% estimations of simulated participants reaching the requisite SE difference rule before item 11. All algorithms had simulated participants reach the stopping rule at least once before the last item of the test, with the number of participants varying per algorithm—the number ranged from between 299 for the random algorithm to 368 for the MLWI algorithm. Among the adaptive algorithms, the worst performance was a tie between the IKLn and IKLPn ISAs, with 326 simulated participants reaching the stopping rule before reaching the last item of the test.

Table 3

Number of Items Administered by Each Algorithm for the Situational Test of Emotional Understanding.

	Number of Items Administered					
Algorithm	4-9	10-17	18-24	25-31	1-31	32
IKL	0	0	124	207	331	357
IKLn	0	0	124	202	326	362
IKLP	0	0	124	207	331	357
IKLPn	0	0	124	202	326	362
MEI	37	230	67	26	360	328
MEPV	36	237	60	19	352	336
MFI	40	236	63	29	368	320
MLWI	70	189	73	24	356	332
MPWI	54	218	66	24	362	326
Random	0	5	82	212	299	389
Total	237	1115	907	1152	3411	3469

Notes. Sample sizes = 688. Item number grouping chosen to highlight differences.

The STEM simulations were less successful in reducing the number of items needed to reach the stopping rule, but test lengths were still significantly reduced. The KL-based algorithms again had identical performance and only started reaching the stopping rule when 23 items were administered. This contrasts with the MFI, MEPV, MLWI, MPWI and MEI

methods, which reached the stopping rule for around 22% of simulated testing sessions before item 23.

Once again, all algorithms had simulated participants reach the stopping rule at least once before the last item of the test. The number of such participants ranged from 172 for the random algorithm to 187 for the MLWI algorithm. Notably, for the STEM, the worst performance was a tie between the KL-based ISAs, which performed worse than the random algorithm.

Table 4

Number of Items Administered by Each Algorithm for the Situational Test of Emotional Management.

	Number of Items Administered			
Algorithm	3–22	23–29	1–29	30
IKL	0	169	169	519
IKLP	0	169	169	519
IKLPn	0	169	169	519
IKLn	0	169	169	519
MEI	153	33	186	502
MEPV	155	28	183	505
MFI	156	31	187	501
MLWI	154	31	185	503
MPWI	155	32	187	501
Random	113	59	172	516

Notes. Sample sizes = 688. Item number grouping chosen to highlight differences.

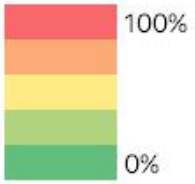
The efficient performance of FI-based algorithms was also observed on Study 2. For the STEU, the best ISAs reached the minimum mean SE on simulations with 26 items. This indicates that, on average, precision is already at its highest value for the best ISAs even before the last six items had been administered. Comparatively, the same mean SE level is only reached by the KL-based algorithms when simulations had administered all items. The MEPV did not perform correctly, as mean SE estimates varied back and forth. These results can be seen in Figure 3.

Figure 3

Mean Standard Error per Number of Items for Each Item Selection Algorithm for the Situational Test of Emotional Understanding.

Standard Error - Mean

	MI	MEPV	MLWI	MPWI	MEI	IKL	IKLP	IKLn	IKLPn	Random
1	1,016	1,016	1,016	1,016	1,016	1,016	1,016	1,016	1,016	1,016
2	0,680	0,798	0,748	0,680	0,773	0,778	0,778	0,778	0,778	0,904
3	0,635	0,597	0,658	0,610	0,550	0,779	0,779	0,779	0,779	0,841
4	0,386	0,496	0,606	0,467	0,491	0,638	0,638	0,638	0,638	0,770
5	0,441	0,470	0,516	0,455	0,453	0,584	0,584	0,584	0,584	0,688
6	0,423	0,448	0,464	0,436	0,423	0,582	0,582	0,582	0,582	0,653
7	0,406	0,433	0,438	0,414	0,409	0,536	0,536	0,536	0,536	0,615
8	0,394	0,424	0,419	0,392	0,395	0,530	0,530	0,530	0,530	0,575
9	0,384	0,398	0,405	0,386	0,388	0,516	0,516	0,516	0,516	0,570
10	0,377	0,411	0,393	0,379	0,381	0,505	0,505	0,505	0,505	0,539
11	0,372	0,377	0,383	0,375	0,374	0,504	0,504	0,504	0,504	0,535
12	0,367	0,400	0,378	0,368	0,370	0,496	0,496	0,496	0,496	0,498
13	0,362	0,366	0,370	0,364	0,366	0,493	0,493	0,493	0,493	0,512
14	0,358	0,400	0,365	0,361	0,365	0,478	0,478	0,478	0,478	0,479
15	0,356	0,358	0,363	0,360	0,361	0,469	0,469	0,469	0,469	0,487
16	0,352	0,390	0,357	0,356	0,358	0,446	0,446	0,446	0,446	0,450
17	0,349	0,353	0,355	0,355	0,355	0,407	0,407	0,407	0,407	0,467
18	0,347	0,384	0,353	0,352	0,352	0,392	0,392	0,392	0,392	0,433
19	0,345	0,350	0,352	0,349	0,350	0,392	0,392	0,392	0,392	0,447
20	0,345	0,383	0,352	0,348	0,348	0,390	0,390	0,390	0,390	0,413
21	0,346	0,349	0,350	0,347	0,346	0,378	0,378	0,378	0,378	0,439
22	0,346	0,382	0,348	0,348	0,346	0,375	0,375	0,375	0,375	0,402
23	0,346	0,347	0,348	0,347	0,346	0,372	0,372	0,372	0,372	0,420
24	0,348	0,377	0,350	0,349	0,346	0,369	0,369	0,369	0,369	0,388
25	0,345	0,345	0,346	0,346	0,345	0,367	0,367	0,367	0,367	0,410
26	0,344	0,377	0,346	0,345	0,344	0,360	0,360	0,360	0,360	0,370
27	0,343	0,344	0,344	0,344	0,344	0,358	0,358	0,358	0,358	0,404
28	0,343	0,378	0,344	0,344	0,343	0,358	0,358	0,358	0,358	0,361
29	0,343	0,344	0,344	0,344	0,344	0,356	0,356	0,356	0,356	0,393
30	0,343	0,377	0,344	0,344	0,344	0,353	0,353	0,352	0,352	0,354
31	0,343	0,344	0,344	0,344	0,344	0,346	0,346	0,346	0,346	0,384
32	0,344	0,378	0,344	0,344	0,344	0,344	0,344	0,344	0,344	0,344



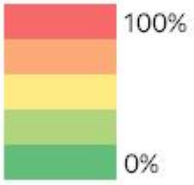
For the STEM, mean SEs took longer to stabilize. The MEI, MFI, MLWI, and MPWI methods reached the stopping rule only when 29 items had been administered. However, the MEPV algorithm did not exhibit the erratic behavior observed on the STEU. Meanwhile, the IKL methods did not stabilize at a low SE pattern until simulations were done with all test items. These results can be seen in Figure 4.

Figure 4

Mean Standard Error per Number of Items for Each Item Selection Algorithm for the Situational Test of Emotional Management.

Standard Error - Mean

	MI	MEPV	MLWI	MPWI	MEI	IKL	IKLP	IKLn	IKLPn	Random
1	0,887	0,887	0,887	0,887	0,887	0,887	0,887	0,887	0,887	0,887
2	0,752	0,742	0,762	0,777	0,742	0,828	0,828	0,828	0,828	0,815
3	0,637	0,651	0,720	0,665	0,669	0,759	0,759	0,759	0,759	0,758
4	0,585	0,593	0,631	0,603	0,620	0,713	0,713	0,713	0,713	0,710
5	0,547	0,555	0,576	0,564	0,569	0,689	0,689	0,689	0,689	0,670
6	0,519	0,525	0,544	0,533	0,534	0,654	0,654	0,654	0,654	0,634
7	0,496	0,503	0,518	0,508	0,507	0,625	0,625	0,625	0,625	0,607
8	0,477	0,483	0,495	0,489	0,486	0,591	0,591	0,591	0,591	0,582
9	0,462	0,467	0,477	0,471	0,471	0,568	0,568	0,568	0,568	0,560
10	0,449	0,453	0,462	0,456	0,457	0,556	0,556	0,556	0,556	0,539
11	0,439	0,441	0,448	0,445	0,444	0,523	0,523	0,523	0,523	0,524
12	0,429	0,431	0,437	0,434	0,433	0,510	0,510	0,510	0,510	0,506
13	0,419	0,422	0,427	0,424	0,424	0,493	0,493	0,493	0,493	0,490
14	0,411	0,413	0,418	0,415	0,415	0,478	0,478	0,478	0,478	0,480
15	0,403	0,406	0,409	0,407	0,407	0,461	0,461	0,461	0,461	0,466
16	0,396	0,399	0,402	0,400	0,400	0,448	0,448	0,448	0,448	0,454
17	0,390	0,392	0,395	0,393	0,393	0,439	0,439	0,439	0,439	0,443
18	0,385	0,387	0,389	0,387	0,388	0,431	0,431	0,431	0,431	0,436
19	0,379	0,381	0,383	0,382	0,383	0,415	0,415	0,415	0,415	0,426
20	0,374	0,376	0,378	0,377	0,377	0,412	0,412	0,412	0,412	0,415
21	0,370	0,372	0,374	0,372	0,373	0,408	0,408	0,408	0,408	0,408
22	0,367	0,368	0,369	0,369	0,368	0,400	0,400	0,400	0,400	0,400
23	0,363	0,364	0,366	0,365	0,365	0,385	0,385	0,385	0,385	0,394
24	0,361	0,361	0,362	0,362	0,361	0,377	0,377	0,377	0,377	0,386
25	0,358	0,359	0,359	0,359	0,359	0,371	0,371	0,371	0,371	0,381
26	0,356	0,357	0,357	0,357	0,357	0,370	0,370	0,370	0,370	0,374
27	0,354	0,355	0,355	0,355	0,355	0,363	0,363	0,363	0,363	0,368
28	0,353	0,353	0,353	0,353	0,353	0,359	0,359	0,359	0,359	0,362
29	0,352	0,352	0,352	0,352	0,352	0,355	0,355	0,355	0,355	0,357
30	0,351	0,351	0,351	0,351	0,351	0,351	0,351	0,351	0,351	0,351



Additionally, mean SEs for experimental ISAs are significantly lower than the negative control (the random algorithm), even from simulations with just two items. In other words, any algorithm that made use of information—whether FI or KL and whether weighted or not—was more effective than the algorithm that randomly selected items, even if only one item is selected.

Discussion

This study aimed to determine whether the Brazilian versions of the STEU and the STEM could be efficiently administered using a CAT mechanism without loss of validity and precision. Given the relationship between increased measurement validity in CAT and reduced testing times. To this end, we employed a methodology based on simulated administrations of real-world data for which test results were previously known. This allowed us to assess the

extent to which administering fewer items could accurately estimate proficiency levels while also including the measurement error expected in the response patterns collected in a typical test administration session.

In the first study simulations were run without establishing a fixed number of items that would be administered to each participant. Different ISAs could administer whichever items they chose, and the number of items administered would be a consequence of the stopping rule being reached, which was the same for all simulations ($\Delta SE < 10^{-3}$). Using this method, it was possible to determine that the correlation between the thetas estimates in the simulations and those estimated using the entire test was over .9 for all ISAs.

Assuming that reducing testing times would be one of the main goals of this step, the homogeneity among the correlation between thetas could hide different performances. For example, even the random algorithm reached elevated levels of correlation, outperforming several other algorithms. This measure alone is inadequate to measure the ISAs' performance, since the values of these correlations does not take into account whether an algorithm's ability to reach ability estimates so close to those obtained using full information was due to its simulations requiring the administration of additional items. As such, this did not yet allow for the study of the expected SE for each number of items administered for every participant response pattern.

To address this issue, the second study involved constraining analyses to an exact number of items administered for each ISA. Analyzing the results of this study helped determine that the seemingly positive results of the random algorithm are associated with it requiring the administration of many more items to reach the stopping rule. In this step, every single ISA had significantly better mean SEs, from the first item they selected all the way to the one before the last. At this point, every ISA, including the random algorithm, reached the minimum SEs possible for the test they were simulating, since all items had been administered.

Indeed, the random algorithm could only reach the minimum SEs for all simulated participants at that point.

Considering that the ultimate goal of an examination is to estimate the examinee's latent trait, the results of the this study suggest that it is possible to achieve optimal, or optimal or near-optimal mean SE levels with significantly fewer items administered in either of the tests, provided that informative ISAs are used. Using the $\Delta SE < 10^{-2}$ stopping rule, which is an often acceptable setting, as many as 15 and 6 fewer items could be administered by the best-performing ISAs in the STEU and the STEM, respectively. This represents a reduction of 21 items out of 62, or 34% of total items answered.

This analysis helps to understand the impact of the mean SE. However, in practical terms, without any loss to the SE, the first study already revealed that more than half of all participants could be expected to need to answer at least one fewer item for the most efficient ISAs. For the MFI algorithm, this was 368 participants, versus 320 who had to answer all items.

In comparing these results with other research, our data does not seem to agree with papers by the authors of nearly every algorithm competing with the MFI algorithm (e.g., van der Linden, 1998; van der Linden & Pashley, 2000; Veerkamp & Berger, 1997). Instead, our findings are consistent with those reported by Choi and Swartz (2009), who found that the MFI, MLWI, and MPWI were roughly close in performance. Excluding the its SE performance at every second item, the MEPV algorithm performed as efficiently as the best performing algorithm, the MFI. A study that compared the MFI, MPWI and MEPV using testlets also reached the conclusion that these algorithms' performance is overall similar (Murphy et al., 2010)

Indeed, considering Choi and Schwartz's (2009) finding that the MEPWI was mathematically identical to the MPWI despite previous findings that it was worse than the MPWI, it is possible that differences in the transformation of the mathematical formulae into

computer algorithms could account for any reported differences in the literature. We hypothesize that a similar issue may be responsible for the issue with the ability estimate at every other item found in the MEPV results for the present study.

Methods such as the MEI (Han, 2018), MLWI, and MPWI (van der Linden, 1998) all work similarly, making use of the FI weighed by a probability or likelihood function and possibly a prior. The MLWI algorithm weighs the same information by the likelihood of theta given the response pattern and can be viewed in Equation 3. Given that all the prior for all calculations was a normal distribution N (0;1), it is expected that these methods would have similar performance. Further studies should examine the performance of these methods under different priors.

Finally, the algorithms based on the KL had the poorest mean SE performance among the non-random ISAs. The IKL algorithm (Cover & Thomas, 1991) was developed to address situations for which the FI would not be adequate (Chang & Ying, 1996). Specifically, its use is recommended during testing at points in which the test's current ability estimate is not necessarily likely to be the true theta, such as at the start of the test, when a generic prior is utilized. This has been implemented by estimating global information at each item for KL-based algorithms, whereas the MFI algorithm is based directly on how much local information each item carries (Chang & Ying, 1996)

Local information refers to the amount of information that items provide at theta values close to the current theta estimate. In contrast, global information encompasses all the information that items provide, including information farther from that point. These concepts are similar and there is a mathematical relationship between them, as shown in Equation 7 (Dabak & Johnson, 2003).

$$\boldsymbol{I_j(\theta) = \frac{\partial^2}{\partial \theta^2} K^{(n)}(\theta||\theta)|_{\theta-\theta_0}} \quad (7)$$

Where $K^{(n)}(\theta||\theta)$ is the KL.

The MFI is defined as the inverse of the expected value of the second derivative of the log-likelihood function with respect to theta (Ly et al., 2017). In contrast, when applied to the item selection (Cover & Thomas, 1991), the Kullback and Leibler (1951) divergence is the expected value of the logarithmic difference between the likelihood and the theta estimation function (Dabak & Johnson, 2003).

Therefore, the KL is the second derivative of the FI. The two will be equal when the KL is calculated for a small interval of theta. Since KL considers information for a larger interval, the MFI's performance will be better when the test taker's true ability is close to the ability estimate. In this way, the KL's focus on global information may lead it to select items that contain information not relevant to the test's theta estimate of the test taker.

However, Chang and Ying's (1996) study revealed that KL-based algorithms performed better in terms of mean squared error and bias under many circumstances. In their study, mean SE was not measured; instead, mean squared error and bias were calculated in relation to the true theta score. In our study, correlations between the various ISA and the true theta score revealed that KL-based algorithms had marginally better bias (mean r = .994) when compared to the FI-based algorithm that had the highest correlation (r = .989).

Another factor that may have influenced the results is the stopping rule used in this study. We adopted the default stopping rule of $\Delta\text{SE} < 10^{-3}$, which was also used in Choi and Swartz's (2009) study. The use of a ΔSE criterion may have contributed to the superior performance of the MFI algorithm. While some studies have examined the effect of stopping rules (e.g., Babcock & Weiss, 2012), no study has assessed their potential effect on ISAs. Future studies that examine ISA performance under different stopping rules, such as those based on the difference between ability estimates, are therefore warranted.

Additionally, this study did not attempt to use content balancing rules. However, their effect is largely independent from the performance of the ISAs except insofar as they control the minimum number of items that must be administered.

There are several limitations to this study. For instance, although using empirical data as response patterns for the simulations has been considered a novel means of accruing justification for further CAT implementations of the STEU and the STEM, no study has investigated whether there is a negative impact of doing so with the same data utilized in parameter calibration. While the utilization of cross-validation methods has been considered largely superseded by IRT fit measures such as the M^2* (Cai & Hansen, 2013) and C^2 (Cai & Monro, 2014) model fit measures, the S-χ^2 (Kang & Chan, 2007) item fit measure, and the traditional measures *Iz* and *Zh* being utilized for person fit (Drasgow et al., 1985; Felt et al., 2017), all of which were employed in the validation study for the STEU and the STEM (Manuscript 2), previous studies have suggested various situations in which repeating samples would be considered a limitation (de Rooiji & Weeda, 2020). For instance, Cunha (2019) showed that constructing a prediction model based on the same data in which the predictor was produced led to estimates below the expected risk value. While the present study did not employ any further parameter calculation using the same covariance matrix, it is possible that some other problem may have arisen.

With respect to making sure that the tests benefit from CAT's many reported benefits (Manuscript 3), one important limitation is the size of the item bank. Several benefits have been associated with large item pools, such as the reduction of the exposure effect, which reduces the bias associated with individuals taking the same test more than once. However, since both the items and IRT parameters have been published (Manuscript 2), further studies may construct additional items which can be easily calibrated using incomplete block designs (Ariel et al., 2006).

In summary, regardless of individual ISA performance, to the extent to which the goal of administering the Brazilian adaptation of the STEU and the STEM is the estimation of the latent traits emotional understanding and emotional management, these results show that the CAT versions of the STEU and the STEM were able to estimate these abilities with the same precision as the completed test versions while benefitting from the advantages CAT confers. This development should allow researchers to achieve increased measurement validity while reaching more participants by offering more attractive research participation opportunities with lower testing times. Finally, this study employed a novel methodology for seeking evidence of benefits that test developers stand to gain when adapting tests to CAT. We hope that psychological test developers will take notice and employ similar techniques.

References

Ariel, A., van der Linden, W. J., & Veldkamp, B. P. (2006). A Strategy for Optimizing Item-Pool Management. *Journal of Educational Measurement, 43*(2), 85–96. https://doi.org/10.1111/j.1745-3984.2006.00006.x

Babcock, B., & Weiss, D. J. (2012). Termination Criteria in Computerized Adaptive Tests: Do Variable-Length CATs Provide Efficient and Effective Measurement? *Journal of Computerized Adaptive Testing, 1*(1–5), 1–18. https://doi.org/10.7333/1212-0101001

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT Item Selection Criteria for Polytomous Items. *Applied Psychological Measurement, 33*(6), 419–440. https://doi.org/10.1177/0146621608327801

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213–229. https://doi.org/10.1177/014662169602000303

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619–632. https://doi.org/10.1007/s11336-009-9123-2

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Wiley.

Dabak, A. G., & Johnson, D. H. (2003). *Relations between Kullback-Leibler distance and Fisher information*. Rice University. http://dhj.rice.edu/files/2014/07/distance.pdf

de Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science, 3*(2), 248–263. https://doi.org/10.1177/2515245919898466

Felt, J. M., Castaneda, R., Tiemensma, J., & Depaoli, S. (2017). Using Person Fit Statistics to Detect Outliers in Survey Research. *Frontiers in Psychology, 8*, 863. https://doi.org/10.3389/fpsyg.2017.00863

Frieden, B. R., & Gatenby, R. A. (2013). Principle of maximum Fisher information from Hardy's axioms applied to statistical systems. *Physical Review E, 88*(4), Article 042144. https://doi.org/10.1103/PhysRevE.88.042144

Han, K. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal Of Educational Evaluation for Health Professions, 15*(1), Article 7. https://doi.org/10.3352/jeehp.2018.15.7

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied Psychological Measurement, 41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Lord, F. M., & Novick, M. R. (1968). *Statistical theory of mental test scores*. Addison-Wesley.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203056615

Luecht, R., & Sireci, S. (2011). *A Review of Models for Computer-Based Testing* (ERIC Document № ED465763). ERIC Database.

Luecht, R. M. (2016). Computer-Adaptive Testing. *Wiley StatsRef*. https://doi.org/10.1002/9781118445112.stat06405.pub2

Ly, A., Marsman, M., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology, 80*, 40–55. https://doi.org/10.1016/j.jmp.2017.05.006

MacCann, C., & Roberts, R. D., (2008). New Paradigms for Assessing Emotional Intelligence: Theory and Data, *Emotion, 8*(4), 540–551. https://doi.org/10.1037/a0012746

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27–45. https://doi.org/10.1037/edu0000205

Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. J. Sluyter (Eds.), *Emotional development and emotional intelligence: Educational implications* (pp. 3–31). Basic Books.

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3*(1), 97–105. https://doi.org/10.1037/1528-3542.3.1.97

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Murphy, D. L, Dodd, B. G., & Vaughn, B. K. (2010). A Comparison of Item Selection Techniques for Testlets. *Applied Psychological Measurement, 34*(6), 424–437. https://doi.org/10.1177/0146621609349804

Oliveira, G. F., Barbosa, G. A., Souza, L. E., Costa, C. L., Araújo, R. C., & Gouveia, V. V. (2009). Satisfação com a vida entre profissionais da saúde: correlatos demográficos e laborais, *Revista de Bioética*, *17*(2), 319–334.

Oliveira, C. M. (2017). *Construção e busca de evidências de validade de um banco de itens de personalidade para testagem adaptativa desenvolvido a partir dos princípios do desenho universal* [Doctoral dissertation, Universidade Federal da Santa Catarina]. Repositório Institucional da UFSC. https://repositorio.ufsc.br/bitstream/handle/123456789/187269/PPSI0766-T.pdf

Passos, M. F. D., & Laros, J. (2015). Construção de uma escala reduzida de Cinco Grandes Fatores de personalidade, *Avaliação Psicológica, 14*(1), 115–123. https://doi.org/10.15689/ap.2015.1401.13

Penfield, R. D. (2006). Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*, 1–20. https://doi.org/10.1207/s15324818ame1901_1

Peres, A. J. de S. (2019). Testagem adaptativa por computador (CAT): Aspectos Conceituais e um Panorama da Produção Brasileira. *Examen: Política, Gestão E Avaliação Da Educação*, *3*(3), 66–86. https://examen.emnuvens.com.br/rev/article/view/10

Reeve, B. B. (2006). Special Issues for Building Computerized-Adaptive Tests for Measuring Patient-Reported Outcomes: The National Institute of Health's Investment in New Technology. *Medical Care, 44*(11, Suppl 3), S198–S204. https://doi.org/10.1097/01.mlr.0000245146.77104.50

Santos, J. S. (2017). *Mensuração de habilidades cognitivas preditoras do desenvolvimento de leitura em crianças através de jogos educacionais* [Master's thesis, Universidade Federal de Campina Grande]. Repositório da UFCG.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201–216. https://doi.org/10.1007/BF02294775

van der Linden, W. J., & Pashley, P. J. (2000). Item Selection and Ability Estimation in Adaptive Testing. In W. J. van der Linden, G. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice*. Springer. https://doi.org/10.1007/0-306-47531-6_1

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some New Item Selection Criteria for Adaptive Testing. *Journal of Educational and Behavioral Statistics*, *22*(2), 203–223. https://doi.org/10.2307/1165378

Vieira-Santos, J., Lima, D. C., Sartori, R. M., Schelini, P. W., & Muniz, M. (2018). Inteligência emocional: revisão internacional da literatura. *Estudos Interdisciplinares em Psicologia*, *9*(2), 78–99. https://doi.org/10.5433/2236-6407.2018v9n2p78

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd ed.). Springer.

Wise, S. L. (2018). Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating. *Education Inquiry, 10*(1), 1–13. https://doi.org/10.1080/20004508.2018.1490127

Discussão Geral

O objetivo deste trabalho foi adaptar um instrumento de inteligência emocional (IE) para o contexto brasileiro, construir um argumento que defenda a validade do instrumento com base em evidências empíricas, e avaliar a viabilidade do instrumento adaptado ser administrado como um teste adaptativo computadorizado. Para alcançar este objetivo, o estudo foi conduzido da seguinte maneira: primeiro, foi realizada a adaptação do teste por meio de um procedimento de tradução reconhecido no meio acadêmico. Em seguida, foram coletadas evidências de validade, e realizou-se a argumentação da validade dos testes adaptados com base em várias fontes de evidência. Por fim, foi feita a demonstração da capacidade da testagem adaptativa computadorizada (CAT) de mensurar os construtos estudados de forma comparável ao teste aplicado de forma tradicional com os benefícios do método de administração adaptativo, incluindo o aumento de eficiência na testagem.

Para subsidiar o procedimento de adaptação, que inevitavelmente contém elementos subjetivos na medida em que inclui decisões editorais na forma como são implementadas as recomendações emitidas pelo painel de especialistas (Sireci et al., 2006), e até mesmo oferecer ferramentas para que os especialistas emitam recomendações e possam tomar decisões informadas de aceite e rejeição de itens, a primeira parte desta tese é composta por uma ampla revisão de literatura que traz o histórico e o desenvolvimento do construto da IE como um conjunto de habilidades de raciocínio. O primeiro manuscrito traz informações que fomentam a justificativa de tomar esta perspectiva no estudo da IE, e, tendo em vista a escassez de testes de desempenho que sejam derivados da teoria Mayer–Salovey–Caruso (MSC; Mayer et al., 2012), justifica a importância de adaptar um novo teste para o contexto brasileiro. Em termos teóricos, essa justificativa é reforçada pelos argumentos que levaram MacCann e Roberts (2008) a construir o teste, que também se aplicam para o Brasil. Além disso, conforme as diversas críticas realizadas ao *Mayer–Salovey–Caruso Emotional Intelligence Test* (MSCEIT), é

necessária a utilização de uma teoria substantiva de emoção para subsidiar as evidências de validade com base no conteúdo do teste. Para construção dos itens de forma sistemática, MacCann e Roberts (2008) empregaram a teoria estrutural de emoções de Roseman (1984; 2001).

No primeiro manuscrito, também é feita a distinção entre o modelo de habilidades da IE e o modelo da IE como traço de personalidade. Tendo em vista que o estudo de validação dos testes adaptados emprega evidências de validade discriminante em relação à personalidade para construir a justificativa de validade, assim como o estudo de validação original dos testes (MacCann & Roberts, 2008), essa distinção justifica teoricamente esta mesma posição sendo tomada no segundo manuscrito.

Ainda no segundo manuscrito, os subsídios para a tomada de decisão no método de adaptação são estabelecidos na introdução, que também estabelece um esboço para as hipóteses que são descritas no método. Outra tarefa deste manuscrito é empregar o método padrão-ouro de análise de estrutura interna (anteriormente conhecida como validade de construto) para obter evidências de validade baseadas nos índices de ajuste, tanto tradicionais, como o NNFI (Bentler, 1990) e o RMSEA, quanto mais recentes, como os índices de ajuste global M^2* (Cai & Hansen, 2013) e C^2 (Cai & Monro, 2014), e os índices de ajuste por item $S\text{-}\chi^2$ (Kang & Chen, 2008) da teoria de resposta ao item (TRI).

As análises dos índices de ajuste do modelo de variável latente conferem mais confiabilidade aos resultados, além de uma análise simples dos coeficientes de dificuldade e discriminação da TRI (Cai & Hansen, 2013). Essas medidas obtiveram valores favoráveis e tornam a interpretabilidade dos resultados da estrutura dos construtos mais concreta, uma vez que não se admite interpretação de erro de mensuração além daquele expresso por estes índices de ajuste. Torna-se possível afirmar que a evidência de validade com base na estrutura interna dos testes é favorável.

Além disso, a evidência de validade com base na relação hipotetizada com os demais construtos mensurados é parcialmente confirmada no que diz respeito às relações conhecidas como validade discriminante (Campbell & Fisk, 1959). Em outras palavras, correlações não encontradas no estudo original dos testes situacionais de IE (MacCann & Roberts, 2008) entre os escores dos testes de compreensão emocional e de gerenciamento emocional e os fatores de personalidade também não foram encontradas no presente estudo. Com base no paradigma de validade convergente-discriminante, que é um tipo de evidência de validade com base na relação com variáveis externas (AERA et al., 2014), o fato de ser possível diferenciar um construto de outro construto diferente tem valor como evidência de validade. A ideia da validade convergente é o contrário, isto é, a correlação entre escores que devem ser relacionados também tem valor como evidência de validade. Neste estudo, a correlação significativa entre o escore do Teste Situacional de Compreensão Emocional (STEU) e do Teste Situacional de Gerenciamento Emocional (STEM), r = 0,501, sugere que eles realmente formam um fator comum: a inteligência emocional em si. Isto também foi observado no estudo de MacCann e Roberts (2008).

Uma forma notável pela qual o Manuscrito 2 divergiu dos resultados do estudo de MacCann e Roberts (2008) foi pela ausência de correlação entre o STEM e a Satisfação com a Vida. Ao mesmo tempo, houve correlação baixa entre Satisfação com a Vida e o STEU que não foi encontrada no estudo original (MacCann & Roberts, 2008). No entanto, a magnitude muito baixa dessas correlações neste estudo e no estudo original, torna provável que as correlações não sejam reais, ou pelo menos não tenham efeito prático. Subscrever a esta hipótese não entra em conflito com estudos descritos no primeiro manuscrito que encontram correlação de magnitude média entre inteligência emocional e satisfação com a vida, tendo em vista que esta correlação era com a inteligência emocional mensurada por escalas de autorrelato que avaliavam a IE pelo modelo de traço de personalidade (Sánchez-Álvarez et al., 2016)

Tendo isto em vista, os resultados são amplamente favoráveis à validade dos testes. É nesse contexto que os manuscritos 3 e 4 passam para a última tarefa proposta para a tese: elucidar a possibilidade de administração eficiente do STEU e do STEM como teste adaptativo computadorizado (CAT) sem prejuízo para validade e precisão. No Manuscrito 3, este argumento é construído ao demonstrar as vantagens associadas à testagem adaptativa e as estratégias que servem para mitigar o efeito negativo que as desvantagens da testagem adaptativa podem causar. Neste contexto, o argumento é reforçado ao estabelecer que a testagem adaptativa possibilita, de várias maneiras, um aumento na validade da testagem (Wise, 2018).

Além disso, a construção de uma justificativa do emprego do CAT nos testes adaptados depende da capacidade de uma aplicação que utiliza os algoritmos de seleção dos itens (ISAs) adaptativos resultar em estimativas de IE equivalentes às de uma aplicação dos testes em formato tradicional. Isto seria adequadamente justificado se essas estimativas equivalentes se beneficiassem das vantagens da testagem adaptativa. Em outras palavras, a ideia é que a administração do teste como CAT resulte em escores com a mesma validade e precisão, mas faça uso das vantagens de aplicação do CAT, como por meio de uma testagem mais eficiente. Dois estudos são descritos Manuscrito 4 que tiveram o objetivo de construir um argumento que mostre exatamente isso.

As simulações do primeiro estudo tiveram como regra de parada o critério padrão de diferença de erro padrão menor que 10^{-3} ($\Delta SE < 10^{-3}$). Os resultados deste estudo revelaram que, mesmo aplicando menos itens em média, todos os algoritmos—incluindo o que selecionou itens aleatoriamente—produziram escores que se correlacionam fortemente ($r \geq 0{,}9$) com os escores obtidos na aplicação do teste completo. O desempenho dos ISAs foi semelhante no STEU e no STEM. Uma análise mais detalhada revelou que os ISAs baseados na informação de Fisher, como o máxima informação de Fisher (MFI; Lord, 1980), mínima variância posterior

esperada (MEPV; Bock & Mislevy, 1984), máxima informação ponderada pela verossimilhança (MLWI) e máxima informação ponderada pela posterior (MPWI; van der Linden, 1993) e máxima informação esperada (MEI; Han, 2018) tiveram os melhores desempenhos, seguido pelos ISAs baseados na informação de Kullback–Leibler (KL), que foram representados pelas diferentes variantes do critério de integração de Kullback-Leibler (IKL; Cover & Thomas, 1991). O algoritmo menos eficiente foi o algoritmo aleatório. Novamente, estes resultados foram iguais nos dois testes.

No segundo estudo, uma segunda etapa de simulações foi realizada na qual cada ISA realizou *k* simulações por teste, sendo *k* o número de itens do teste. A regra de parada de cada simulação foi definida como o número de itens que deveria ser aplicado, de 1 a *k*, sendo k o número total de itens do teste que estava sendo simulado (32 para o STEU, e 30 para o STEM). Desta forma, além de identificar que os ISAs alcançaram o erro padrão usado como critério com um determinado número de itens um determinado número de vezes—conforme descrito acima—seria possível calcular a média de erro padrão para cada ISA para cada número de itens respondidos.

Esta análise demonstrou mais claramente o desempenho dos ISAs. Foi possível concluir que, utilizando o algoritmo de melhor desempenho, seria possível, em média, interromper o STEU 15 itens (46,9%) antes de administrar todos os itens, e o STEM 6 itens (20%) antes, com um prejuízo de apenas 0,01 no erro padrão. Apesar de não ser, em todas as situações, insignificativa, este prejuízo no erro padrão é pequeno o bastante que há subsídios para ser ignorada, isto é, para definir a regra de parada como 0,01, e não 0,001 (Chalmers, 2012; Finch, 2010). No entanto, não é necessário reduzir esta tolerância, que é o valor comumente considerado adequado para a regra de parada dos CATs (Chalmers, 2012). Ao invés disso, esta análise teve como objetivo apenas informar acerca dos efeitos médios do emprego dos ISAs.

Em termos práticos, a primeira etapa da análise mostrou que, com o emprego do melhor ISA, mais da metade (368 ou 53,5%) dos participantes responderiam pelo menos um item a menos no STEU. Além disso, seria possível reduzir a quantidade média de itens administrados no STEU em 5 (15,6%), sem prejuízo para o erro padrão. Apesar de não haver uma redução média de itens aplicados mantendo o erro padrão em todos os sujeitos, a análise revelou que pelo menos 187 (27,18%) dos participantes simulados teriam terminado o teste tendo respondido no mínimo um item a menos, e no máximo 26 itens a menos.

Por outro lado, os resultados também evidenciaram uma série de limitações. Por exemplo, a quantidade reduzida de instrumentos de IE habilidade disponíveis para pesquisa no Brasil limitou a capacidade de obtenção de evidências de validade com base na relação com variáveis externas do tipo validade convergente com outras medidas de gerenciamento emocional e de compreensão emocional. Além disso, idealmente seriam desenvolvidas medidas dos quatro "ramos" da teoria MSC. No entanto, esta limitação é combatida pela combinação das fortes evidências de validade com base na estrutura do teste e com base no conteúdo do teste, que, dentro da lógica argumentativa de *validity by design* (Mislevy, 2007), dão subsídio para um bom julgamento de validade quando o teste está nesta exata situação: quando possui evidências de validade derivadas do processo de construção e/ou adaptação mas carece de um teste com o qual aplicar para coleta de evidências de validade convergente. Mislevy (2007) defende a possibilidade de julgar a validade do teste com base na teoria por meio da avaliação da representação do construto.

Neste sentido, as evidências de validade são favoráveis. Por um lado, o processo de adaptação utilizou-se de um procedimento já tradicional: a retrotradução, inicialmente descrita por Brislin (1970) e posteriormente incrementadas por sugestões de Sireci et al. (2006) e da *International Test Commission* (2017). Ao mesmo tempo, o procedimento de construção dos testes situacionais de IE foi uma resposta às críticas a falta de sustentação teórica no

desenvolvimento dos itens do MSCEIT. Neste contexto, MacCann e Roberts (2008) utilizaram-se de uma fundamentação sólida: a teoria estrutural de emoções de Roseman (2001) e o paradigma do teste situacional (Motowidlo et al., 1990).

Futuros estudos podem complementar as evidências de validade dos testes por meio de técnicas de exame do processo de resposta, bem como de relação com variáveis externas, inclusive de desempenho escolar e ocupacional. Além da complementação da teoria MSC, o estudo dos processos de resposta dos testes situacionais pode ser utilizado para testar hipóteses da teoria de Roseman (2001) que ainda estão em estudo (Lerner et al., 2015).

Outra limitação relevante deste estudo diz respeito ao número de itens dos testes adaptados. Os testes avaliados não possuem tamanho reduzido, mas há possibilidade de construção de muito mais itens. Por exemplo, esforços como o Banco Internacional de Itens de Personalidade já juntaram mais de 3000 itens publicados juntamente com seus parâmetros de TRI (*Oregon Research Institute*, s.d.). Apesar de ter sido possível executar plenamente o objetivo do Manuscrito 4, seria possível realizar estudos para avaliar os benefícios do CAT de diferentes formas com um banco de itens extensivo. No entanto, a publicação dos itens juntamente com os parâmetros de TRI, descritos no Manuscrito 2, permitirá que pesquisadores empreguem o delineamento de blocos incompletos balanceados para calibrar novos itens de forma eficiente.

Em resumo, os resultados dos estudos realizados permitem afirmar que os testes STEU e STEM adaptados possuem evidências de validade majoritariamente positivas e podem ser administrados num formato de CAT, tomando parte nos benefícios da testagem adaptativa, sem prejuízo à precisão e validade dos escores. Foi possível mostrar, por exemplo, uma melhora substancial no tempo de resposta da maioria dos participantes.

Além das consequências diretas do estudo, o sucesso do emprego do CAT promove uma discussão dos métodos tradicionais de testagem sendo usados no Brasil, tanto na testagem

psicológica como na educacional. Por meio de replicação dos métodos de adaptação de um teste tradicional para um CAT descritos neste estudo, torna-se possível que diversos outros testes tomem parte nos extensos benefícios deste método de testagem. O presente estudo mostrou que mesmo a adaptação de uma medida tradicional é beneficiada pela metodologia do CAT, mantendo a validade da testagem a despeito da aplicação reduzida de itens. Espera-se que estudos possam apresentar possíveis benefícios a medidas já existentes de forma similar.

Na testagem educacional, por exemplo, o emprego de CAT teria um impacto positivo no acompanhamento da evolução dos alunos e na eficiência do trabalho docente (Luecht, 2016). Na testagem psicológica, o impacto da propagação dessa metodologia tem desdobramentos na área de saúde mental (Carlo et al., 2021), no recrutamento em seleção (Kantrowitz & Dawson, 2011), e na realização de pesquisa em termos gerais, inclusive na ética associada às exigências de tempo que se faz dos participantes. Desta forma, espera-se que os recursos disponibilizados por este estudo, incluindo o material suplementar que contém o código-fonte das análises realizadas, resultem numa maior adoção do CAT na psicometria e em outras áreas do conhecimento.

Referências

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*(2), 248–287. https://doi.org/10.1016/0749-5978(91)90022-L

Bar-On, R. (2000). Emotional and social intelligence: Insights from the Emotional Quotient Inventory. Em R. Bar-On & J. D. A. Parker (Orgs.), *The handbook of emotional intelligence: Theory, development, assessment, and application at home, school, and in the workplace* (pp. 363–388). Jossey-Bass/Wiley.

Bentler, P. M. (1990). Comparative fit indexes in structural model, *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431–444. https://doi.org./10.1177/0146621682006004

Brislin, R. W. (1970). Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216. https://doi.org/10.1177/135910457000100301

Bru-Luna, L. M., Martí-Vilar, M., Merino-Soto, C., & Cervera-Santiago, J. L. (2021). Emotional Intelligence Measures: A Systematic Review. *Healthcare, 9*(12), Artigo 1696. https://doi.org/10.3390/healthcare9121696

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*, 245–276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Cai, L., & Monro, S. (2014). *A new statistic for evaluating item response theory models for ordinal data* (ERIC Document № ED555724). ERIC Database.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Carlo, A. D., Barnett, B. S., & Cella, D. (2021) Computerized Adaptive Testing (CAT) and the Future of Measurement-Based Mental Health Care. *Administration and Policy in Mental Health and Mental Health Services Research*, 48, 729–731. https://doi.org/10.1007/s10488-021-01123-9

Chalmers, R., P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cancino-Montecinos, S., Björklund, F., & Lindholm, T. (2018). Dissonance reduction as emotion regulation: Attitude change is related to positive emotions in the induced compliance paradigm. *PLOS ONE, 13*(12), Artigo e0209012. https://doi.org/10.1371/journal.pone.0209012

Cancino-Montecinos, S., Björklund, F., & Lindholm, T. (2020). A General Model of Dissonance Reduction: Unifying Past Accounts via an Emotion Regulation Perspective. *Frontiers in Psychology, 11*, Artigo 540081. https://doi.org/10.3389/fpsyg.2020.540081

Chalmers, R. P. (2016). Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *Journal of Statistical Software*, 71(5), 1–38. https://doi.org/10.18637/jss.v071.i05

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Wiley.

Evans, T. R., Hughes, D. J., & Steptoe-Warren, G. (2019). A Conceptual Replication of Emotional Intelligence as a Second-Stratum Factor of Intelligence. *Emotion*, *20*(3), 507–512. https://doi.org/10.1037/emo0000569

Elfenbein, H. A., & MacCann, C. (2017). A closer look at ability emotional intelligence (EI): What are its component parts, and how do they relate to each other? *Social and Personality Psychology Compass, 11*(7), Artigo e12324. https://doi.org/10.1111/spc3.12324

Finch, H. (2010). Item Parameter Estimation for the MIRT Model: Bias and Precision of Confirmatory Factor Analysis–Based Models. *Applied Psychological Measurement, 34*(1), 10–26. https://doi.org/10.1177/0146621609336112

Gignac, G. E. (2005). Evaluating the MSCEIT V2.0 via CFA: Comment on Mayer et al. (2003). *Emotion, 5*(2), 233–235. https://doi.org/10.1037/1528-3542.5.2.233

Gonzaga, A. R., & Monteiro, J. K. (2011). Inteligência emocional no Brasil: um panorama da pesquisa científica. *Psicologia: Teoria e Pesquisa, 27*(2), 225–232. https://doi.org/10.1590/S0102-37722011000200013

Grubb III, W., & Mcdaniel, M. (2007). The Fakability of Bar-On's Emotional Quotient Inventory Short Form: Catch Me if You Can. *Human Performance, 20*, 43–59. https://doi.org/10.1080/08959280709336928

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255-282. https://doi.org/10.1007/BF02288892

Han, K. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal Of Educational Evaluation for Health Professions*, *15*(1), Artigo 7. https://doi.org/10.3352/jeehp.2018.15.7

Hogan, T. P. (2007). *Psychological Testing: A Practical Introduction* (2nd ed.). Wiley.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16–25. https://doi.org/10.1111/j.1745-3992.2001.tb00066.x

International Test Commission. (2017). *The ITC Guidelines for Translating and Adapting Tests (2nd edition)*. Recuperado de https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

Oregon Research Institute. (s.d.). *International Personality Item Pool. A Scientific Collaboratory for the Development of Advanced Measures of Personality and Other Individual Differences*. IPIP Home. https://ipip.ori.org/

Kang, T., & Chen, T. T. (2007). *An investigation of the performance of the generalized S-χ² item-fit index for polytomous IRT models*. ACT Research Report.

Kantrowitz T. M., Dawson C. R., Fetzer M. S. (2011). Computer adaptive testing (CAT): A faster, smarter, and more secure approach to pre-employment testing. *Journal of Business and Psychology, 26*, 227–232.

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology, 66*, 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203056615

Luecht, R., & Sireci, S. (2011). *A Review of Models for Computer-Based Testing* (ERIC Document № ED465763). ERIC Database.

Luecht, R. M. (2016). Computer-Adaptive Testing. *Wiley StatsRef*. https://doi.org/10.1002/9781118445112.stat06405.pub2

MacCann, C., &, Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: theory and data. *Emotion*, 8(4), 540–551. https://doi.org/10.1037/a0012746

Mayer, J. D., Caruso, D. R., & Salovey, P. (2000). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, *27*(4), 267–298. https://doi.org/10.1016/S0160-2896(99)00016-1

Mayer, J. D., Salovey, P., & Caruso, D. (2002). *Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT© V2.0) User's Manual*. MHS Publishers.

Mayer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology*, *59*, 507–536. https://doi.org/10.1146/annurev.psych.59.103006.093646

Miguel, F. K., & Primi, R. (2014). Estudo psicométrico do Teste Informatizado de Percepção de Emoções Primárias. *Avaliação Psicológica, 13*(1), 1–9.

Miguel, F. K., Giromini, L., Colombarolli, M. S., Zuanazzi, A. C., & Zennaro, A. (2016). A Brazilian Investigation of the 36- and 16-Item Difficulties in Emotion Regulation Scales. *Journal of Clinical Psychology, 73*(9), 1146–1159. https://doi.org/10.1002/jclp.22404

Mislevy, R. J. (2007). Validity by Design. *Educational Researcher, 36*(8), 463–469. https://doi.org/10.3102/0013189X07311660

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75*(6), 640–647. https://doi.org/10.1037/0021-9010.75.6.640

O'Connor, P. J., Hill, A., Kaya, M., & Martin, B. (2019). The Measurement of Emotional Intelligence: A Critical Review of the Literature and Recommendations for Researchers and Practitioners, *Frontiers in Psychology, 10*, Artigo 1116. https://doi.org/10.3389/fpsyg.2019.01116

Palmer, B. R., Gignac, G., Manocha, R., & Stough, C. (2005). A psychometric evaluation of the Mayer-Salovey-Caruso Emotional Intelligence Test Version 2.0. *Intelligence, 33*(3), 285–305. https://doi.org/10.1016/j.intell.2004.11.003

Petrides, K. V., & Furnham, A. (2001). Trait Emotional Intelligence: Psychometric Investigation with Reference to Established Trait Taxonomies. *European Journal of Personality, 15*, 425–448. https://doi.org/10.1002/per.416

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. In P. Shave (Ed.), *Review of Personality & Social Psychology* (Vol. 5, pp. 11–36). Sage Publications. https://www.researchgate.net/publication/245683603_Cognitive_Determinants_of_Emotion_A_Structural_Theory

Roseman, I. J. (2001). A model of appraisal in the emotion system: Integrating theory, research, and applications. In K. R. Scherer & A. Schorr (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 68–91). Oxford University Press.

Rossen, E., Kranzler, J. H., & Algina, J. (2008). Confirmatory factor analysis of the Mayer–Salovey–Caruso Emotional Intelligence Test V 2.0 (MSCEIT). *Personality and Individual Differences, 44*(5), 1258–1269. https://doi.org/10.1016/j.paid.2007.11.020

Santos, M. V., Nakano, T. C., & Silva, T. F. (2015). *Competências socioemocionais: análise da produção científica nacional e internacional* [Apresentação de Trabalho, Psicologia USP, São Paulo]. https://www.researchgate.net/publication/283777647

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. Em D. P. Flanagan, E. M. McDonough. (Orgs.). *Contemporary Intellectual Assessment: Theories, Tests and Issues* (4th ed.). Guildford Publications.

Sireci, S. G., Yang, Y., Harter, J., & Ehrlich, E. J. (2006). Evaluating Guidelines for Test Adaptations: A Methodological Analysis of Translation Quality. *Journal of Cross-Cultural Psychology, 37*(5), 557–567. https://doi.org/10.1177/0022022106290478

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201–216. https://doi.org/10.1007/BF02294775

Wise, S. L. (2018). Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating. *Education Inquiry, 10*(1), 1–13. https://doi.org/10.1080/20004508.2018.1490127

Woyciekoski, C., & Hutz, C. S. (2009). Inteligência emocional: teoria, pesquisa, medida, aplicações e controvérsias. *Psicologia: Reflexão e Crítica, 22*(1), 1–11. https://doi.org/10.1590/S0102-79722009000100002