

Teacher Behavior Checklist e Implicit Relational Assessment Procedure na avaliação de professores

Raquel Neves Balan¹, Verônica Bender Haydu², João Henrique de Almeida³,
Marcelo Henrique Oliveira Henklain⁴ y Marcela Roberta Jacyntho Zacarin⁵
*Universidade Estadual de Londrina- Brasil y Universidade Federal de
Roraima-Brasil*

A avaliação de professores tem sido conduzida com instrumentos como o Teacher Behavior Checklist (TBC) cuja evidência de validade de conteúdo foi obtida por relato verbal. A relação entre seis itens do TBC e o estímulo “Bom Professor” foi avaliada com o IRAP, e foi avaliada a correlação entre a nota dos participantes em uma disciplina da graduação e a avaliação do professor. Participaram 40 estudantes universitários que responderam aos dois instrumentos e informaram a nota recebida na disciplina. As médias dos D-IRAP escores foram estatisticamente significativas para “bom professor-positivo-verdadeiro” e “mau professor-negativo-verdadeiro”. Os índices de correlações entre as notas dos participantes na disciplina e a avaliação do docente não foram estatisticamente significativos.

Palavras-chave: Teacher Behavior Checklist, Implicit Relational Assessment Procedure, ensino eficaz.

¹ Graduada em Psicologia pela Universidade Estadual de Londrina e bolsista da iniciação científica pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Direção postal: Rua Gago Coutinho 34, Jardim Caravelle, Londrina, Paraná, Brasil. Contato: raquelineves0207@gmail.com.

² Mestrado e doutorado pela Universidade de São Paulo. Professora da Universidade Estadual de Londrina. Direção postal: Rua Duque de Caxias, 1235, Jardim Caviúna, Rolândia, Paraná, Brasil. Contato: veronichaydu@gmail.com. <https://orcid.org/0000-0002-4522-8338>

³ Doutor em Psicologia pela Universidade Federal de São Carlos. Professor na Universidade Estadual de Londrina. Direção Postal: Rua Duque de Caxias, 1235, Jardim Caviúna, Rolândia, Paraná, Brasil. Contato: joaodealmeida@uel.br

⁴ Doutor em Psicologia pela Universidade Federal de São Carlos. Professor do curso de Psicologia da Universidade Federal de Roraima, sendo colaborador externo do Instituto Nacional de Ciência e Tecnologia em Comportamento, Cognição e Ensino (INCT-ECCE). Direção postal: Rua do Cajueiro, 133, Caçari, Boa Vista, Roraima, Brasil. Contato: marcelo.henklain@ufrr.com.br ORCID: <https://orcid.org/0000-0001-9884-8592>

⁵ Mestrado e doutoranda no programa de pós-graduação em Análise do Comportamento da Universidade Estadual de Londrina. Psicóloga clínica na Psicomciência – psicoterapia e ciência comportamental contextual. Direção postal: Rua Guararapes 76, apartamento 404, Jardim Higienópolis, Londrina, Paraná, Brasil. Contato: psimzacarin@gmail.com.

Teacher Behavior Checklist e Implicit Relational Assessment Procedure en la evaluación docente

La evaluación de docentes se ha realizado con instrumentos como el Teacher Behavior Checklist (TBC) cuya evidencia de validez de contenido se ha obtenido mediante el reporte verbal. Se evaluó la relación entre seis ítems de TBC y el estímulo “Buen maestro” mediante el IRAP, y la correlación entre la puntuación de los participantes en un curso de pregrado y la evaluación del maestro. Participaron cuarenta estudiantes, quienes respondieron ambos instrumentos e informaron la calificación recibida en la disciplina. Los promedios de los puntajes del D-IRAP fueron estadísticamente significativos para “buen maestro-positivo-verdadero” y “mal maestro-negativo-verdadero”. Los índices de correlación entre las calificaciones de los participantes en la disciplina y la evaluación del maestro no fueron estadísticamente significativos.

Palabras clave: Teacher Behavior Checklist, Implicit Relational Assessment Procedure, enseñanza eficaz.

Teacher Behavior Checklist and Implicit Relational Assessment Procedure in teacher evaluation

The evaluation of teachers has been conducted with instruments such as the Teacher Behavior Checklist (TBC), whose evidence of content validity was obtained by verbal reporting. The relationship between six TBC items and the “Good Teacher” stimulus was assessed using the IRAP, and the correlation between the score of participants in an undergraduate course and the teacher’s assessment was assessed. Forty university students participated, who answered both instruments and reported the grade received in the discipline. The D-IRAP scores’ averages were statistically significant for “good teacher-positive-true” and “bad teacher-negative-true”. The correlation indexes between the participants’ grades in the discipline and the way they evaluated the teacher were not statistically significant.

Keywords: Teacher Behavior Checklist, Implicit Relational Assessment Procedure, effective teaching.

Teacher Behavior Checklist e Implicit Relational Assessment Procedure dans l'évaluation des enseignants

L'évaluation des enseignants a été menée à l'aide d'instruments tels que la liste de contrôle du comportement des enseignants (TBC) dont la preuve de la validité a été obtenue par le biais de rapports verbaux. La relation entre six items TBC et le stimulus “bon professeur” a été évaluée avec le IRAP, et la corrélation entre le score des participants dans une discipline et l'évaluation de l'enseignant. Quarante élèves ont participé qui ont répondu aux deux instruments et ont rapporté la note reçue dans la discipline. Les moyennes des scores D-IRAP étaient statistiquement significatives pour “bon enseignant-positif-vrai” et “mauvais enseignant-négatif-vrai”. Les indices de corrélation entre les notes des participants dans la discipline et l'évaluation du professeur n'étaient pas statistiquement significatifs.

Mots-clés: Teacher behavior checklist Implicit Relational Assessment Procedure, enseignement efficace.

A avaliação de comportamentos de docentes pode ocorrer de diferentes formas, dentre as quais existem a autoavaliação, avaliação pelos discentes ou avaliação externa. Para autores como Wilson e Ryan (2012), a avaliação realizada pelos alunos é benéfica porque há uma grande amostra de indivíduos com percepções variadas sobre o professor. Além disso, os estudantes entram em contato com diferentes momentos do ensino do docente avaliado. Esses dois benefícios podem diminuir a magnitude do erro na avaliação. Eles não estão presentes no caso de avaliadores externos porque normalmente uma única pessoa é designada a avaliar o desempenho do docente em um momento específico e não ao longo do tempo da disciplina. Outro benefício da avaliação feita pelos próprios alunos é que o avaliador não será uma pessoa totalmente desconhecida ou superior hierarquicamente, evitando influências como estresse ou nervosismo na performance do professor. Considerando esses benefícios, buscar meios para que essa avaliação realizada pelos alunos seja viabilizada de forma a considerar variáveis como desejabilidade social, maneiras de se mensurar o fenômeno e apresentar os dados de forma clara, torna-se algo relevante.

Um dos instrumentos que pode ser utilizado para avaliação de professores é o *Teacher Behavior Checklist* (TBC), desenvolvido por Buskist et al. (2002), com evidências psicométricas preliminares identificadas por Keeley, Smith e Buskist (2006). O TBC é formado por 28 itens que representam qualidades de professores eficazes e exemplos desses comportamentos. Existem diferentes maneiras pelas quais o TBC pode ser utilizado na avaliação de docentes, a depender dos objetivos do estudo. A primeira é avaliar quais itens do TBC são relevantes para definir um ensino eficaz (Buskist et al., 2002); a segunda é avaliar quão frequente são as características descritas nos itens do TBC em um professor que o respondente considera excelente (Schaeffer et al., 2003); e a terceira é avaliar o desempenho de um docente específico

pelas características descritas no TBC (Keeley et al., 2006). A partir dos objetivos da linha de pesquisa, o tipo de instrumento para a coleta de dados é especificado.

Os instrumentos geralmente utilizados para a aplicação do TBC são escalas Likert com cinco ou sete pontos. Por exemplo, Liu, Keeley e Buskist (2015) utilizaram o TBC com uma escala Likert de cinco pontos em um estudo com participantes chineses. O objetivo foi avaliar em que medida o TBC possui características que realmente representam professores eficazes. Para isso, os estudantes classificaram se os itens do instrumento eram apresentados por professores considerados eficazes, utilizando a escala de “1= nunca apresenta” a “5=sempre apresenta”. Os estudantes indicaram que todas as qualidades são importantes, atribuindo pontuação maior do que quatro para as características descritas nos itens do TBC. Outro estudo de validação (Lammers, Savina et al., 2010), teve como objetivo comparar em que medida participantes estadunidenses e russos consideravam as características do TBC importantes em docentes eficazes, por meio de uma classificação de “1 = de algum modo importante” a “7 = extremamente importante”. Os resultados permitiram observar que em 21 dos 28 itens do TBC não houve diferenças significativas entre os participantes, demonstrando evidência de validade de conteúdo do TBC em diferentes culturas.

Apesar de vários estudos (e.g., Keeley, Christopher & Buskist, 2012; Lammers et al., 2010; Buskist & Keeley, 2018) indicarem a validade do TBC como instrumento de avaliação de desempenho de professores, deve-se considerar que a medida desse questionário é o heterorrelato (avaliação dos discentes em relação aos docentes) e é uma medida explícita cujos dados podem estar sob controle da presença do experimentador ou do aplicador. Por exemplo, um estudante pode relatar o que é socialmente aceitável em detrimento de como ele se sente em relação a um determinado item de avaliação. Portanto, para além de utilizar o TBC com o intuito de examinar quais são as qualidades mais importantes em professores eficazes, é necessário que os estudos realizem procedimentos buscando minimizar variáveis sociais que podem estar controlando a resposta dos participantes.

Uma alternativa para avaliação de comportamentos evitando o viés social é possibilitada pelo *Implicit Relational Assessment Procedure* (IRAP) (Barnes-Holmes et al., 2006), um instrumento capaz de avaliar a força do responder relacional dos indivíduos por meio das latências de respostas. Durante a aplicação do IRAP, os participantes devem relacionar estímulos de maneira consistente e inconsistente com sua história pré-experimental. Isso possibilita a comparação das latências das respostas nos blocos consistentes e inconsistentes. A expectativa teórica é de que os participantes relacionem mais prontamente os estímulos que foram fortemente relacionados na história dos participantes (latências menores nos blocos consistentes), permitindo, assim, avaliar a força do responder relacional.

O IRAP pode ser utilizado para avaliar o responder relacional de uma variedade grande de comportamentos observados em diferentes contextos, como o educacional, o clínico, o social (cf. Golijani-Moghaddam, Hart, & Dawson, 2013; Mizael & de Almeida, 2019; Vahey, Nicholson, & Barnes-Holmes, 2015). No contexto educacional, Barnes-Holmes et al. (2006) avaliaram comportamentos de professores em relação à estudantes autistas. Os participantes desse estudo foram distribuídos em três grupos: grande experiência de trabalho com autistas, experiência moderada e nenhuma experiência. Todos eles responderam a dois questionários para avaliar suas atitudes em relação a essas crianças e responderam ao IRAP. Os estímulos do IRAP foram: (a) “Transtorno do Espectro Autista” ou “Desenvolvimento normal”; (b) adjetivos positivos ou negativos; e (c) “similar” e “oposto” como opções de resposta. Os professores deveriam responder tanto de forma consistente com viés negativo para autistas (autismo-negativo-similar, autismo-positivo-oposto, normal-negativo-oposto e normal-positivo-similar) quanto de forma inconsistente (autismo-positivo-similar, autismo-negativo-oposto, normal-positivo-oposto e normal-negativo-similar). Os resultados do questionário de autorrelato demonstraram que apenas os participantes do grupo sem experiência possuíam viés negativo para estudantes autistas. No entanto, no IRAP, todos os grupos demonstraram latências menores ao relacionar desenvolvi-

mento normal com coisas boas em comparação ao desenvolvimento atípico. Assim, pode-se entender que o IRAP permitiu avaliar um responder implícito que não havia sido identificado em um questionário de avaliação explícita.

Ainda em relação à utilização do IRAP no contexto educacional, Kelly e Barnes-Holmes (2013) realizaram um estudo para avaliar vieses positivos ou negativos sobre a utilização de reforço ou punição na presença de comportamentos “bons” ou “ruins”. Os participantes foram professores que trabalhavam com crianças com atraso no desenvolvimento e professores que trabalhavam com crianças com desenvolvimento típico. Todos os professores responderam à instrumentos explícitos de avaliação (*Behavior Intervention Rating Scale* e *Treatment Acceptability Rating Form—Revised*) e ao IRAP. No IRAP, os participantes deviam relacionar estímulos que representam comportamentos bons (e.g., estudar sentado) e comportamentos ruins (e.g., chutar objetos da sala de aula) às palavras “reforço” e “punição”, por meio da escolha entre “verdadeiro” ou “falso”. Em blocos consistentes, deveriam responder como se comportamentos bons fossem relacionados ao reforço e comportamentos ruins fossem relacionados à punição. Nos blocos inconsistentes, o contrário deveria ser feito. Os resultados do estudo demonstraram que nas medidas explícitas utilizadas, ambos os grupos demonstraram vieses antipunição tanto para comportamentos bons quanto para comportamentos ruins. No entanto, no IRAP, o grupo de professores de crianças com desenvolvimento típico demonstrou um viés pró-punição para comportamentos considerados ruins (latências pequenas ao selecionar verdadeiro diante da relação entre comportamentos ruins com a palavra punição). A partir do estudo, conclui-se que o IRAP, um instrumento implícito, foi eficaz para identificar vieses não identificados anteriormente pelos instrumentos explícitos utilizados (questionários).

Os estudos de Barnes-Holmes et al. (2016) e Kelly e Barnes-Holmes (2013) foram realizados comparando percepções de professores no contexto da educação especial. Docentes avaliaram alunos com características específicas ou processos de ensino presentes em sala de aula. No entanto, o IRAP também pode ser utilizado para os alunos avaliarem

docentes e os processos de ensino usado pelos professores durante as aulas. Henklain et al. (2019), por exemplo, avaliaram em que medida o responder relacional de 47 participantes no IRAP seria compatível com estudos prévios de validade de conteúdo do TBC que indicam que as 28 qualidades do instrumento são importantes para avaliar docentes. Na Etapa 1, os participantes responderam a seis categorias do TBC (acessibilidade, comunicação efetiva, entusiasmo, conhecimento, preparação e esforço para ser melhor) de acordo com uma escala Likert de 7 pontos (“1= totalmente irrelevante” a “7 = totalmente relevante”). Na Etapa 2, os estudantes responderam ao IRAP. Os estímulos apresentados no IRAP foram: (a) Bom Professor ou Mau Professor (estímulos-alvo); (b) características positivas (acessível, comunicador eficaz, entusiasmado, domina o conteúdo, preparado e busca aperfeiçoar-se) ou características negativas (inacessível, comunicador ineficaz, desanimado, desconhece o conteúdo, despreparado e acomodado) como estímulos-rótulo e (c) as opções de resposta Verdadeiro ou Falso. Nos blocos consistentes, os participantes deveriam responder como se bom professor fosse positivo e mau professor fosse negativo. Nos blocos inconsistentes, o contrário deveria ser feito. As latências das respostas foram transformadas em quatro tipos de D-IRAP escores: bom-professor-positivo, bom-professor-negativo, mau-professor-negativo e mau-professor-positivo. Verificou-se que os D-IRAP escores foram maiores para a relação entre bom-professor-positivo e mau-professor-negativo (latências menores ao responder diante dessas relações), o que permite sugerir que são relações mais bem estabelecidas em suas histórias. Portanto, os resultados do IRAP evidenciaram que a relação entre Bom Professor e as características descritas no TBC apresentado aos participantes era verdadeira.

Uma das sugestões de Henklain et al. (2019) foi de que estudos subsequentes avaliassem o efeito de ordem de aplicação dos instrumentos sobre os resultados, uma vez que se levantou a hipótese de que essa fosse uma variável relevante. Assim, para adicionar evidências aos resultados do estudo de Henklain et al., foi realizada uma replicação, utilizando os mesmos instrumentos com os mesmos estímulos. Os objetivos do presente estudo foram: (a) avaliar o efeito da ordem de

realização dos instrumentos (um grupo respondeu o TBC primeiro e o outro grupo que respondeu o IRAP primeiro) sobre os resultados dos dois instrumentos, (b) investigar se as características descritas nos seis itens do TBC são relacionadas a “Bom professor” no IRAP, (c) verificar se há correlação entre os resultados do TBC e do IRAP para os tipos de tentativa bom professor-positivo e mau professor-negativo, (d) verificar se há correlação entre as notas dos participantes nas matérias e a maneira como avaliaram o docente no TBC.

Método

Participantes

Participaram do estudo 40 estudantes universitários com idades entre 18 a 40 anos do curso de Psicologia. Eles foram distribuídos em dois grupos com 20 participantes para controle do efeito de ordem de execução do IRAP e o TBC. Os participantes foram distribuídos de forma randomizada de acordo com a ordem em que se candidataram a participar (o/a primeiro/primeira a se candidatar foi para o Grupo 1, o/a segundo/segunda a se candidatar para o Grupo 2 e assim por diante). Cada grupo ficou com 20 participantes em cada um.

Local e Materiais

O estudo foi realizado em um laboratório contendo cubículos nos quais havia uma mesa e uma cadeira. Os materiais utilizados foram o Termo de Consentimento Livre e Esclarecido (TCLE), uma adaptação do *Teacher Behavior Checklist* realizada para o presente estudo e o *software* GO IRAP (disponível em <https://go-rft.com/go-irap/>) instalado em um *notebook*.

Teacher Behavior Checklist adaptado para o estudo: foram utilizados seis itens avaliativos do TBC com uma Escala Likert de 1 a 5, em que as pontuações variam entre Totalmente Falso, Parcialmente Falso, Nem falso nem verdadeiro, Parcialmente verdadeiro e Totalmente verdadeiro,

respectivamente. Os seis itens selecionados foram os mesmos do estudo de Henklain (2019, p. 4): “(1) Acessível/disponível (Informa horário de trabalho; disponibiliza horário para atender os estudantes; disponibiliza seu contato telefônico, de WhatsApp e de e-mail; responde ao contato dos estudantes); (2) Comunicador(a) eficaz (Fala claramente e de modo que todos consigam ouvir e compreender; utiliza a língua portuguesa corretamente; fornece exemplos claros e convincentes); (3) Entusiasmado(a) pelo ensino e pelo tema que ensina (Sorri durante a aula; prepara atividades de sala de aula interessantes; utiliza gestos e expressa emoções para enfatizar pontos importantes; não se atrasa para a aula); (4) Domina o tema ensinado (Responde às perguntas dos estudantes com facilidade; ao ministrar aulas, não se limita a ler diretamente de livros ou anotações; utiliza exemplos claros e compreensíveis); (5) Preparado(a) (Traz os materiais necessários para a aula; dificilmente se atrasa para a aula; fornece diretrizes para as discussões em classe) e (6) Busca ser um(a) professor(a) melhor (Solicita avaliações dos estudantes sobre suas habilidades como professor(a); busca aprendizado/aperfeiçoamento contínuo [participa de seminários, etc. sobre educação]; utiliza novos métodos de ensino).” As instruções do instrumento foram:

“O Teacher Behavior Checklist (TBC) possui seis itens. Cada um deles representa uma qualidade docente, que está acompanhada por exemplos concretos de comportamentos do(a) professor(a) localizados ao lado do nome de cada qualidade. Esses exemplos deverão ajudá-lo(a) a compreender o significado de cada uma dessas qualidades.

A sua tarefa será: avalie o quão verdadeira ou falsa a característica especificada nesse item é evidenciada no ensino de um professor que ministra uma disciplina específica que você está cursando na graduação. Selecione a sua resposta na caixa de opções ao lado de cada item. A escala que você utilizará possui os seguintes graus de: 1 = Totalmente falso, 2 = Parcialmente falso, 3 = Nem falso, nem verdadeiro, 4 = Parcialmente verdadeiro e 5 = Totalmente verdadeiro.”

Software IRAP de pré-treino: A configuração desse instrumento foi realizada com a utilização de palavras comumente consideradas boas

ou ruins na comunidade verbal dos participantes. A tela do IRAP é configurada por um estímulo no topo da tela (estímulo-rótulo), um no centro da tela (estímulo-alvo) e duas opções de respostas no inferior da tela (“verdadeiro” ou “falso”). Os estímulos-rótulo desse IRAP foram “prazeroso” e “desprazeroso”, os estímulos-alvo foram atributos considerados positivos ou negativos no meio cultural dos participantes (sorriso, risada, dor de cabeça, vomito) e as opções de resposta foram “verdadeiro” ou “falso”. Esse instrumento foi composto por dois pares de blocos consistentes (atributos positivos-prazeroso-verdadeiro, atributos negativos-prazeroso-falso, atributos negativos-desprazeroso-verdadeiro e atributos positivos-desprazeroso-falso) e inconsistentes (atributos positivos-prazeroso-falso, atributos negativos-prazeroso-verdadeiro, atributos negativos-desprazeroso-falso e atributos positivos-desprezoso-verdadeiro), totalizando oito blocos sem exigência de critérios de latência ou precisão. Os estímulos foram retirados do estudo de Vahey, Boles e Barnes-Holmes (2010).

Software IRAP de avaliação dos itens que representam características de professores: instrumento computadorizado com a programação de tentativas de treino e teste de relações condicionais entre estímulos. A precisão e latências das respostas dos participantes foram registradas e armazenadas pelo *software*. O IRAP, em cada tentativa, apresenta um estímulo no topo da tela (estímulo-rótulo), um estímulo no centro da tela (estímulo-alvo) e duas opções de resposta na parte inferior. Os estímulos ficam na tela até que o participante aperte as teclas “D” e “K”, que correspondem a “Verdadeiro” ou “Falso” em posições aleatorizadas ao longo do experimento. A programação específica que blocos de prática são apresentados até que: o participante atinja um critério de acerto igual ou maior que 85% e latências de respostas igual ou menor que 2000 ms. Os blocos de teste subsequentes são compostos com exatamente os mesmos estímulos que os blocos de prática. Tanto na fase de prática quanto de teste, os blocos são alternados entre consistentes (professor bom-positivo-verdadeiro, professor bom-negativo-falso, mau professor-negativo-verdadeiro e mau professor-positivo-falso) e inconsistentes (professor bom-positivo-falso, professor bom-negativo-verdadeiro,

mau professor-negativo-falso e mau professor-positivo-verdadeiro). Um exemplo de como o IRAP pode ser programado encontra-se na Figura 1.

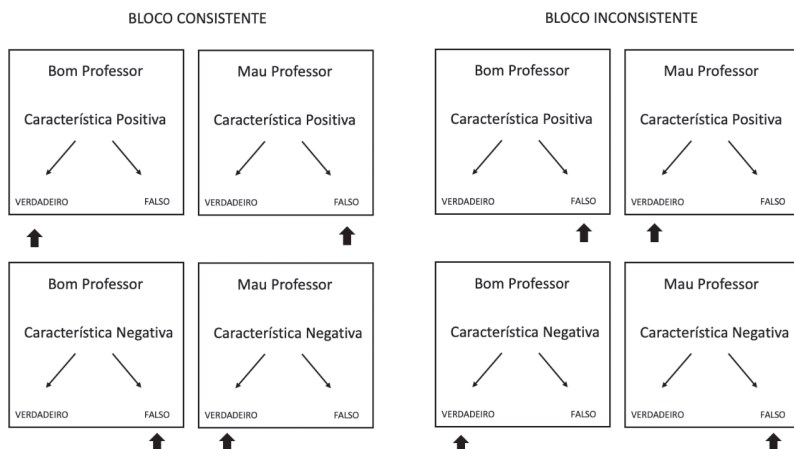


Figura 1. Exemplo do arranjo de estímulos no IRAP em blocos consistentes e inconsistentes. As setas indicam as respostas consideradas corretas nos blocos consistente (esquerda) e inconsistente (direita).

No IRAP do presente estudo, os estímulos-rótulo foram “Bom Professor” e “Professor Ruim”, os estímulos-alvo foram características positivas (Acessível, Comunicador Eficaz, Entusiasmado, Domina o Conteúdo, Preparado e Busca Aperfeiçoar-se) ou características negativas (Inacessível, Comunicador Ineficaz, Desanimado, Desconhece o Conteúdo, Despreparado e Acomodado). As opções de respostas foram “Verdadeiro” ou “Falso”. A Tabela 1 estão especificados os estímulos utilizados no IRAP.

Tabela 1

Estímulos que foram apresentados no IRAP

Estímulos-Modelo	
Bom professor	Mau professor
Estímulos-alvo	
Acessível	Inacessível
Comunicador eficaz	Comunicador ineficaz
Entusiasmado	Desanimado
Domina o conteúdo	Desconhece o conteúdo
Preparado	Despreparado
Busca aperfeiçoar-se	Acomodado

Procedimento

Os participantes do estudo foram convidados pela primeira autora deste estudo, tendo sido informados que sua participação ocorreria em dois momentos: durante uma disciplina ministrada por um professor específico de seu curso e após o término dessa disciplina. Inicialmente, os participantes receberam o Termo de Consentimento Livre Esclarecido (TCLE), aprovado pelo Comitê de Ética em Pesquisa da instituição da primeira autora para leitura e assinatura caso concordassem em participar. Na Etapa 1, os participantes do Grupo 1 receberam a versão adaptada do TBC para preenchimento e, em seguida, foi realizada a aplicação do IRAP de pré-treino, seguida pela aplicação do IRAP de avaliação do professor. Na Etapa 2, responderam novamente o TBC e foi solicitada a informação da nota recebida na disciplina do professor avaliado. Os participantes do Grupo 2, na Etapa 1, responderam primeiramente o IRAP de pré-treino e o IRAP de avaliação de professores, e depois preencheram o TBC adaptado. Na Etapa 2, após o término da disciplina, foi solicitado que respondessem novamente o TBC, e foi perguntado ao participante a nota recebida na disciplina desse professor.

Os participantes dos dois grupos, antes da realização do TBC, foram instruídos a responder pensando em um professor de uma disciplina específica que estavam cursando. Na segunda vez que os participantes responderam esse instrumento, foi solicitado que eles pensassem no mesmo professor e que informasse a nota final recebida nessa disciplina. No TBC, as respostas deveriam ser apresentadas considerando o quão verdadeira ou falsa é a relação entre a característica especificada no item e o repertório do professor avaliado.

Em relação ao IRAP de treino, os participantes responderam a dois pares de blocos consistentes e inconsistentes de maneira alternada. Caso os participantes respondessem de maneira incorreta, um “X” na tela era apresentado. A próxima tentativa só aparecia se a resposta correta fosse selecionada. Caso os participantes emitissem a resposta especificada pelos experimentadores (considerada a resposta correta), a próxima tela seria apresentada após 400 ms. No IRAP de avaliação de professores, os participantes passaram por no máximo oito pares de blocos de prática até atingirem os critérios programados. Se os critérios não fossem atingidos, a pesquisadora sinalizada que o estudo seria encerrado nesta etapa, e agradecia a colaboração do participante. Os dados dos blocos de prática realizados não foram considerados na análise dos dados. Caso o participante atingisse esses critérios, responderia a três pares de blocos de teste com os mesmos critérios de acertos e tempo de reação. Cada bloco foi composto por 24 tentativas com *feedback* para as respostas incorretas, após as quais apareceria o “X” vermelho na tela. A tentativa seguinte era apresentada após um intervalo de 400 ms.

Resultados

Os registros de latência obtidos foram convertidos em quatro pontuações médias de D-IRAP escores, uma para os quatro tipos de tentativas apresentadas no programa: bom professor-positivo, mau professor-positivo, bom professor-negativo, mau professor-negativo. Portanto, quatro D-IRAP escores para cada participante. A amostra

apresentou distribuição normal ($w=.995$, $p=.852$) e homogeneidade da variância ($F(3.156)=.989$, $p=.258$) verificada pelo teste de Levene. Para verificar se a ordem de aplicação do IRAP e do TBC afetou os D-IRAP escores e os resultados do TBC, foram feitas duas análises em separado para cada instrumento. Os dados do TBC não aprearam distribuição normal ($p=.021$), assim, foi feita uma análise por meio do teste de Welsh, tendo-se observado que não houve diferença estatisticamente significativa entre os dados dos dois grupos ($t = 1.80$; $p = .08$), isso é, não houve efeito de ordem. Quanto aos dados do IRAP, verificou-se que os escores D-IRAP apresentaram distribuição normal ($p=.20$), igualdade de variâncias do erro (Teste de Levene) e que se pode assumir que os dados apresentam esfericidade (Teste de Mauchly). Assim, considerando a interação entre os grupos e as relações testadas (bom professor-positivo, mau professor-positivo, mau professor-negativo, bom professor-negativo), verificou-se que não há diferença estatística significativa ($p = .33$), a partir do Modelo Linear Geral com medidas repetidas, podendo-se afirmar que não houve efeito da ordem de aplicação dos instrumentos.

Para verificar se houve relação estatisticamente significativa entre os seis itens do TBC com “Bom professor” no IRAP, e entre o antônimo desses seis itens com “Mau professor” no IRAP, realizou-se um Teste t. Constatou-se que no caso do grupo que respondeu o TBC primeiro, os dados são estatisticamente significativos em relação às tentativas bom professor-positivo-verdadeiro ($t=5.83$; $p=.0001$), mau professor-positivo-falso ($t=2.75$; $p=.0127$) e mau professor-negativo-verdadeiro ($t=3.80$; $p=.0012$). Para o grupo que respondeu o IRAP primeiro, os resultados foram estatisticamente significativos para os tipos de tentativas bom professor-positivo-verdadeiro ($t=4.36$; $p=.0003$) e mau professor-negativo-verdadeiro ($t=2.55$; $p=.0196$). Ao realizar-se o Teste t dos dados dos dois grupos de participantes para avaliar se os resultados são estatisticamente diferentes de zero, os tipos de tentativa estatisticamente significativos foram bom professor-positivo-verdadeiro ($t=7.13$; $p<.001$), bom professor-negativo-verdadeiro ($t=2.930$; $p=.006$) e mau

professor-negativo-verdadeiro ($t=4.397$; $p < .001$). Esses resultados são apresentados na Figura 2.

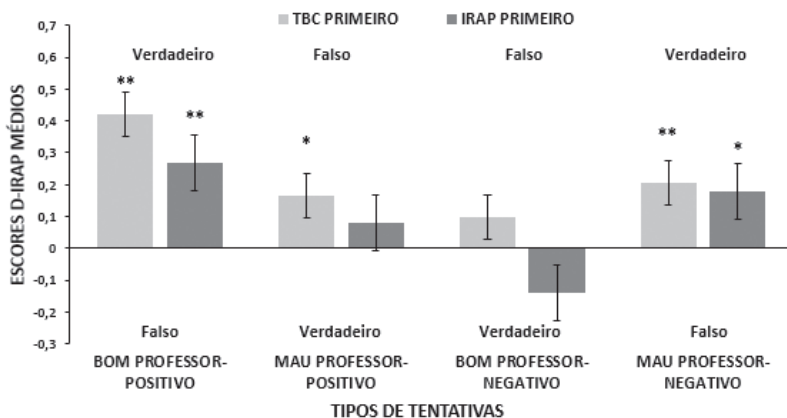


Figura 2. Distribuição dos escores D-IRAP médios para os quatro tipos de tentativa (bom professor-positivo, mau professor-positivo, bom professor-negativo, mau professor-negativo).
 $p < .05 = *$; $p < .01 = **$

A ANOVA com medidas repetidas e o cálculo para tamanho do efeito demonstrou que há diferenças no tamanho do efeito comparando os quatro tipos de tentativas do IRAP em cada um dos dois grupos. Os resultados são apresentados na Tabela 2.

Tabela 2

Resultados da ANOVA para os dois grupos de participantes, com os valores dos graus de liberdade (gl), o valor p e o tamanho do efeito (h^2 parcial)

	<i>F</i>	<i>Gl</i>	<i>p</i>	<i>h² parcial</i>
TBC primeiro	4.54	3	.006**	.15
IRAP primeiro	6.80	3	<.0001**	.21
Dois grupos	10.1	3	<.001**	.16

$p < .01 = **$

Foram realizados testes *post hoc* Bonferroni para verificar as diferenças entre os tipos de tentativas para os participantes que responderam o TBC primeiro, para os que responderam o IRAP primeiro e para os dois grupos juntos. Os resultados indicaram que o tipo de tentativa bom professor-positivo diferiu do tipo de tentativa mau professor-positivo ($p=.047$) com uma diferença média de .2425, e que o tipo de tentativa bom professor-positivo diferiu do e do tipo bom professor-negativo ($p=.005$) com uma diferença média de .3227. Para o grupo de participantes que responderam o IRAP primeiro, os resultados também foram estatisticamente significativos, $F(3.76)=6.80$, $p<.0001$, $h^2_{\text{partial}} = .212$, um tamanho de efeito moderado. O *post hoc* Bonferroni indicou que o tipo de tentativa bom professor-positivo diferiu do tipo de tentativa bom professor-negativo ($p= <.001$) com uma diferença média de .4062, e que o tipo de tentativa bom professor-negativo diferiu do tipo de tentativa mau professor-negativo ($p=.008$) com uma diferença média de -.3154. O *post hoc* Bonferroni indicou que o tipo de tentativa bom professor-positivo diferiu do tipo de tentativa mau professor-positivo ($p= .008$) com uma diferença média de .2277; que o tipo de tentativa bom professor-positivo diferiu do tipo de tentativa bom professor-negativo ($p= <0,001$) com uma diferença média de 03644; e que o tipo de tentativa bom professor-negativo diferiu do tipo de tentativa mau professor-negativo ($p=.012$) com uma diferença média de -.2123.

Conforme descrito no procedimento, os participantes avaliaram um bom professor ou um mau professor de acordo com a média das pontuações no TBC para verificar se a nota recebida pelo estudante na matéria afeta a maneira como esse estudante avaliou o professor no TBC. O índice de correlação de *Pearson* entre a média dos D-IRAP escores do tipo de tentativa bom professor-positivo com os escores do TBC para os participantes que avaliaram um bom professor no TBC não foi estatisticamente significativo ($r = .218$, $p= .343$) e para os que avaliaram um mau professor também não foi estatisticamente significativo ($r = -.155$, $p = .527$). Os índices de correlação entre a nota dos participantes que avaliaram um bom professor na disciplina e o escore com o qual avaliaram o docente no TBC ($r = .165$, $p = .489$), e a nota

dos participantes que avaliaram um mau professor na disciplina e o escore com o qual avaliaram esse professor no TBC ($r = -.009$, $p = .971$) não foram estatisticamente significativos.

Discussão

Os objetivos do presente estudo foram replicar o estudo de Henklain et al. (2019) e avaliar se a ordem de aplicação dos instrumentos (IRAP e TBC) afeta os resultados. Além disso, avaliou-se se as características do TBC estariam fortemente relacionadas à “Bom professor” no IRAP, e se há correlação entre a nota recebida pelos participantes em uma disciplina e a pontuação atribuída ao professor no TBC. Com relação ao primeiro objetivo pode-se afirmar que não houve diferença estatisticamente significativa entre os dois grupos, demonstrando que não houve efeito de ordem da aplicação do TBC e do IRAP. Esse resultado corrobora a avaliação realizada por Farrel e McHugh (2017), em que os pesquisadores aplicaram dois instrumentos de avaliação implícita (o IRAP e o IAT) com o intuito de identificar vieses de gênero em profissões consideradas masculinas. Ao analisar o efeito de ordem de aplicação dos instrumentos, não foram encontradas diferenças estatisticamente significativas entre os grupos. Outro estudo que avaliou o efeito de ordem de aplicação foi realizado por Keeley, Furr e Buskist (2010) com o TBC em que estudantes avaliaram um mau professor, um bom professor e o professor mais recente. Os pesquisadores controlaram a ordem de aplicação do TBC para a avaliação das diferentes categorias (bom professor, mau professor e professor recente). Os resultados demonstraram que no caso dos estudantes da universidade de Auburn, no Alabama (Grupo 1) não houve efeito de ordem. No caso dos estudantes da universidade de Appalachian, na Carolina do Norte (Grupo 2) houve efeito de ordem, mas o tamanho do efeito foi pequeno. Assim, sugere-se que ao comparar instrumentos de avaliação implícita e explícita a ordem de aplicação provavelmente não é relevante.

Com relação às características do TBC estarem relacionadas às do IRAP verificou-se que as características positivas descritas no TBC

foram fortemente relacionadas com “Bom professor” e as características negativas (antônimos dos adjetivos descritos no TBC) foram fortemente relacionadas com “Mau professor” para ambos os grupos. Para o grupo que respondeu o TBC primeiro, houve uma forte relação entre mau professor-positivo-falso. Esses resultados em relação ao IRAP são similares aos do estudo de Henklain et al. (2019), em que as médias dos D-IRAP escores foram estatisticamente significativas para os tipos de tentativas bom professor-positivo-verdadeiro e mau professor-negativo-verdadeiro. O IRAP do presente estudo foi programado com os mesmos seis estímulos do TBC de Henklain et al., o que indica uma replicação da evidência da validade de conteúdo desses seis itens em condições que minimizam a influência de variáveis sociais. Ao comparar os dados dos tipos de tentativas com valores estatisticamente significativos, verifica-se que, em ambos os estudos, as médias dos D-IRAP escores foram maiores para o tipo de tentativa bom professor-positivo em comparação ao tipo de tentativa mau professor-negativo. Isso indica que as qualidades positivas de bons professores são mais fortemente relacionadas do que a qualidades negativas de maus professores. Finn, Barnes-Holmes e McEnteggart (2018) apoiam essa hipótese ao afirmarem que a história pré-experimental não é o único fator responsável pelo efeito sobre o D-IRAP escore, segundo eles, a alta coerência entre os estímulos também influencia as respostas dos participantes. Com relação aos dados do presente estudo, sugere-se que a relação entre “Bom professor”, “qualidades positivas” e “verdadeiro”, é uma valência positiva (há alta coerência entre os estímulos) na língua portuguesa. Em contrapartida, isso não ocorre na relação entre os estímulos “Mau professor”, “qualidades negativas” e “verdadeiro”.

Em relação ao objetivo de verificar se houve correlações estatisticamente significativas entre as notas recebidas pelos estudantes e a maneira como avaliaram o docente, verificou-se que o índice de correlação não foi significativo. Isso pode ter sido observado porque os fatores que fazem um professor ser considerado bom ou ruim não estão fortemente relacionados às notas que atribuem aos alunos. Esses resultados corroboram com os do estudo realizado por Spooren e Mortelmans

(2006) em que foi avaliado se a nota recebida em um curso influenciou a maneira como os professores foram avaliados pelos alunos em um instrumento de 31 itens. Spooren e Mortelmans verificaram que notas maiores no curso estão fortemente relacionadas a uma melhor avaliação da disciplina, mas não necessariamente à avaliação do professor, indicando que outros fatores além da nota influenciam na avaliação do docente. No presente estudo, observou-se, por meio de relatos não sistematicamente registrados de alguns participantes, que determinadas qualidades consideradas por eles como importantes em um bom professor não estavam contempladas nos seis itens do TBC. Eles relataram que aspectos como empatia do docente em relação aos alunos é uma variável muito mais determinante para considerar um bom professor do que as notas atribuídas por ele. Essa informação é condizente com os resultados do estudo de Shevlin et al. (2000), que demonstraram que a característica de carisma de um professor influencia fortemente a maneira como a efetividade do ensino é avaliada. Logo, pode-se considerar que no caso do presente estudo, os professores que atribuíram notas ruins aos seus alunos podem ter sido avaliados de maneira positiva no TBC devido às outras características do repertório desse docente.

Outro fator a ser destacado em relação à valorização de outras características de docentes que poderiam não estar contempladas no TBC é que todos os participantes do presente estudo foram estudantes de Psicologia. De acordo com Jensen e Fischer (2006), cujo estudo teve como objetivo avaliar as percepções de alunos sobre seus professores em dois cursos diferentes, a preferência por características específicas em docentes é influenciada pelo curso que os alunos estão realizando. Por exemplo, estudantes de psicologia, no estudo Jensen e Fischer, valorizaram mais habilidades interpessoais dos docentes, como ser engraçado, tratar alunos com respeito e interagir com alunos de uma maneira positiva. Enquanto estudantes de administração valorizaram mais características dos docentes como oferecer informações atuais e *feedbacks* construtivos. Assim, uma possível razão adicional pela ausência de correlação entre notas recebidas e a maneira como o professor foi avaliado no TBC pode ter sido a valorização das habilidades interpessoais dos docentes em detrimento da nota que recebem.

Outro objetivo do presente estudo foi verificar se há correlação entre dois tipos de tentativas do IRAP (bom professor-positivo e mau professor-negativo) e os resultados da avaliação de professores bons e ruins no TBC. Verificou-se que não há correlação estatisticamente significativa entre o D-IRAP score do tipo de tentativa bom professor-positivo e valores no TBC dos participantes que avaliaram um professor considerado bom. Não foi verificada uma correlação entre o D-IRAP score do tipo de tentativa mau professor-negativo e valores no TBC dos participantes que avaliaram um professor ruim. Isso pode ser explicado pelo fato de que o IRAP exige um tempo resposta pequeno e uma porcentagem de acertos alta para o desempenho dos participantes, controlando a variável social no momento da aplicação. Ao impossibilitar que o participante responda sob controle do que imagina que o pesquisador espera ou do que é socialmente desejável, o IRAP é capaz de acessar as propriedades relevantes do responder relacional que não são identificadas em instrumentos explícitos de avaliação. Conforme citado por Cullen, Barnes-Holmes e Stewart (2009), essa diferença entre as respostas dos participantes entre as medidas implícitas e entre as medidas explícitas é explicada pelo Modelo de Elaboração e Coerência Relacional (*REC Model*) (Barnes-Holmes et al., 2010). De acordo com esse modelo, os instrumentos explícitos permitem que os participantes emitam respostas longas e elaboradas, e os instrumentos implícitos exigem que os participantes emitam respostas breves e imediatas, coerentes com sua história de aprendizagem pré-experimental, o que pode ter produzido um índice de correção estatisticamente não significativo entre os resultados do TBC e os D-IRAP escores.

Outro fator que pode ter sido responsável pelo índice de correlação estatisticamente não significativo entre os resultados do TBC e do IRAP foi que o fato de que no TBC os estudantes estavam avaliando um professor específico, que poderia ter algumas qualidades positivas e algumas qualidades negativas, à despeito de ser considerado bom ou ruim (e.g., um professor ruim pode, ainda assim, ter uma boa comunicação). Em contrapartida, no IRAP, os participantes avaliaram o conceito de bom e de mau professor, sem considerar características

individuais, o que levou a um índice estatisticamente não significativo. Essa informação corrobora os dados de Henklain (2017), que demonstrou haver variabilidade nas respostas de 676 participantes ao avaliarem bons e maus professores, apesar de haver tendência para avaliações mais positivas ou negativas (e.g., os participantes, ao avaliarem maus professores, não fizeram apenas avaliações negativas em todos os itens do TBC). No estudo de Henklain et al. (2019), foi encontrada uma correlação entre o tipo de teste bom professor-negativo-falso e os escores dos participantes no TBC. Ressalta-se, no entanto, que nesse estudo, os participantes não estavam avaliando professores específicos, mas sim o conceito de bom ou mau professor (tanto no IRAP quanto no TBC).

Para próximos estudos, sugere-se que a amostra seja maior e mais diversificada, uma vez que os participantes do presente estudo foram somente estudantes universitários do curso de Psicologia. Além disso, estudos futuros podem aplicar os instrumentos com a participação de docentes, como foi realizado por Barnes-Holmes et al. (2016) e Kelly e Barnes-Holmes (2013), para avaliar a diferença na percepção de professores sobre características importantes de docentes eficazes em relação à percepção de alunos. Por fim, considerando que o TBC original possui 28 itens, sugere-se que outros itens do TBC sejam avaliados, como a relação entre “Bom professor” e características, como “Bom ouvinte”, “Encoraja e se importa com os estudantes”, “Bom humor e atitudes positivas”, “Respeitoso”, “Compreensivo” e “Humilde”.

O presente estudo indica uma evidência de validade de conteúdo de seis itens do TBC, pois esses estímulos foram fortemente relacionados com “Bom professor” no IRAP. Os resultados também sugerem a eficácia do IRAP para avaliação de comportamentos característicos de professores eficazes e não eficazes, já que as médias dos D-IRAP escores indicaram uma coerência entre características positivas e “Bom professor”, bem como características negativas e “Mau professor”. Foi possível concluir, também, que a ordem de aplicação dos instrumentos não afetou os resultados do IRAP e do TBC, pois não foram identificadas diferenças estatisticamente significativas na comparação dos resultados dos dois grupos. Verificou-se também que o índice de correlação não foi estatis-

ticamente significativo entre os D-IRAP escores e os valores do TBC. Isso provavelmente ocorreu porque, no IRAP, os estudantes tiveram que emitir respostas breves e imediatas coerentes com sua história de aprendizagem pré-experimental, enquanto no TBC puderam emitir respostas com maior latência e elaboradas, as quais poderiam estar sob controle de variáveis sociais. Além disso, destaca-se que no IRAP, os alunos estavam avaliando o conceito de bom ou mau professor, e no TBC estavam avaliando docentes específicos que mesmo sendo considerados bons ou ruins, poderiam ter características positivas e negativas concomitantemente. A partir do índice de correlação estatisticamente não significativo entre a nota recebida na disciplina e a nota que o docente recebeu no TBC, sugere-se que, para além dos resultados das avaliações na disciplina, variáveis como empatia do professor em relação ao aluno podem interferir na avaliação dos docentes por parte dos estudantes.

Referências

- Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the implicit relational assessment procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist*, (32)7, 169-177. Recuperado de <https://opensiuc.lib.siu.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1040&context=tp>
- Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the Relational Elaboration and Coherence (REC) model. *The Psychological Record*, 60(3), 527-542. <https://doi.org/10.1007/BF03395726>
- Buskist, W., & Keeley, J. W. (2018). Searching for universal principles of excellence in college and university teaching. *New Directions for Teaching and Learning*, 156, 95-105. <https://doi.org/10.1002/tl.20321>

- Buskist, W., Sikorski, J., Buckley, T., & Saville, B. K. (2002). Elements of master teaching. In S. F. Davis & W. Buskist (Eds.), *The teaching of psychology: Essays in honor of Wilbert J. McKeachie and Charles L. Brewer* (pp. 30-39). New York, NY: Psychology Press. Recuperado de <https://psycnet.apa.org/record/2002-00149-000>
- Cullen, C., Barnes-Holmes, D., Barnes-Holmes, Y., & Stewart, I. (2009). The Implicit Relational Assessment Procedure (IRAP) and the malleability of ageist attitudes. *The Psychological Record*, 59(4), 591-620. <https://doi.org/10.1007/BF03395683>
- Farrell, L., & McHugh, L. (2017). Examining gender-STEM bias among STEM and non-STEM students using the Implicit Relational Assessment Procedure (IRAP). *Journal of Contextual Behavioral Science*, 6(1), 80-90. <https://doi.org/10.1016/j.jcbs.2017.02.001>
- Finn, M., Barnes-Holmes, D., & McEnteggart, C. (2018). Exploring the single-trial-type-dominance-effect in the IRAP: Developing a differential arbitrarily applicable relational responding effects (DAARRE) model. *The Psychological Record*, 68(1), 11-25. Recuperado de <https://doi.org/10.1007/s40732-017-0262-z>
- Golijani-Moghaddam, N., Hart, A., & Dawson, D. L. (2013). The Implicit Relational Assessment Procedure: Emerging reliability and validity data. *Journal of Contextual Behavioral Science*, 2, 105-119. <https://doi.org/10.1016/j.jcbs.2013.05.002>
- Henklain, M. H. O. (2017). *Aplicações do Teacher Behavior Checklist à formação e avaliação de professores de nível superior: Contribuições analítico-comportamentais e psicométricas*. Tese de doutorado, Universidade Federal de São Carlos (UFSCAR), São Carlos, São Paulo, Brasil. Recuperada de <https://repositorio.ufscar.br/handle/ufscar/9886>
- Henklain, M.H.O., Haydu, V. B., Carmo, J. S. Muniz, M., & Perez, W. F. (2019). Expanding the evidence of content validity for the Teacher Behavior Checklist using the IRAP. *The Psychological Record*, 69, 205-214. <https://doi.org/10.1007/s40732-019-00334-9>

- Jensen, W., & Fischer, B. (2006). Student perspectives on the qualities/behaviors of faculty that contribute most significantly to student learning. *International Journal of Construction Education and Research*, 2(2), 67-73. <https://doi.org/10.1080/15578770600775850>
- Keeley, J., Christopher, A. N., & Buskist, W. (2012). Emerging evidence for excellent teaching across borders. In J. E. Groccia, M. Al - Sudairy, & W. Buskist (Eds.s), *Handbook of college and university teaching: Global perspectives* (pp. 374-390). Sage.
- Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating psychology students' perceptions of teachers using the teacher behavior checklist. *Teaching of Psychology*, 37, 16-20. <https://doi.org/10.1080/00986280903426282>
- Keeley, J., Smith, D., & Buskist, W. (2006). The teacher behaviors check-list: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology*, 33(2), 84-91. https://doi.org/10.1207/s15328023top3302_1
- Kelly, A., & Barnes-Holmes, D. (2013). Implicit attitudes towards children with autism versus normally developing children as predictors of professional burnout and psychopathology. *Research in developmental disabilities*, 34(1), 17-28. <https://doi.org/10.1016/j.ridd.2012.07.018>
- Lammers, W. J., Savina, E., Skotko, D., & Churlyeva, M. (2010). Faculty and student perceptions of outstanding university teachers in the USA and Russia. *Educational Psychology: An International Journal of Experimental Educational Psychology*, 30(7), 803-815. <https://doi.org/10.1080/01443410.2010.512382>
- Liu, S., Keeley, J., & Buskist, W. (2015). Chinese college students' perceptions of characteristics of excellent teachers. *Teaching of Psychology*, 42(1), 83-86. <https://doi.org/10.1177/0098628315620888>
- Mizael, T. M., & de Almeida, J. (2019). Revisão de estudos do Implicit Relational Assessment Procedure sobre vieses raciais. *Acta Comportamental*, 27, 437-461. Recuperado de <https://www.redalyc.org/jatsRepo/2745/274561551003/index.html>

- Schaeffer, G., Epting, K., Zinn, T., & Buskist, W. (2003). Student and faculty perceptions of effective teaching: A successful replication. *Teaching of Psychology*, 30(2), 133-136. Recuperado de <https://goo.gl/7BGLmE>
- Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assessment & Evaluation in Higher Education*, 25(4), 397-405. [https://doi.org/ 10.1080/713611436](https://doi.org/10.1080/713611436)
- Spooren, P., & Mortelmans, D. (2006). Teacher professionalism and student evaluation of teaching: Will better teachers receive higher ratings and will better students give higher ratings? *Educational Studies*, 32(2), 201-214. <https://doi.org/10.1080/03055690600631101>
- Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry*, 48, 59-65. <https://doi.org/10.1016/j.jbtep.2015.01.004>
- Vahey, N., Boles, S., & Barnes-Holmes, D. (2010). Measuring adolescents' smoking-related social identity preferences with the Implicit Relational Assessment Procedure (IRAP) for the first time: A starting point that explains later IRAP evolutions. *International Journal of Psychology and Psychological Therapy*, 10(3), 453-474. Recuperado de <https://psycnet.apa.org/record/2014-13028-010>
- Wilson, J. H., & Ryan, R. G. (2012). Formative teaching evaluations: Is student input useful? In M. E. Kite (Org.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 22-29). Recuperado de <https://goo.gl/pSx91e>

Recibido: 2020-10-08

Revisado: 2021-07-30

Aceptado: 2021-11-28