UNIVERSIDADE DE ÉVORA

ESCOLA DE CIÊNCIAS SOCIAIS

DEPARTAMENTO DE PSICOLOGIA

Medidas Individualizadas de Mudança: Análise das Propriedades Psicométricas do PSYCHLOPS

Daniel Serpa Branco Guerra

Orientação: Prof.ª Doutora Célia Sales, Universidade de Évora

Coorientação: Prof. Doutor Luís Faísca, Universidade do Algarve

Mestrado em Psicologia

Área de especialização: *Psicologia Clínica e da Saúde*

Dissertação

Évora, 2016

UNIVERSIDADE DE ÉVORA

Escola de Ciências Sociais

Mestrado em Psicologia
Especialização em Psicologia Clínica e da Saúde

Medidas Individualizadas de Mudança: Análise das Propriedades Psicométricas do PSYCHLOPS

Daniel Serpa Branco Guerra

Orientação:
Prof.ª Doutora Célia Sales
Coorientação:
Prof. Doutor Luís Faísca

Évora, outubro de 2016

Agradecimentos

Os agradecimentos aqui feitos pretendem permitir o reconhecimento público de contribuições fundamentais para a conclusão do presente trabalho, mas não esgotam a gratidão pelo crescimento pessoal que, neste momento, se manifesta na conclusão desta dissertação. Esses agradecimentos continuarão a ser feitos pessoalmente.

Em primeiro lugar, agradeço aos meus orientadores, a Professora Célia Sales e o Professor Luís Faísca, pelo acompanhamento próximo e individualizado que me ofereceram, e pela paciência e disponibilidade na consideração de ideias por vezes confusas e pouco convencionais. Obrigado pela enorme liberdade dada na seleção de caminhos e de ritmos de trabalho. Quero deixar um agradecimento muito especial ao Professor Luís Faísca, por ter aceitado trabalhar à distância, em horários pouco agradáveis, por ter sempre tido a paciência de debater as ideias que iam surgindo, e pelas aprendizagens que me proporcionou ao longo deste tempo.

Deixo um agradecimento ao Dr. Mark Ashworth pela disponibilização das amostras do Reino Unido e da Polónia, e pela disponibilidade no esclarecimento de dúvidas referentes ao PSYCHLOPS, à Dra. Paula Alves pela disponibilização da segunda amostra portuguesa, e ao Dr. Helgi Héðinsson pela disponibilização da amostra Islandesa, e pelo trabalho na verificação da tradução da base de dados. Agradeço também à Fundação Romão de Sousa pela disponibilização dos dados recolhidos na Casa de Alba, que infelizmente não puderam ser utilizados no presente estudo.

Quero deixar um agradecimento profundo aos meus pais, pelo afeto e carinho incondicionais, expressos também na motivação que foram dando para a conclusão deste trabalho. Ao meu pai, em particular, agradeço a leitura atenta e as sugestões feitas a este trabalho. Obrigado à Cristina, minha namorada, pelo apoio multivalente dado ao longo de todo o processo, pelas revisões, e pelo debate de ideias.

Por fim, quero agradecer ao Eduardo Maia e à Rita Antunes, meus amigos, pela partilha do caminho e pelo apoio.

Medidas Individualizadas de Mudança: Análise das propriedades psicométricas do PSYCHLOPS

Resumo

Neste estudo apresentam-se dados referentes à fiabilidade, validade e sensibilidade para detetar mudança de uma medida individualizada de mudança (MIM), o PSYCHLOPS. Os dados foram recolhidos em quatro países, e dividem-se por seis amostras, totalizando 1053 participantes. Pretende-se aprofundar o conhecimento acerca das propriedades psicométricas e da sua adequabilidade, em MIM no geral, e no PSYCHLOPS, em particular. Após avaliar as propriedades psicométricas separadamente por amostra, procedeu-se à sua integração mediante métodos de meta-análise. O valor global de *alfa* de .82 indica uma boa consistência interna, mas a sua interpretação é dificultada pela não equivalência de itens. A correlação entre o PSYCHLOPS e o CORE-OM é de .63, e a correlação entre os *scores* de mudança das duas medidas é .67. Os dados suportam as hipóteses referentes à validade, mas suscitam dúvidas quanto à estrutura do PSYCHLOPS, e quanto à adequabilidade da avaliação psicométrica tradicional em MIM.

Palavras-chave: PSYCHLOPS, Medidas individualizadas de mudança, psicometria, meta-análise.

Client-Generated Outcome Measures: Analysis of the Psychometric Properties of PSYCHLOPS

Abstract

Various evidence regarding reliability, validity, and responsiveness of PSYCHLOPS, a patient-generated outcome measure (PGOM), are presented. Data comes from three countries, in a total of seven data sets comprising 1053 participants. The purpose of this study is to deepen the understanding of psychometric properties in PGOMs, and in PSYCHLOPS in particular. Following psychometric analysis by sample, meta-analysis procedures were used to calculate overall values. Overall mean internal consistency (alpha) was .82. Overall mean correlation between PSYCHLOPS and CORE-OM is r = .63, and overall mean correlation between change scores is r = .67. Overall mean internal consistency is good, but item non-equivalency makes interpretation difficult. Data support hypotheses regarding correlations, but cast doubt over PSYCHLOPS' internal structure, and over suitability of classical psychometric evaluation in PGOM.

Keywords: PSYCHLOPS, Patient-Generated Outcome Measures, Psychometric analysis, Meta-analysis.

Índice

1. Introdução

A necessidade atual que a comunidade científica tem de desenvolver tratamentos empiricamente suportados, ou validados, obriga a que os métodos utilizados para avaliação da eficácia de um tratamento estejam sujeitos a um maior escrutínio (Sales, Gonçalves, Fragoeiro, Noronha, & Elliott, 2007). Apesar da grande disseminação de medidas nomotéticas estandardizadas, a sua utilização exclusiva para medir os resultados da psicoterapia pode ser posta em causa porque estas não permitem que o paciente faça parte da escolha de assuntos, problemas e dificuldades que servirão para avaliar a eficácia de um tratamento para os seus problemas particulares (Evans et al., 2010; Robinson, Ashworth, Shepherd, & Evans, 2006). Alternativamente, as medidas individualizadas de mudança (MIM), informam sobre a mudança em problemas e dificuldades indicados pelos próprios pacientes (Sales, Gonçalves, Fragoeiro, Noronha, & Elliott, 2007). Estão, também, em consonância com a importância crescente atribuída à consideração da perspetiva do paciente no que concerne à avaliação do tratamento que é recebido (Robinson et al., 2006; Sales & Alves, 2012). Concomitantemente, a consideração de problemas e dificuldades identificadas pelos pacientes pode contribuir para a robustez de estratégias de avaliação distintas (ou complementares) em ensaios clínicos aleatorizados (Sales & Alves, in press).

A utilização de métodos alternativos terá de ser suportada por dados empíricos oriundos de investigações rigorosas. Nesse sentido, o presente estudo pretende contribuir para um maior conhecimento das características psicométricas das MIM, em geral, e, em particular, do PSYCHLOPS (Ashworth et al., 2004). Assim, para além da avaliação das propriedades psicométricas deste instrumento, com recurso a dados provenientes de diferentes países e contextos de intervenção, será feita uma revisão teórica do conhecimento referente às propriedades psicométricas em instrumentos nomotéticos. O propósito central desta revisão será o de refletir acerca da adequabilidade das propriedades psicométricas à avaliação de MIM, particularmente no que concerne aos métodos estatísticos de avaliação dessas propriedades. Serão também consideradas questões teóricas e conceptuais subjacentes a propriedades psicométricas e que coloquem desafios à avaliação de MIM. A comparação com uma medida nomotética de avaliação da mudança robusta, o CORE-OM (Evans et al., 2000), servirá como base para a avaliação das propriedades do PSYCHLOPS.

2. Enquadramento teórico

2.1. Medidas de avaliação de *outcome* da psicoterapia

Ainda que exista um consenso geral referente à importância da psicoterapia enquanto forma indispensável de tratamento de diferentes problemas de saúde mental, a produção de conhecimento científico referente a instrumentos de avaliação de mudança tem sido preterida em favor de outras questões de investigação (Doucette & Wolf, 2009; Ogles, 2013). O desenvolvimento de novas formas de tratamento para uma determinada perturbação, ou a identificação de características e processos do paciente e do terapeuta associados a intervenções bem-sucedidas, são, mais frequentemente, foco de investigação do que o desenvolvimento de medidas de mudança com qualidade elevada (Ogles, 2013). Há, todavia, evidência de que a medição do progresso do paciente oferece benefícios significativos a terapeutas e pacientes (Nordal, 2012), o que justifica o desenvolvimento de medidas de avaliação robustas. Considerar a facilitação da mudança benéfica no paciente como propósito da psicoterapia (Doucette & Wolf, 2009) torna central o desenvolvimento de medidas de avaliação da mudança para validação de uma intervenção (Ogles, 2013)

Pode atribuir-se a crescente importância da avaliação de resultados em psicoterapia à necessidade que os serviços de saúde mental têm de comprovar a qualidade das intervenções que realizam (Barkham et al., 2001). Contudo, o aumento particular da importância atribuída a métodos de avaliação da mudança que incluam o ponto de vista do paciente deve-se à crescente importância atribuída ao cuidado centrado no paciente (Deshpande, Rajan, Sudeepthi, & Nazir, 2011; Sales, 2016). Recomendado em diretrizes referentes à qualidade da prestação de cuidados de saúde (como as recomendações *NICE*¸ ou do *Institute of Medicine*), os cuidados centrados no paciente envolvem o recurso a *patient-reported outcome measures* (PROMs) para planeamento da intervenção e avaliação de resultados (Fitzpatrick, Davy, Buxton, & Jones, 1998). As PROMs definem-se como um conjunto de medidas que permitem ao paciente avaliar aspetos como o seu estado de saúde, ou o seu funcionamento, a partir de questões referentes a sintomas, dificuldades funcionais, bem-estar, ou à qualidade de vida (Staniszewska, Haywood, Brett, & Tutton, 2012). Estas medidas permitem a inclusão do ponto de vista do paciente no processo de avaliação, e dividem-se entre medidas de variáveis genéricas presentes em populações clínicas, e medidas específicas de avaliação de perturbações ou problemas concretos (Sales, 2016). Adicionalmente, estas medidas podem ser utilizadas para monitorização do progresso do paciente, de forma a permitir informar o

terapeuta acerca da resposta ao tratamento (Fitzpatrick, 2012; Overington & Ionita, 2012).

2.1.1. Abordagens nomotéticas

A avaliação de resultados em psicoterapia é maioritariamente feita através de PROMs estandardizadas, compostas por itens escolhidos por profissionais ou provenientes da literatura (Ashworth, Evans, & Clement, 2009). Apelidadas de medidas nomotéticas, estas escalas consistem num conjunto pré-determinado de itens que devem ser respondidos pelos pacientes (Héðinsson, Kristjánsdóttir, Ólason, & Sigurðsson, 2012), e permitem o estabelecimento de normas e de pontos de corte (Evans, Margison, & Barkham, 1998). Uma vez que todos os sujeitos devem responder ao mesmo conjunto de itens, nas medidas nomotéticas assume-se que os resultados representam variabilidade numa dimensão onde todos podem obter uma classificação (Ashworth, Robinson, Evans, Shepherd, Conolly, & Rowlands, 2007). Para além de permitir a comparação do paciente com normas retiradas de uma amostra da população (Ogles, 2013), a utilização destas medidas pode ajudar na redução da sintomatologia e na melhoria do progresso clínico através do *feedback* recebido pelo terapeuta, referente ao progresso clínico (Sales & Alves, 2012). A importância do *feedback* recebido através de medidas nomotéticas reflete-se, por exemplo, na tomada de decisão clínica referente à dosagem do tratamento ou aos assuntos a abordar em sessão (Kelley, Athay, Hargraves, Andrade, Tempesti, & Bickman, 2011), bem como no incremento da motivação para o tratamento (Rottger, Ruber, & Lutz, 2011), existindo resultados preliminares que apontam para uma eficácia maior em terapeutas que recorrem a sistemas de *feedback* (Reese et al., 2009). A título de exemplo, Lambert e colaboradores (2001) conduziram um estudo para perceber se pacientes cujos terapeutas receberam *feedback* do seu progresso (com recurso a uma medida nomotética, o OQ-45) melhoravam mais do que pacientes cujos terapeutas não obtiveram *feedback*. Globalmente, nos casos em que o terapeuta recebeu *feedback,* verificou-se uma maior percentagem de mudança clinicamente significativa, uma menor percentagem de casos em que os pacientes pioraram, e uma redução nos custos do tratamento decorrente da diminuição da dosagem do tratamento em casos com boa evolução, e do ajustamento do tratamento em casos mais graves.

As medidas nomotéticas são globalmente precisas e objetivas, permitem antecipar o comportamento, e são aceites pela comunidade científica (Ashworth et al. 2007), em parte porque existe uma longa tradição de desenvolvimento deste tipo de escalas. Contudo, as medidas nomotéticas apresentam limitações como medidas da

mudança em psicoterapia, particularmente no que concerne à adequação ao contexto clínico. Em primeiro lugar, este tipo de medidas, por produzirem resultados relativos a itens pré-definidos, referentes a problemas comuns em populações clínicas (Evans, Ashworth, & Peters, 2010), pode não ser exaustivo no que concerne aos problemas apresentados individualmente, por cada paciente (Hunter et al., 2004). Em segundo lugar, há o risco de que itens potencialmente adequados para a avaliação de uma determinada dimensão sejam descartados durante o desenvolvimento do instrumento, por não ser provável que todos os pacientes possam responder-lhes, não sendo por isso incluídos na versão final (Evans et al., 2010). Em terceiro lugar, a possível escassez de itens relevantes para os problemas particulares do paciente pode resultar numa menor sensibilidade para a mudança (Héðinsson et al., 2012). Por fim, o conteúdo de instrumentos nomotéticos resulta geralmente da opinião de profissionais e não da visão dos pacientes acerca dos seus problemas (Evans et al., 2010), o que contrasta com a reconhecida importância da consideração da perspetiva dos utilizadores em relação aos seus problemas (Robinson, Ashworth, Shepherd, & Evans, 2006). A utilização de um outro tipo de medidas, as idiográficas, representa uma alternativa às medidas nomotéticas que pode ajudar na identificação de problemas que são significativos para o paciente que não estão incluídos nas medidas nomotéticas (Héðinsson et al., 2012).

2.1.2. Abordagens idiográficas e o papel do envolvimento do utilizador

Os termos “idiográfico” e “nomotético” foram originalmente criados pelo filósofo alemão Wilhelm Windelband (1848-1915), para permitir a distinção entre duas formas diferentes de conhecimento baseado em evidência (Robinson, 2011). Sucintamente, enquanto o conhecimento nomotético pretende encontrar leis gerais e teorias que são comuns a um determinado conjunto de sujeitos, o conhecimento idiográfico pretende descrever e explicar fenómenos particulares (Robinson, 2011). Ainda que tenha sido Münsterberg a introduzir a distinção entre os termos “idiográfico” e “nomotético” no campo da Psicologia (Hurlburt & Knapp, 2006; Robinson, 2011), é ao psicólogo estadunidense Gordon Allport (1937, 1960) que geralmente se atribui a introdução das estratégias idiográficas em Psicologia (Elliott, Wagner, Sales, Rodgers, Alves, & Café, 2016). De acordo com o autor, nesta área é essencial considerar aquilo que é particular em cada indivíduo, e não apenas as assunções universais referentes à espécie humana. Por sua vez, Pascal e Zax (1956), ao selecionarem comportamentos individuais como critério para avaliação da mudança em três tipos de terapia,

encontram-se entre os primeiros a desenvolver e utilizar medidas idiográficas em contexto clínico.

A avaliação idiográfica inclui métodos, estratégias e instrumentos selecionados para um determinado sujeito, e define-se como a medição de variáveis e de relações funcionais individualmente selecionadas ou decorrentes de estímulos ou contextos de avaliação individualizados (Haynes, Mumma, & Pinson, 2009). No que concerne à avaliação de *outcome* em psicoterapia, os instrumentos idiográficos consistem em medidas individualizadas, dirigidas à avaliação da mudança psicológica (Sales & Alves, 2012). As medidas individualizadas de mudança (MIM), também apelidadas de *patient-generated outcome measures* (PGOMs), ou *individualized patient-reported outcome measures* (IPROMs) (Sales & Alves, in press), correspondem a instrumentos onde é permitido ao paciente selecionar domínios ou assuntos que o/a preocupam, que devem ser atendidos durante o tratamento, e que podem não ter sido selecionados previamente pelos investigadores que desenvolveram determinado questionário (Fitzpatrick et al., 1998). A seleção de domínios e/ou assuntos é feita pelo paciente, mediante a definição dos itens que constarão do instrumento (Ashworth et al., 2004), mantendo uma estrutura estandardizada que define o formato do instrumento (Sales & Alves, in press).

As MIM permitem a avaliação do progresso clínico de acordo com tópicos relevantes para os pacientes (Sales & Alves, 2012), o que significa que possam ser mais sensíveis à mudança do que as medidas nomotéticas (Ashworth et al., 2009; Ogles, 2013). Adicionalmente, estas medidas evitam que seja desperdiçado tempo a avaliar sintomas irrelevantes (Wagner & Elliott, 2001), e são flexíveis ao permitir que o paciente inclua itens referentes a variáveis individuais como a personalidade, o estatuto socioeconómico ou o contexto cultural na avaliação (Sales & Alves, 2012). A utilização de MIM contribui também para a realização de boas práticas clínicas ao permitir que o paciente diga, com as suas palavras, o que verdadeiramente o preocupa (Sales & Alves, in press). A liberdade na expressão dos problemas do paciente facilita a clarificação dos seus objetivos, o que é particularmente importante se se considerar que existe uma maior disposição para o envolvimento no tratamento quando os objetivos refletem as necessidades individuais do paciente e quando são percecionados como significativos (Turner-Stokes, 2009).

A consideração da perspetiva do paciente acerca dos seus problemas, característica das MIM, é consonante com o reconhecimento da importância da perspetiva do utilizador no planeamento, implementação e avaliação dos cuidados de saúde (Robinson et al., 2006). O conceito de envolvimento do utilizador, ou *user involvement*, é referente à tentativa de aumentar a influência real exercida pelos

utilizadores de serviços de saúde nas decisões referentes ao seu tratamento (Storm & Edwards, 2013), e trata-se de um tópico central de políticas sociais de serviços de saúde mental em vários países (Petersen, Hounsgaard, Borg, & Nielsen, 2012), e em planos de saúde mental internacionais (Storm & Edwards, 2013). Da mesma forma, a promoção de uma abordagem médica centrada no paciente (designada por *patient-centered care*), tem sido incluída nos objetivos centrais de melhoria dos cuidados de saúde, em países como o Reino Unido, ou os Estados Unidos (Mead & Bower, 2000; Zill, Scholl, Härter, & Dirmaier, 2015). Ainda que não de forma consensual, as várias definições de cuidado centrado no paciente colocam a consideração das preferências, valores, e necessidades no centro da tomada de decisão acerca da prestação de cuidados médicos, o que torna o conceito indissociável do conceito de envolvimento do utilizador (Gerteis, Edgman-Levitan, Daley, & Delbanco, 1993; Berwick, 2002).

Fundamental para a reabilitação em saúde mental (Roberts, Davenport, Holloway, & Tartan, 2006), o incremento do envolvimento do utilizador reflete a passagem de um modelo paternalista de cuidado, centrado na autoridade do prestador de cuidados (Storm & Edwards, 2013), para um modelo baseado na parceria entre o utilizador e o/a prestador/a de cuidados (Petersen et al., 2012). Baseada no respeito mútuo e no reconhecimento das vantagens que a parceria terá no cumprimento dos objetivos, esta parceria deve refletir-se no trabalho colaborativo, de forma a atingir os objetivos definidos (Coutler, 1999). Assim, o utilizador deve também estar envolvido na avaliação do seu progresso e na seleção de medidas de mudança que avaliem a eficácia do tratamento (Entwistle, Renfrew, Yearley, Forrester, & Lamont, 1998).

A diferença entre a forma como os utilizadores e os prestadores de cuidados avaliam a eficácia de uma intervenção (e.g. Rothwell, McDowell, Wong, & Dorman, 1997) justifica o envolvimento do utilizador no desenvolvimento e na escolha de uma medida de mudança, particularmente porque esta diferença pode ser mais vincada no que respeita à saúde mental (Faulkner & Thomas, 2002). Nesse sentido, Crawford e colaboradores (2011) realizaram um estudo onde utilizadores de um serviço de saúde mental com perturbações psicóticas e/ou afetivas apresentaram a sua perspetiva acerca de diferentes medidas de mudança. Os utilizadores demonstraram, em relação a algumas medidas, preocupação acerca da capacidade de captação das suas experiências individuais, tendo as PROMs sido globalmente avaliadas como mais relevantes e apropriadas do que medidas preenchidas por profissionais, por observação do paciente. No fundo, parece ser possível ordenar de forma contínua as medidas de mudança de acordo com o grau de envolvimento do paciente na avaliação da mudança (Fitzpatrick et al., 1998), sendo que este influencia a confiança dos pacientes na capacidade dos instrumentos na captação de experiências individuais.

Assim, se as PROMs são avaliadas como mais relevantes do que medidas preenchidas por profissionais, porque mais capazes de captar experiências individuais (Crawford et al., 2011), será razoável considerar que as MIM permitam captar essas experiências ainda melhor, porque permitem um maior envolvimento do utilizador. Ao mesmo tempo, podem estar ausentes dificuldades associadas à extensão dos itens, à complexidade, ou a enviesamentos culturais, que podem suscitar reações negativas e comprometer a validade e utilidade de medidas nomotéticas (Blount, Evans, Birch, Warren, & Norton, 2002).

As MIM podem promover o envolvimento do utilizador e dar uma resposta adequada à dificuldade de captação das experiências individuais dos pacientes, uma vez que há evidência de que é frequente que pacientes indiquem preocupações nestas medidas que não estão incluídas nas medidas nomotéticas (Ashworth et al., 2007; Hunter et al., 2004; Wagner & Elliott, 2001). Concomitantemente, o facto de o conteúdo das MIM exigir uma interação com o paciente (Sales & Alves, in press) é vantajoso tendo em conta que o processo de recolha de dados é, para os pacientes, tão ou mais importante do que o conteúdo dos instrumentos (Crawford et al., 2011).

Uma revisão de literatura recente (Sales & Alves, in press), identificou a existência de três MIM em saúde mental, as quais, apesar de terem em comum a possibilidade de geração do seu conteúdo pelo paciente, têm características e estruturas distintas: o *Goal Attainment Scaling* (Kiresuk & Sherman, 1968), que requer a identificação e a descrição dos objetivos que o paciente quer atingir em terapia, bem como as expetativas quanto ao *outcome* (Turner-Stokes, 2009); o *Simplified Personnal Questionnaire* (Elliott, Mack & Shapiro, 1999), que consiste numa lista de itens construídos com a linguagem do paciente, retirados de uma entrevista semiestruturada na qual o paciente reporta os problemas que motivaram o pedido de acompanhamento (Elliott et al., 2016); o PSYCHLOPS (Ashworth et al., 2004), um instrumento de autorrelato no qual o paciente identifica os dois problemas que mais o afetam, bem como as atividades que, consequentemente, tem maior dificuldade em realizar (será também necessário responder a uma questão nomotética referente ao bem-estar).

Ainda que o valor das MIM seja suportado por vários estudos (e.g. Ashworth et al., 2009; Elliott et al., 2016; Sorenson, Gorsuch, & Mintz, 1985), há desvantagens que devem ser consideradas. Primeiramente, a utilização de algumas destas medidas (o *Personnal Questionnaire* e o *Goal Attainment Scaling*) é dificultada pela extensão e complexidade da sua administração, sendo necessária a orientação de um entrevistador treinado (Ashworth et al., 2005). Adicionalmente, o conteúdo temático individualizado destas medidas torna incerto o significado das pontuações e das normas (Ashworth et al. 2007), o que as leva a serem acusadas de apenas permitir a

avaliação da mudança intraindividual, não permitindo a comparação dos resultados com populações clínicas (Lacasse, Wong, & Guyatt, 1999). Ao mesmo tempo, a comparação dos resultados apenas pode ser feita globalmente, e não por dimensões. Por fim, a comparação de casos é também dificultada pela variabilidade no número de itens (como no caso do *Personnal Questionnaire*), que exige acrescentar à comparação de conteúdos e respostas a comparação do número de itens (Sales e Alves, in press).

Devido a uma elaboração reduzida e pouco frequente dos princípios teóricos que subjazem a avaliação idiográfica, as propriedades psicométricas destas medidas não têm sido sistematicamente sujeitas a uma avaliação rigorosa (Haynes et al., 2009). Para tal contribui, por um lado, uma adoção pouco abrangente das MIM em contextos de saúde mental primária, o que dificulta a condução de estudos referentes às propriedades psicométricas (Ashworth et al., 2009). Por outro lado, as características deste tipo de medidas colocam desafios aos pressupostos da análise psicométrica tradicional. Para além da incerteza associada ao significado das pontuações (Ashworth et al., 2007), a principal dificuldade na análise psicométrica destas medidas reside na inadequação da abordagem clássica de medida em Psicologia, que assenta na comparação de pontuações obtidas por diferentes sujeitos nos mesmos itens (Haynes et al., 2009). A incompatibilidade deste pressuposto com a avaliação idiográfica liga-se à seleção ou geração idiossincrática dos itens por parte do paciente, o que impede a sua integração em medidas compósitas. Por este motivo, enquanto classicamente os instrumentos nomotéticos recorrem a métodos estatísticos quantitativos (Evans et al., 1998), a análise das MIM contemplam a utilização combinada de métodos quantitativos e qualitativos (Ashworth et al., 2007).

Assim, é pela ausência de investigação referente às propriedades psicométricas de MIM que se justifica o presente estudo: para que seja possível complementar a evidência qualitativa, referente à especificidade dos conteúdos incluídos nestas medidas (Ashworth et al., 2007; Hunter et al., 2004; Wagner & Elliott, 2001), com informação psicométrica que permita compreender se existe sustentação para a utilização destas medidas em prática clínica, e como forma de avaliação de serviços de saúde mental.

2.2. Uma medida individualizada de mudança - PSYCHLOPS

O PSYCHLOPS (Psychological Outcome Profiles) (Ashworth et al., 2004) é uma medida individualizada de mudança, de autopreenchimento, desenvolvida para utilização em serviços de saúde mental. Iniciado em 1999, o seu desenvolvimento foi

impulsionado pela procura de um instrumento capaz de captar aspetos do progresso do paciente, ausentes de instrumentos convencionais (Ashworth, Kordowicz, & Schofield, 2012). Concretamente, alguns terapeutas reportaram casos de pacientes que aparentavam ter resolvido grande parte dos problemas que motivaram o acompanhamento, mas cujas melhorias não eram captadas por instrumentos convencionais de avaliação do progresso clínico, tendo o PSYCHLOPS sido desenvolvido para permitir captar a mudança individual, de forma simples (Ashworth et al., 2009).

O desenvolvimento do PSYCHLOPS teve como base uma medida médica de avaliação da saúde física geral, o MYMOP (Paterson, 1996). Consistindo num questionário simplificado de uma página, o tempo de administração do MYMOP não excede o tempo de uma consulta de um médico de clínica geral, permitindo medir aspetos e efeitos de determinada doença considerados pelo paciente como sendo os mais importantes (Ashworth et al., 2004). Corresponde a uma medida individualizada de mudança de administração simples e breve (Héðinsson et al., 2012), cuja validação foi feita através da comparação com a escala SF-36 (Paterson, 1996), e que é sensível à mudança (Ashworth et al., 2004).

À semelhança do MYMOP, o PSYCHLOPS tem uma página e contém quatro questões referentes a três domínios: duas questões relativas ao domínio Problemas; uma questão referente ao domínio Funcionamento; uma questão relativa ao Bem-estar (Ashworth et al., 2009). Nas duas primeiras questões, referentes ao domínio Problemas, o paciente deverá descrever livremente o(s) problema(s) que mais o afetam (Ashworth et al., 2005), sendo também necessário avaliar a intensidade dos mesmos (Ashworth et al., 2012), bem como a sua duração (Ashworth et al., 2005). Na terceira questão, referente ao domínio Funcionamento, é pedido ao paciente que identifique uma tarefa cuja realização se tenha tornado difícil como consequência do(s) problema(s) identificado(s) (Ashworth et al., 2004), e que avalie o grau de dificuldade sentido na realização da tarefa. Por fim, a questão referente ao domínio Bem-estar, por carecer de uma componente de descrição livre, é nomotética, devendo o paciente avaliar como se sentiu consigo mesmo na semana anterior (Ashworth et al., 2012). As questões referentes à severidade dos problemas, à dificuldade na realização das tarefas identificadas, e ao bem-estar são respondidas com recurso a uma escala ordinal de seis níveis cujos valores variam entre 0 e 5. Os três domínios incluídos no PSYCHLOPS decorrem de um modelo que descreve uma sequência empírica de causalidade: são os problemas que despoletam dificuldades funcionais, as quais, por sua vez, resultam numa diminuição do bem-estar (Ashworth et al., 2012).

O PSYCHLOPS está disponível em três versões distintas: uma versão pré-terapia, uma versão intermédia (a ser administrada durante a terapia), e uma versão pós-terapia (Ashworth et al., 2012), estando todas elas disponíveis em Inglês, Francês, Português, Espanhol, Polaco, Islandês, Árabe, Sorâni, Farsi, e Coreano, no site www.psychlops.org. Neste site é também possível aceder a versões para utilização com crianças, pré e pós-terapia, e para utilização com adolescentes, encontrando-se neste caso também disponível a versão para utilização durante a terapia. Tendo inicialmente sido criado para administração pré e pós-terapia, foi posteriormente desenvolvida uma versão intermédia que permite administrações repetidas e a análise dos dados de acordo com um modelo de medidas repetidas (Czachowski, Seed, Schofield, & Ashworth, 2011). Esta versão permite também o acompanhamento do progresso clínico ao longo da terapia (Ashworth et al., 2012). A versão intermédia distingue-se da versão pré-terapia por incluir uma questão adicional acerca do surgimento de novos problemas, devendo o paciente descrevê-los e avaliar a sua intensidade (Czachowski et al., 2011). Por sua vez, a versão pós-terapia, para além das quatro questões presentes na versão pré-terapia, inclui duas questões nomotéticas: uma referente aos efeitos de problemas que surgiram ao longo da terapia, e outra onde é perguntado ao paciente como se sente consigo mesmo no fim da terapia. Contudo, o *score* total é obtido através da soma dos resultados das quatro primeiras perguntas (caso o paciente apenas identifique um problema, o *score* deste deve ser duplicado), servindo as questões adicionais incluídas nas versões intermédia e pós-terapia apenas para fornecer informação clínica ao terapeuta.

Diferentes estudos avaliaram as propriedades psicométricas do PSYCHLOPS (e.g. Ashworth et al., 2005; Ashworth et al., 2009; Czachowski et al., 2011; Evans et al., 2010). Concretamente, foi avaliada a consistência interna e a sensibilidade à mudança, medida através da magnitude do efeito (Ashworth et al., 2005; Ashworth et al., 2008; Czachowski et al., 2011), a estabilidade teste-reteste (Evans et al., 2010), e a validade convergente através da comparação com o CORE-OM (Ashworth et al., 2005) e com o HADS (Ashworth et al., 2008). A perceção de terapeutas em relação ao PSYCHLOPS também foi avaliada, tendo sido possível compreender que o instrumento é percecionado como oferecendo informação complementar a instrumentos nomotéticos (Ashworth et al., 2005b). A possibilidade de utilização do PSYCHLOPS em diferentes problemas é evidente, por exemplo, num estudo realizado por Davy e colaboradores (2012), no qual o instrumento foi utilizado para avaliação da mudança em pacientes com insónia. Embora o PSYCHLOPS esteja traduzido em diferentes línguas (Ashworth et al., 2012), existe apenas um estudo de validação e replicação (que pretendeu averiguar se o instrumento mostra um desempenho

semelhante numa amostra de um país diferente) referente a uma população clínica islandesa (Héðinsson et al., 2012). O PSYCHLOPS foi incluído no sistema de avaliação do progresso do paciente CORE Net (que pode ser acedido em http://www.coreims.co.uk/).

2.3. As propriedades psicométricas das MIM

O desenvolvimento de um instrumento de avaliação psicológica inclui a importante tarefa de analisar as suas propriedades psicométricas (Ogles, 2013). Contudo, apesar da popularidade crescente das MIM, há uma escassez de dados referentes às propriedades psicométricas destas medidas (Elliott et al., 2016). Independentemente das dificuldades associadas à análise psicométrica das MIM (Haynes et al., 2009), as propriedades psicométricas que devem ser avaliadas neste tipo de instrumentos são comuns a outras medidas de mudança, e são referentes aos domínios: fiabilidade; validade; sensibilidade para detetar mudança (Fitzpatrick et al., 1998; Mokkink et al., 2010). Existem ainda propriedades ligadas à utilidade clínica de um instrumento (como a aceitabilidade ou a *feasibility*) que vão para além do âmbito do presente estudo.

2.3.1. Fiabilidade

A fiabilidade de uma medida diz respeito à capacidade que esta tem para produzir os mesmos resultados quando aplicada a alvos estruturalmente iguais (Marôco & Garcia-Marques, 2006). Referente à consistência interna, à fiabilidade teste-reteste e ao erro de medida, a avaliação do domínio fiabilidade permite compreender até que ponto a medida proporcionada por um instrumento está isenta de erro aleatório (Fitzpatrick et al., 1998; Mokkink et al., 2010), ou seja, em que grau esse instrumento permite conhecer o verdadeiro *score,* capacidade, classificação ou medida do objeto (Marôco e Garcia-Marques, 2006). A fiabilidade corresponde a um importante domínio de qualquer medida de mudança uma vez que é essencial que este tipo de instrumentos permita concluir que mudanças observadas ao longo de uma intervenção se devem à mesma e não a variações aleatórias nas medidas fornecidas pelo instrumento (Fitzpatrick et al., 1998).

Consistência Interna

A consistência interna é a dimensão da fiabilidade referente à correlação entre os itens de um determinado questionário (Terwee, et al., 2007). Para que um instrumento possua consistência interna é necessário que os itens que o compõem meçam diferentes aspetos de um mesmo constructo, e não diferentes constructos (Streiner & Norman, 1995). A necessidade de garantir esta homogeneidade dos itens decorre de um princípio básico da teoria clássica da medida, que afirma que várias observações produzem uma estimativa mais fiável (isenta de erro aleatório) do que uma observação isolada, sendo por esse motivo que geralmente um constructo se avalia através de mais do que um item num questionário (Fitzpatrick et al., 1998). Classicamente, a consistência interna é atingida através de uma definição adequada do constructo a medir, de itens bem construídos que operacionalizem esse constructo, e com recurso a técnicas de análise fatorial, para determinar se os itens do questionário compõem uma única dimensão (Terwee et al., 2007).

A consistência interna pode ser medida de diferentes formas, sendo o cálculo do coeficiente alfa de Cronbach o método utilizado com maior frequência (Field, 2009). Baseado no método *split-half,* que consiste na divisão aleatória dos itens de um instrumento em dois grupos para avaliação do grau de concordância entre ambos os grupos (Fitzpatrick et al., 1998), o alfa de Cronbach consiste essencialmente na estimação do nível de concordância médio de todas as partições *split-half* possíveis (Cronbach, 1951). O cálculo do alfa de Cronbach deve ser efetuado após ter sido determinado o número de subescalas homogéneas que um instrumento integra, devendo ser calculado individualmente para cada uma delas (Terwee, 2007). Valores elevados neste coeficiente indicam uma consistência interna elevada (i.e., uma correlação alta entre os itens), enquanto valores baixos apontam para a inexistência de correlação (Fitzpatrick et al., 1998; Terwee et al., 2007). Contudo, uma correlação perfeita entre os itens de uma escala é indicadora de redundância, motivo pelo qual se sugere que o valor do alfa de Cronbach não deva ser inferior a .70 nem superior a .90 (Nunnally & Bernstein, 1994). A avaliação da consistência interna pode ser realizada também através da análise da correlação entre os itens individuais e a pontuação total da escala sem o item avaliado (correlação item-total), não devendo o valor dessas correlações ser inferior a .20 (Steiner & Norman, 1995).

Apesar da importância e utilidade do cálculo do alfa de Cronbach para avaliação da consistência interna de uma medida, a sua interpretação deve ser elaborada com cuidado uma vez que o seu valor é influenciado pelo número de itens do instrumento (Field, 2009). Concretamente, valores elevados de alfa podem ser

obtidos mesmo quando a correlação média entre itens é baixa ou quando existe multidimensionalidade, bastando para isso que o instrumento inclua um número elevado de itens (Cortina, 1993). Assim, a interpretação deste coeficiente deve ter em conta o nível médio da correlação entre itens, devendo a unidimensionalidade da escala ser verificada *a priori* mediante técnicas de análise fatorial (Cortina, 1993), uma vez que o alfa de Cronbach não informa sobre dimensionalidade (Marôco & Garcia-Marques, 2006). Independentemente da importância de garantir a consistência interna no desenvolvimento de uma escala, é relevante notar que a enfatização excessiva desta característica psicométrica pode significar que itens importantes sejam descartados no processo de construção da escala, particularmente aqueles que refletem a complexidade do fenómeno em avaliação, por não permitirem alcançar um coeficiente de fiabilidade elevado (Donovan, Frankel, & Eyles, 1993).

As dificuldades na avaliação da consistência interna em MIM decorrem de os diferentes sujeitos não estarem a ser avaliados pelos mesmos itens. Considerando que a consistência interna estima o erro de medida associado a um item, mediante a análise da variabilidade desse item numa amostra de sujeitos (Marôco & Garcia-Marques, 2006), a geração idiossincrática dos itens que compõem as MIM dificulta a estimação do erro de medida e, consequentemente, a avaliação da fiabilidade através da consistência interna. Não obstante, apesar de não ser possível calcular o coeficiente de correlação para itens individuais, alguns autores apresentam o valor do alfa de Cronbach obtido no estudo das propriedades psicométricas de MIM (e.g. Ashworth et al., 2005; Elliott et al., 2016; Héðinsson et al., 2012). Por exemplo, Elliott e colaboradores (2016), num estudo sobre as propriedades psicométricas do *Personal Questionnnaire* (Elliott et al., 1999), avaliaram a consistência interna através do cálculo do alfa de Cronbach, tanto com base na variação interindividual (recorrendo à correlação entre as avaliações dos diferentes itens gerados através do PQ numa determinada amostra), como com base na variação intraindividual (recorrendo à correlação entre as avaliações repetidas de itens gerados pelo mesmo sujeito em aplicações sucessivas do PQ). Neste caso, apenas os procedimentos referentes à variação intraindividual dão, em parte, resposta às dificuldades decorrentes da não equivalência dos itens gerados por MIM.

Fiabilidade Teste-reteste

A fiabilidade teste-reteste é referente à consistência temporal das pontuações obtidas ao administrar um instrumento em dois momentos distintos (Polit, 2014). Enquanto medida de fiabilidade, a fiabilidade teste-reteste corresponde à proporção da

variância total que se deve a diferenças reais entre sujeitos (Mokkink et al., 2010). No entanto, distingue-se da consistência interna, por ser aplicável a medidas baseadas num único item ou baseadas em vários itens não correlacionados entre si. Esta aplicabilidade é possível uma vez que a fiabilidade teste-reteste assenta em replicações dos mesmos itens em momentos diferentes, e não na resposta num único momento, no qual se considera que os itens são replicações uns dos outros (Polit, 2014). Portanto, na fiabilidade teste-reteste a variação analisada é entre momentos de administração distintos, e não entre diferentes itens. Tendo em conta que a consistência interna de uma medida pode ser inflacionada através do aumento do número de itens de uma medida, o que não acontece na fiabilidade teste-reteste, alguns autores consideram esta propriedade como mais importante do que a consistência interna (e.g. de Vet, Terwee, Mokkink, & Knol. 2011; Polit & Yang, 2014). Ainda que não exista consenso quanto à duração do intervalo entre administrações, é necessário garantir que tenha passado tempo suficiente para que a probabilidade de os sujeitos se recordarem das suas respostas seja baixa, mas não ao ponto de terem ocorrido mudanças no constructo avaliado (Fitzpatrick et al., 1998). Nesse sentido, Streiner e Norman (1995) sugerem um intervalo entre dois e catorze dias, sendo contudo mais importante que exista uma justificação teórica para o intervalo escolhido (Terwee et al., 2007).

A necessidade de demonstrar que um instrumento é consistente ao longo do tempo opõe-se à necessidade de utilizá-lo para medir a mudança, o que torna complexa a avaliação da fiabilidade teste-reteste nas medidas de mudança, do ponto de vista conceptual e prático (Ogles, 2013). Um problema central é a ocorrência de mudança real durante o intervalo entre administrações, e consequente interpretação da mesma como erro de medida, o que pode resultar num coeficiente de fiabilidade teste-reteste reduzido (Polit, 2014). Concomitantemente, pode acontecer que um instrumento estável (i.e., com boa fiabilidade teste-reteste) seja pouco sensível à mudança, ou, por outro lado, que um instrumento sensível à mudança não seja fiável por detetar mudanças resultantes de outras variáveis que não o tratamento ou intervenção (Ogles, 2013). A solução pode passar não só pela mensuração de constructos que são indicadores importantes de mudança, estáveis ao ponto de permitir medições fiáveis, mas também pelo cuidado na escolha do intervalo entre administrações (Ogles, 2013; Polit, 2014).

Apesar de ser geralmente avaliada através do coeficiente de correlação de Pearson (Fitzpatrick et al., 1998), o Coeficiente de Correlação Intraclasse é o indicador estatístico mais adequado para avaliação da fiabilidade teste-reteste em medidas de natureza contínua (Terwee et al., 2007). A utilização do coeficiente de correlação de

Pearson é insuficiente, pois resultados provenientes de duas administrações diferentes podem estar altamente correlacionados mas ser sistematicamente diferentes (Fitzpatrick et al., 1998), e o coeficiente de correlação de Pearson não é sensível a diferenças sistemáticas (Streiner & Norman, 1995; Terwee et al., 2007). Por sua vez, o Coeficiente de Correlação Intraclasse permite determinar que fração da variabilidade dos resultados se deve a diferenças entre sujeitos, e que fração se deve a variabilidade na medição (Fitzpatrick et al., 1998). Ou seja, o Coeficiente de Correlação Intraclasse permite avaliar até que ponto os sujeitos de uma amostra podem ser distinguidos independentemente do erro de medida (Polit, 2014). Um valor de .70 é geralmente considerado como o valor mínimo aceitável para a fiabilidade teste-reteste, devendo este valor ser obtido numa amostra não inferior a 50 pacientes (Terwee et al., 2007). Por sua vez, Nunnally e Bernstein (1994) defendem que a amostra deve ter, pelo menos, entre 200 e 300 sujeitos, uma vez que a confiança numa estimativa de fiabilidade é influenciada pela dimensão da amostra, da qual a estimativa foi obtida (Eliasziw & Donner, 1987), enquanto Streiner e Norman (1995) afirmam que amostras de dimensão inferior a 200 já permitem obter intervalos de confiança com amplitude .10, aproximadamente. Contudo, a forte influência que a dimensão da amostra exerce sobre o valor do Coeficiente de Correlação Intraclasse pode levar a que se alcancem os valores mínimos aceitáveis com amostras suficientemente grandes, sem que esses valores reflitam uma fiabilidade teste-reteste real, pelo que poderá ser mais importante avaliar a fiabilidade teste-reteste em diferentes amostras do que alcançar um valor pré-determinado para esse coeficiente (Fitzpatrick et al., 1998). Portanto, o estudo desta propriedade deve ser levado a cabo em populações diferentes para aumentar a confiança nas estimativas de fiabilidade de um instrumento (Williams & Naylor, 1992).

Erro de medida

A terceira propriedade de medida do domínio da fiabilidade é o erro de medida, que diz respeito à quantidade do *score* de um sujeito que não pode ser atribuída a mudanças efetivas no constructo subjacente (Mokkink et al., 2010). O erro de uma medida pode resultar de uma combinação de processos sistemáticos ou aleatórios (Nunnally & Bernstein, 1994), e é adequadamente avaliado pelo cálculo do erro-padrão da medida (van der Linde et al, 2015). O erro-padrão da medida corresponde ao desvio padrão de medidas repetidas de um só sujeito, e no seu cálculo podem ser consideradas não só a componente aleatória do erro, mas também as diferenças sistemáticas entre medições (Terwee et al., 2007; van der Linde et al., 2015).

Adicionalmente, o erro-padrão da medida tem a vantagem de corresponder a um atributo da medida e não da amostra sob estudo (i.e., não é dependente da amostra) (Crosby, Kolotkin, & Williams, 2003). Uma vez que os valores do erro-padrão da medida são expressos na unidade de medida do instrumento, não é possível definir diretrizes referentes a valores aceitáveis, devendo a interpretação da magnitude do erro de medida ser feita de acordo com a experiência na utilização do instrumento (de Vet, Terwee, Mokkink, & Knol, 2007). Uma das desvantagens do erro-padrão da medida é assumir que o valor do erro é constante ao longo de toda a amplitude da variável que está a ser medida (Crosby et al., 2003), existindo evidência que não confirma esta premissa (Stratford et al., 1996). Não obstante, a importância do erro-padrão da medida estende-se não só à possibilidade de ser utilizado no cálculo de *smallest detectable change* (o valor mínimo de mudança que um sujeito deve apresentar para garantir que a mudança observada está para além do erro de medida) (de Vet et al., 2011; van der Linde et al., 2015), mas também ao cálculo do *reliable change index*, que permite perceber se a magnitude da mudança observada num determinado sujeito é estatisticamente fiável (Jacobson & Truax, 1991).

2.3.2. Validade

A validade de uma medida concerne ao grau com que um instrumento mede os constructos que se propõe medir (Mokkink et al., 2011), e é um domínio fundamental de qualquer instrumento de avaliação por ditar a confiança no significado dos valores obtidos com o instrumento (Ogles, 2013). Não correspondendo a uma propriedade fixa (uma vez que existem várias formas de medi-la), a validade remete para a utilidade científica de um instrumento (Fitzpatrick et al., 1998; Nunnally & Bernstein, 1994). Apesar de não ser raro dizer que determinada medida se encontra validada, a validação é referente à utilização da medida e não à medida em si mesma (Nunnnally & Bernstein, 1994), ou seja, a validade deve ser avaliada em relação a um objetivo e contexto específicos (Fitzpatrick et al., 1998). Mesmo beneficiando de uma definição relativamente simples, a predominância de constructos inobserváveis em Psicologia torna difícil atingir um consenso quanto à forma de avaliar a validade (de Vet et al., 2011). Na verdade, a própria definição de validade, enquanto grau com que um instrumento mede o que se propõe medir, é controversa, já que definições clássicas de validade definem-na também como a correspondência entre as relações empíricas que um instrumento estabelece com outras medidas e as relações teóricas numa rede nomológica (Cronbach & Meehl, 1955). Estas diferentes definições podem resultar em formas diferentes de estudar a validade. Por um lado, o conceito de rede nomológica

implica a avaliação da validade mediante a análise de matrizes de correlação entre os resultados de um teste e outras medidas teoricamente relevantes. Por outro lado, a perspetivação da validade, enquanto capacidade de um instrumento medir o que se propõe medir, subentende a existência de uma relação de causalidade entre o constructo pretensamente medido e os valores de medida obtidos através desse instrumento, passando a ser fundamental estudar os processos que sustentam tal relação (Borsboom, Mellenbergh, & van Heerden, 2004; de Vet et al., 2011). Nesse sentido, Borsboom e colaboradores defendem que, por exemplo, o peso não deve ser implicitamente definido através da sua relação com a altura, o que significa que não faz sentido estudar a validade de uma balança com recurso exclusivo à correlação com valores obtidos pela utilização de uma fita métrica. Passa, assim, a ser de maior importância desenvolver e testar uma teoria que explique a forma através da qual variações no constructo (e.g., o peso) resultem, causalmente, em mudanças no instrumento de medida (e.g., a balança). Não obstante, são raros os exemplos de estudos de validação que seguem esta estratégia (de Vet et al., 2011).

Globalmente, podem ser distinguidos três grandes tipos de validade: a validade de conteúdo (que inclui a *face validity*); a validade de critério (que engloba a validade preditiva); e a validade de constructo (onde se enquadra a validade convergente e discriminante) (de Vet et al., 2011; Fitzpatrick et al., 1998).

Validade de conteúdo

Um primeiro aspeto da validade de conteúdo refere-se ao grau com que, com base no conteúdo manifesto dos itens, uma medida parece refletir o constructo a ser medido, ou seja, a *face validity* (Fitzpatrick et al., 1998; Mokkink et al., 2010; Ware, Brook, Davies & Lohr, 1981). Uma vez que corresponde a uma avaliação subjetiva resultante das primeiras impressões do conteúdo dos itens de uma medida, a avaliação da *face validity* carece de normas, e a impossibilidade na sua quantificação impede a aplicação de procedimentos de análise estatística (de Vet et al., 2011; Fitzpatrick et al., 1998). A validade de conteúdo, enquanto propriedade de medida propriamente dita, concerne ao grau com que o constructo a ser avaliado é abrangido pelo conteúdo do instrumento, ou seja, até que ponto está garantida a adequabilidade do processo de amostragem que levou à seleção dos itens (Guyatt, Feeney, & Patrick, 1993; Nunnally & Bernstein, 1994). De acordo com de Vet e colaboradores (2011), a validação do conteúdo de uma medida para uma determinada população implica cinco passos: a definição clara do constructo que a medida pretende avaliar, a descrição das condições de utilização em relação à população alvo, e a descrição do propósito da

medida; a disponibilização, de forma detalhada, de informação acerca da medida, do seu desenvolvimento, e dos procedimentos de administração; a seleção de um painel de peritos, preferencialmente composto por investigadores e, no caso das medidas de mudança, por pacientes, representantes da população alvo ou utilizadores da medida, para avaliação da correspondência entre o conteúdo da medida e o constructo; o recurso a um quadro de referência teórico para avaliar a correspondência entre o instrumento e o constructo.

Considerando que uma componente importante da validade de conteúdo é referente ao envolvimento dos utilizadores da medida no desenvolvimento e elaboração do seu conteúdo (Fitzpatrick et al., 1998), e uma vez que o conteúdo das MIM é desenvolvido pela pessoa a quem a medida será administrada (Ashworth et al., 2004), há indícios de que a validade de conteúdo destas medidas possa ser considerável (Ashworth et al., 2007). Não obstante, apesar de terem sido realizados alguns estudos sobre as propriedades psicométricas de MIM (e.g. Ashworth et al., 2005; Elliott et al., 2016), a validade de conteúdo não é frequentemente avaliada. Se, por um lado, tal pode resultar de uma desvalorização desta propriedade decorrente da dificuldade em recorrer a modelos estatísticos para a sua avaliação (Feinstein, 1987), por outro lado há que considerar a especificidade das MIM, em particular no que diz respeito à geração de itens idiossincráticos. Concretamente, a multiplicidade de constructos abrangidos pelas MIM dificulta a avaliação da validade de conteúdo enquanto grau com que o conteúdo de uma medida reflete adequadamente o constructo a medir (Mokkink et al., 2010). A contradição evidencia-se quando se considera que Davy e colaboradores (2012) avaliaram a validade de conteúdo do PSYCHLOPS mediante a presença de itens relacionados com problemas de sono, o que indica que seria este o constructo em avaliação. Contudo, o PSYCHLOPS é descrito como sendo uma medida do sofrimento psicológico (Ashworth, 2007). Ou seja, pode relacionar-se a dificuldade em avaliar a validade de conteúdo (enquanto abrangência do constructo pelo conteúdo da medida) com a possibilidade de utilizar MIM para avaliar diferentes constructos.

Validade de critério

A validade de critério define-se como o grau com que os resultados de um instrumento se correlacionam com os resultados de outro instrumento globalmente aceite como mais válido, ou considerado como variável critério (Fitzpatrick et al., 1998; Mokkink et al., 2010). Portanto, a avaliação da validade de critério apenas é possível quando está disponível um critério independente com o qual os resultados do novo

instrumento possam ser comparados (de Vet et al., 2011; Ogles, 2013). Genericamente referente às relações funcionais entre um novo instrumento e um critério, a validade de critério pode ser subdividida de acordo com a relação temporal entre ambos, concretamente, em validade concorrente e em validade preditiva (Nunnally & Bernstein, 1994). Quando o instrumento sob estudo serve propósitos diagnósticos e avaliativos é comum avaliar-se a validade concorrente, considerando os resultados do instrumento e do critério obtidos no mesmo momento (de Vet et al., 2011). Por outro lado, quando se espera que os resultados do novo instrumento se correlacionem com resultados futuros do critério, avalia-se a validade preditiva (Fitzpatrick et al., 1998). Em ambos os casos é recomendável que o nível de concordância requerido entre o novo instrumento e o critério seja definido previamente (para prevenir a utilização de dados pouco convincentes como base para conclusões positivas referentes à validade de critério), e é necessário que os resultados sejam obtidos de forma independente. Por fim, é preciso determinar a força da relação entre os resultados do instrumento e os resultados do critério, com recurso à análise da sensitividade e da especificidade (quando os resultados do instrumento e do critérios são dicotómicos), ao cálculo de *Receiver Operating Characteristic Curves* (caso a escala do instrumento seja de natureza ordinal ou contínua), ou ao cálculo do Coeficiente de Correlação Intraclasse (quando as escalas do instrumento e do critério são de natureza contínua e expressas na mesma unidade) (de Vet et al., 2011).

Uma vez que existe uma escassez de medidas precisas ao ponto de ser possível considerá-las como critério perfeito, o que resulta em parte da preponderância de constructos inobserváveis, há abordagens indiretas cuja utilização para a avaliação da validade de medidas de mudança é preferível à comparação dos resultados de um instrumento com um critério (de Vet et al., 2011; Fitzpatrick et al., 1998; Patrick & Erickson, 1993). No que diz respeito às MIM, a avaliação da validade de critério é dificultada pela premissa de que cada paciente apresenta uma condição clínica e um conjunto de problemas únicos (Elliott et al., 2016), oposta ao pressuposto nomotético de que a maioria dos pacientes pode obter um resultado numericamente comparável a outros pacientes utilizando a mesma medida (Evans et al., 2010). Concretamente, a necessidade de encontrar um instrumento que possa ser considerado como critério apenas é conciliável com a perspetivação da condição do paciente como única, caso seja possível obter como critério outra medida individualizada de mudança com precisão suficiente para que seja possível considerá-la como critério para determinada pessoa. Assim, subentende-se que a avaliação da validade de critério em MIM requer a seleção de um critério específico para cada sujeito. Contudo, esta dificuldade não é particularmente significativa se atendermos às

limitações inerentes à avaliação da validade através do recurso a um critério, tais como a incapacidade de facilitar o desenvolvimento de teoria, a impossibilidade de desenvolver medidas superiores ao critério, ou a ausência de avaliação independente na seleção do critério. Estas limitações potenciaram o desenvolvimento do conceito de validade de constructo, e retiraram importância à validade de critério (Strauss & Smith, 2009). Deve, todavia, sublinhar-se que, no que concerne a medidas de avaliação do estado de saúde, a avaliação da validade de critério é fundamental no desenvolvimento e validação de versões reduzidas de questionários, uma vez que se pretende saber se a versão reduzida é tão válida como a versão completa, tomada como critério (Ware, Kosinski, & Keller, 1996).

Validade de constructo

O desenvolvimento de teorias integrativas referentes à cognição, à personalidade, ou à psicopatologia, pode atribuir-se a conhecimento proveniente de metodologias assentes na validade de critério. Não obstante, as dificuldades inerentes a estas metodologias e a necessidade de desenvolver uma teoria de validação que suportasse os avanços teóricos emergentes contribuíram significativamente para o desenvolvimento do conceito de validade de constructo (Stauss & Smith, 2009). A validade de constructo define-se como o grau com que os resultados de um instrumento são consistentes com hipóteses referentes a relações com os resultados de outros instrumentos, a relações internas ao instrumento, ou a diferenças entre grupos (Mokkink et al., 2010). Esta definição remete, em primeiro lugar, para a articulação de teorias concretas que descrevem as relações entre processos psicológicos, tal como referido por Cronbach e Meehl (1955), o que implica a avaliação do sistema interligado de leis que constituem uma teoria mediante a análise das relações quantitativas entre constructos (Fitzpatrick et al., 1998; Strauss & Smith, 2009). Ou seja, o primeiro passo no estudo da validade de constructo de uma medida assenta numa descrição detalhada do constructo sob estudo, preferencialmente enquadrado num modelo conceptual (de Vet et al., 2011), uma vez que a avaliação da validade de constructo é simultaneamente o teste do significado atribuído aos resultados de um instrumento e o teste da teoria que define o constructo (Messick, 1995; Strauss & Smith, 2009). Em segundo lugar, a avaliação da consistência dos resultados com hipóteses referentes a relações com resultados de outros instrumentos remete também para os conceitos de validade convergente e validade discriminante, propostos por Campbell e Fiske (1959). Enquanto a validade convergente é demonstrada através de associações entre procedimentos de medida independentes,

desenvolvidos para a avaliação de constructos semelhantes, a validade discriminante exige correlações ténues, ou inexistentes, com medidas de constructos conceptualmente distantes, ainda que dentro do mesmo domínio conceptual (Campbell & Fiske, 1959; de Vet et al., 2011; Strauss & Smith, 2009). Em ambos os casos, as hipóteses devem ser detalhadamente estabelecidas antes da recolha de dados (de Vet et al., 2011). Este método corresponde a uma das formas mais sofisticadas de avaliação da validade de constructo, e apelida-se de *Multitrait-Multimethod Matrix* (Campbell & Fiske, 1959; Fitzpatrick et al., 1998). Em terceiro lugar, a consistência dos resultados de um instrumento com hipóteses referentes a relações internas remete para a validade estrutural. Este aspeto da validade de constructo define-se como o grau através do qual os resultados de um instrumento refletem adequadamente a dimensionalidade do constructo em estudo, sendo avaliado mediante análise fatorial confirmatória (quando existem hipóteses prévias quanto às dimensões do constructo), ou exploratória (quando não há hipóteses concretas quanto à dimensionalidade) (de Vet et al., 2011; Mokkink et al., 2010). Por fim, a definição de validade de constructo inclui ainda a possibilidade de colocar hipóteses referentes a diferenças esperadas entre subgrupos específicos de sujeitos, ou seja, a validação discriminativa (de Vet et al., 2011).

A possibilidade de avaliar diferentes constructos ao utilizar MIM torna a avaliação da validade de constructo destes instrumentos particularmente complexa. Uma vez que a validade de constructo é comumente avaliada a partir da análise de relações de um constructo com um conjunto de variáveis (Fitzpatrick et al., 1998), e considerando a necessidade de validar novamente um instrumento quando o contexto ou o propósito da sua utilização mudam (de Vet et al., 2011; Strauss & Smith, 2009), o recurso a MIM para avaliar diferentes constructos exige estudos de validade distintos. A título de exemplo, a análise da validade de constructo da utilização do PSYCHLOPS, enquanto medida do constructo “sofrimento psicológico”, exige uma descrição detalhada do constructo e a formulação de hipóteses testáveis acerca dos resultados expectáveis, tendo em conta uma teoria que explicite as relações com outros constructos, e que seja testada empiricamente (de Vet et al., 2011; Kane, 2001). Consequentemente, a utilização do PSYCHLOPS, como medida do constructo “problemas de sono”, (Davy et al., 2012) exige uma descrição do constructo e da rede nomológica em que este se insere, distinta da que terá sido elaborada para o constructo “sofrimento psicológico”. Adicionalmente, a utilização de uma medida individualizada de mudança para a medição de constructos distintos pode significar que, teoricamente, se possa falar da existência de instrumentos distintos. Assim é porque a exigência de desenvolver um novo estudo de validação de uma medida

utilizada numa situação diferente, ou para um propósito distinto , apenas considera diferentes populações, idiomas, ou formas de avaliação, mas não a utilização da medida para avaliação de outro constructo (de Vet et al., 2011). Esta dificuldade não está presente em medidas de mudança nomotéticas, uma vez que um instrumento como o CORE-OM (Evans et al., 2002), por exemplo, é utilizado como medida dos constructos "sofrimento psicológico" e "risco" (Lyne, Barrett, Evans, & Barkham, 2006; Skre et al., 2013), independentemente da amostra, da linguagem ou do contexto de utilização (e.g. Elfström et al., 2013; Evans et al., 2002; Jenkins & Turner, 2014; Palmieri et al., 2009; Sales, Moleiro, Evans & Alves, 2013; Skre et al., 2013).

A dificuldade em avaliar a validade de constructo de MIM é também evidenciada pela distinção proposta por Fayers, Hand, Bjordal e Groenvold (1997) entre modelos reflexivos e modelos formativos. De uma maneira geral, a construção de escalas psicométricas assenta no pressuposto de que todos os itens de uma escala refletem o constructo latente que a escala pretende medir (Boehmer & Luszcynska, 2006). Ou seja, presume-se que o constructo se manifesta nos itens, de tal forma que são as variações no constructo que causam variações nos resultados dos itens (de Vet et al., 2011; Fayers et al., 1997). Quando, num quadro conceptual, é à variável latente (constructo) que se atribui prioridade causal sobre os seus indicadores (itens), o modelo é reflexivo e os itens da escala correspondem a indicadores de efeito, porque indicam o nível (ou quantidade) do constructo a medir (Bohemer & Luszcynska, 2006; Bollen & Ting, 2000; de Vet et al., 2011; Edwards & Bagozzi, 2000). Contudo, existe a possibilidade de que a direção da causalidade seja invertida, de tal forma que são os indicadores (os itens) que formam a variável latente (MacCallum & Browne, 1993), isto é, são as mudanças nos indicadores que determinam a variação no valor da variável latente, e não o inverso (Diamantopoulos & Siguaw, 2006). O constructo *stress* corresponde a um exemplo de um modelo formativo, já que serão as mudanças em itens referentes a aspetos como a perda de emprego, a morte de um familiar, ou a um divórcio, que resultarão em mudanças no constructo, e não o inverso (i.e., não é o aumento do *stress* que faz com que a resposta aos itens mencionados se altere) (de Vet et al., 2011). Este modelo é um modelo formativo uma vez que os itens formam o constructo, sendo, por isso, indicadores causais e não indicadores de efeito (Edwards & Bagozzi, 2000; Fayers et al., 1997). Ainda que a relação entre os indicadores e o constructo possa ser mais complexa, na medida em que um indicador de efeito pode ser, em parte, um indicador causal (Fayers et al., 1997), os pressupostos da teoria psicométrica clássica não se aplicam na totalidade a modelos formativos (Bollen & Lennox, 1991), o que implica que os procedimentos de seleção dos itens e de análise

das propriedades psicométricas sejam diferentes nos dois tipos de modelo (Diamantopoulos & Siguaw, 2006).

No que diz respeito à seleção dos itens, a ênfase que a abordagem de desenvolvimento de escalas tradicionais coloca na correlação positiva entre os itens, na variância comum, na unidimensionalidade e na consistência interna significa que esta é a abordagem mais adequada quando o modelo é reflexivo. Contudo, tendo em conta que os indicadores causais podem apresentar correlações negativas, ou inexistentes, entre si, uma abordagem que parte da premissa de que os indicadores de um constructo se encontram positivamente correlacionados entre si não parece adequada quando o modelo é formativo (onde itens representam indicadores causais). Assim, será mais adequado recorrer a estratégias de construção de índices do que a estratégias baseadas no desenvolvimento de escalas tradicionais (Diamantopoulos & Winklhofer, 2001; Diamantopoulos & Siguaw, 2006; Bollen & Lennox, 1991). Apesar de ser expectável que os itens gerados pelas duas abordagens não sejam significativamente divergentes, as diferenças entre a abordagem de desenvolvimento de escalas e a estratégia de construção de índices poderão resultar em conjuntos finais de itens distintos: na abordagem de construção de escalas a seleção dos itens enfatizará as correlações entre itens, para aumento da consistência interna; na construção de índices, a influência distinta de cada um dos itens é valorizada uma vez que não são esperadas correlações elevadas entre indicadores causais (Bollen, 1989; Bollen & Lennox, 1991; Diamantopoulos & Siguaw, 2006; Fayers & Hand, 1997).

A diferença no sentido da causalidade em modelos reflexivos e modelos formativos tem consequências na avaliação das propriedades psicométricas, em particular no que diz respeito à avaliação da validade estrutural (Fayers et al., 1997). Concretamente, a utilização de análise fatorial exploratória para o estudo da dimensionalidade do constructo não se adequa a indicadores causais, uma vez nesta análise se assume que os fatores são compostos por indicadores de efeito, o que não acontece em modelos formativos (Fayers & Hand, 1997). A extração de fatores a partir de indicadores causais não poderá resultar num constructo unidimensional e significativo do ponto de vista conceptual, não devendo por isso ser levada a cabo, pois a análise fatorial exploratória não permite desenvolver modelos a partir de indicadores causais (Bohemer & Luszczynska, 2006; Fayers & Hand, 1997). A determinação da variável latente pelos indicadores causais também não deverá ser realizada a partir da análise de componentes principais uma vez que o modelo de medida com indicadores causais tem apenas uma variável latente para *q* indicadores, e não *q* componentes latentes (Bollen & Lennox, 1991).

Uma vez que modelos reflexivos e formativos de medida significam diferenças tanto ao nível do conteúdo da medida como ao nível do estudo das suas propriedades psicométricas, é fundamental que a escolha do modelo métrico se baseie em considerações teóricas referentes à direção causal entre o constructo e os seus indicadores (Diamantopoulos & Siguaw, 2006; Jarvis, Mackenzie, & Podsakoff, 2003). Neste sentido, a avaliação da validade de constructo, em particular da validade estrutural, em MIM vai ser dificultada, uma vez que o processo de geração dos itens idiossincráticos não contempla a distinção entre indicadores de efeito e indicadores causais, o que pode resultar em medidas compostas pelos dois tipos de indicadores. A razoabilidade desta hipótese é suportada pelas categorias temáticas identificadas a partir da análise das respostas de pacientes ao domínio "problemas" do PSYCHLOPS, onde há simultaneamente categorias formadas por indicadores causais e categorias formadas por indicadores de efeito do constructo "sofrimento psicológico" que o instrumento se propõe medir (Ashworth et al., 2004; Robinson et al., 2006). Assim, não é possível avaliar a dimensionalidade do constructo mediante a realização de análise fatorial exploratória, podendo também estar comprometida a avaliação da consistência interna uma vez que as correlações entre indicadores causais não são necessariamente positivas (Diamantopoulos, Riefler, & Roth, 2008; Bollen & Lennox, 1991).

2.3.3. Sensibilidade para detetar mudança

Para além de cumprirem requisitos referentes à fiabilidade e à validade, as medidas de mudança devem permitir a identificação de mudanças, ou seja, é necessário que um instrumento seja sensível à mudança (Guyatt, Kirshner, & Jaeschke, 1992; Ogles, 2013). A sensibilidade para detetar mudança, ou *responsiveness*, define-se como a capacidade de um instrumento em detetar a mudança ao longo do tempo no constructo sob estudo, havendo contudo uma grande diversidade de definições desta propriedade (Mokkink et al., 2010; Terwee et al., 2007). As definições existentes diferem no tipo de mudança que deve ser detetado através da administração de uma medida como, por exemplo, se são mudanças clinicamente importantes, mudanças devidas aos efeitos do tratamento, mudanças no valor real do constructo subjacente, ou, simplesmente, mudanças (de Vet et al., 2011; Terwee, Dekker, Wiersinga, Prummel, & Bossuyt, 2003). Quando se tem em conta que a utilização de uma medida de mudança pretende detetar a ocorrência de uma mudança real, e não de ruído ou de mudança num constructo diferente do pretendido, a deteção de uma mudança geral, definida enquanto mudança estatisticamente

significativa, não é suficiente (de Vet et al., 2011). Desta forma, e uma vez que a avaliação da sensibilidade para detetar mudança implica o teste da hipótese de que mudanças no constructo avaliado resultam em mudanças correspondentes na medida, é preciso garantir que as mudanças que se pretendem avaliar se refiram ao constructo sob estudo (Terwee, 2014). Adicionalmente, a avaliação de uma mudança clinicamente importante implica uma definição daquilo que é clinicamente importante, o que é referente à interpretação do resultado de mudança, e não à validade do mesmo (de Vet et al., 2011). Considerando a proximidade da definição apresentada com o conceito de validade, alguns autores defendem incluir a sensibilidade para detetar mudança dentro do domínio métrico da validade (Ogles, 2013). Contudo, a importância da sensibilidade para detetar mudança, as diferenças em relação às análises de avaliação da validade, e a distinção entre a validade de um resultado único e de um resultado de mudança justificam a manutenção da distinção (de Vet et al., 2011).

A variedade de definições do conceito de sensibilidade para detetar mudança implica que existam, também, vários métodos de calcular esta propriedade (Terwee et al., 2003). Alguns autores (e.g. Husted, Cook, Farewell, & Gladman, 2000) distinguem entre dois tipos de sensibilidade para detetar mudança – a sensibilidade à mudança interna e a sensibilidade à mudança externa –, os quais justificam a utilização de alguns métodos em detrimento de outros. A *sensibilidade para detetar mudança interna* define-se como a capacidade de uma medida para detetar mudança num determinado intervalo de tempo, e é referente à possibilidade de detetar qualquer tipo de mudança estatisticamente fiável (Prous, Salvanés, & Ortells, 2008). Os indicadores estatísticos mais frequentemente utilizados para avaliar a *sensibilidade para detetar mudança interna* são o teste *t* para medidas emparelhadas (através do qual é testada a hipótese de que, em média, não existe mudança nas medidas em dois momentos distintos), e o cálculo da magnitude do efeito (Husted et al., 2000). Ainda que seja relativamente consensual considerar insuficiente a utilização de testes *t* emparelhados (por não ser um teste da validade dos *scores* de mudança, mas da significância estatística dessa mudança, a qual é dependente da dimensão da amostra e da variabilidade da medida), há algum debate acerca da adequabilidade dos índices de magnitude do efeito na avaliação da sensibilidade para detetar mudança (de Vet et al., 2011; Husted et al., 2000). Por um lado, as medidas padronizadas de magnitude do efeito permitem medir a relação entre a magnitude da mudança e a variabilidade dos resultados, independentemente da dimensão da amostra (no caso da *standardized response mean*) (Eisen, Ranganathan, Seal, & Spiro, 2007; Prous et al., 2008), superando assim as limitações inerentes ao uso do teste *t*. Por outro lado, a magnitude

do efeito avalia o efeito de um determinado tratamento, mas não a sensibilidade para detetar mudança do instrumento que está a medir esse efeito. Além disso, as medidas de magnitude do efeito são dependentes do desvio-padrão, tomando valores mais altos em populações relativamente homogéneas ou quando existe pouca variabilidade no efeito do tratamento (no caso da *standardized response mean*) (Becker, 2000; de Vet et al., 2011; Husted et al., 2007; Prous et al., 2008). Adicionalmente, a ocorrência de um efeito de teto, ou a escassez de itens relevantes, pode resultar em estimativas da magnitude do efeito que subestimam a mudança real, ficando a sensibilidade para detetar mudança dependente tanto do efeito do tratamento como da medida (de Vet et al., 2011; Prous et al., 2008).

A *sensibilidade para detetar mudança externa* define-se como o grau com que as mudanças numa medida estão correlacionadas com mudanças correspondentes noutra medida (Prous et al., 2008). Contrariamente à *sensibilidade interna,* a *sensibilidade externa* depende exclusivamente da escolha do instrumento para comparação, e não do tratamento, motivo pelo qual este tipo de sensibilidade corresponde a uma propriedade da medida sob estudo (Husted et al., 2000; Prous et al., 2008). Esta perspetiva coaduna-se com a conceptualização da sensibilidade para detetar mudança dentro do domínio da validade, portanto, distinta da abordagem experimental que relaciona a sensibilidade para detetar mudança com a magnitude do efeito de um tratamento (Terwee et al., 2003). Neste caso, a sensibilidade para detetar mudança será avaliada à semelhança do que ocorre na avaliação da validade, de acordo com uma abordagem de critério, ou de acordo com uma abordagem de constructo (de Vet et al., 2011).

A abordagem de critério, adequada quando é possível aceder a uma medida de mudança critério, permite avaliar o grau com o qual as mudanças numa medida refletem adequadamente mudanças na medida critério. Ainda que os passos necessários para levar a cabo este tipo de avaliação da sensibilidade para detetar mudança sejam semelhantes aos passos necessários para a avaliação da validade de critério, a dificuldade em encontrar uma medida que possa ser utilizada como critério pode não ser tão relevante na avaliação da sensibilidade. Tal pode dever-se ao recurso a uma *escala de avaliação global*, que consiste numa única questão referente à mudança no constructo de interesse ao longo do tratamento. Contudo, ainda que esta escala beneficie de elevada *face validity*, não é claro que a sua fiabilidade esteja garantida, ou que seja possível assegurar que o constructo medido seja o mesmo que a medida sob estudo avalia, ficando assim a abordagem de critério principalmente adequada à avaliação de versões reduzidas de uma medida (de Vet et al., 2011; Norman, Stratford, & Rehegr, 1997). Os testes utilizados para avaliação da

sensibilidade para detetar mudança de acordo com a abordagem de critério dependem da natureza da escala utilizada na medida critério, sendo adequado recorrer a correlações quando se trata de variáveis contínuas, e a *receiver operating characteristic curves* quando a medida critério é dicotómica (de Vet et al., 2011).

A inexistência de uma medida critério implica que a avaliação da sensibilidade para detetar mudança recorra a uma abordagem de constructo, dependente do teste de hipóteses referentes a diferenças esperadas entre os resultados de mudança obtidos em grupos diferentes, ou a correlações esperadas entre os resultados de mudança do instrumento sob estudo e outros instrumentos que avaliam constructos semelhantes. Nesse sentido, é fundamental formular hipóteses concretas acerca da direção esperada das correlações e acerca da magnitude das mesmas, antes do início da recolha de dados, para que sejam evitados enviesamentos na análise de dados. Adicionalmente, importa considerar que a significância estatística, ou o *p value,* são menos importantes e informativos do que a verificação das hipóteses referentes à direção e magnitude das correlações observadas (de Vet et al., 2011).

3. Objetivos e hipóteses

Contribuir para um maior conhecimento acerca das características psicométricas do PSYCHLOPS é o principal objetivo deste estudo. Adicionalmente, pretende-se analisar a aplicabilidade e adequabilidade das propriedades psicométricas clássicas na avaliação de MIM, mediante revisão da literatura. Nesse sentido, proceder-se-á a um estudo comparativo da validade, da fiabilidade, e da sensibilidade para detetar mudança, com recurso a seis amostras, provenientes de diferentes países europeus. O recurso a várias amostras prende-se com o objetivo de realizar um estudo abrangente e comparativo das propriedades psicométricas do PSYCHLOPS. A análise destas propriedades será efetuada separadamente para cada amostra, sendo então estimados os efeitos globais com recurso a uma abordagem meta-analítica.

A partir da revisão da literatura e dos objetivos supramencionados, é possível estabelecer as hipóteses a seguir enumeradas, referentes às propriedades psicométricas do PSYCHLOPS:

1. Relativamente à consistência interna da medida compósita proporcionada pelo PSYCHLOPS (avaliada mediante o coeficiente alfa de Cronbach), espera-se que o valor global estimado, a partir das amostras, seja bom ($.8 \leq \alpha \leq .9$), ainda que inferior ao valor global esperado para o CORE-OM. Para sustentar esta hipótese, foram tidos em conta estudos anteriores realizados com o PSYCHLOPS (Ashworth et al., 2005; Ashworth et al., 2009; Héðinsson et al., 2012), e com o CORE-OM (Evans et al., 2002).

2. No que diz respeito à validade de constructo, espera-se que o valor médio da correlação entre os resultados do PSYCHLOPS e os resultados do CORE-OM sejam moderadas ($.40 < r < .70$). Esta hipótese tem em conta estudos anteriores (Ashworth et al., 2005; Ashworth et al., 2009; Héðinsson et al., 2012), bem como o facto de o CORE-OM avaliar um constructo semelhante ao avaliado pelo PSYCHLOPS.

   2.1. A respeito da validade convergente, espera-se que as correlações entre domínios correspondentes do PSYCHLOPS e do CORE-OM sejam de moderadas a fortes ($r > .40$).

   2.2. Em relação à validade discriminante, é esperado que as correlações entre as subescalas do PSYCHLOPS e a subescala Risco do

CORE-OM sejam fracas (r < .40), uma vez que esta subescala avalia uma dimensão do sofrimento psicológico que não é contemplada no PSYCHLOPS.

2.3. Relativamente à validade estrutural, espera-se que as correlações entre os dois itens da subescala Problemas do PSYCHLOPS sejam fortes (r > .70), enquanto as correlações com os itens restantes serão moderadas (.40 < r < .70). A avaliação da validade estrutural restringiu-se ao teste exclusivo desta hipótese uma vez que as restantes escalas do PSYCHLOPS são compostas por um único item, pelo que não é possível formular hipóteses semelhantes referentes a outros domínios. Adicionalmente, a ausência de teoria referente à estrutura do PSYCHLOPS não permite testar hipóteses quanto a correlações entre domínios.

3. No que concerne à sensibilidade para detetar mudança, espera-se que o valor global da correlação entre os *change scores* do PSYCHLOPS e do CORE-OM seja moderada a forte (r > .40), tendo em conta que os constructos avaliados pelos dois instrumentos são semelhantes. Não existem, contudo, estudos anteriores que avaliem a sensibilidade para detetar mudança enquanto próxima do conceito de validade, pelo que a hipótese colocada não é suportada por estudos anteriores.

4. Método

4.1. Participantes

O presente estudo inclui seis amostras distintas, provenientes de quatro países: Portugal (n = 80 e n = 94), Islândia (n = 225), Polónia (n = 208), e Reino Unido (n = 336 e n = 110). As amostras são compostas por grupos clínicos distintos, que incluem utilizadores de cuidados de saúde mental primária e utilizadores de serviços hospitalares de psiquiatria e saúde mental. As seis amostras incluídas neste estudo perfazem um total de 1053 participantes, cujas distribuições referentes à idade e ao sexo podem ser consultados na Tabela 1. Excetuando as amostras portuguesas, foram publicados artigos que utilizaram os dados das restantes amostras, tendo as mesmas sido disponibilizadas pelos autores e incluídas no presente estudo com as devidas autorizações. A referência de cada um dos artigos será incluída na descrição das amostras.

Tabela 1: Distribuição dos participantes por sexo e por idade para as seis amostras em estudo.

Amostra	**Sexo**			**Idade**		**Total**
	Masculino	**Feminino**	***Missing***	**Média (Desvio-padrão)**	***Missing***	
Amostra 1 (POR)	19 (23,8%)	61 (75,3%)	0	44,1 (14,78)	1	80
Amostra 2 (POR)	55 (59,1%)	38 (40,9%)	0	42,9 (11,03)	4	94
Amostra 3 (ISL)	55 (24,4%)	169 (75,1%)	1	38,6 (12,21)	2	225
Amostra 4 (POL)	44 (21,2%)	153 (73,6%)	11	41,1 (11,86)	11	208
Amostra 5 (RU)	30 (27,3%)	75 (68,2%)	5	39,6 (11,97)	7	110
Amostra 6 (RU)	125 (37,1%)	199 (59,1%)	13	37,6 (11,82)	3	336
Total	328	695	30	39,9 (12,79)	28	1053

Amostra 1

A recolha de dados referentes a pacientes adultos do Departamento de Psiquiatria e Saúde Mental do Hospital do Espírito Santo de Évora permitiu a composição de uma amostra de 80 sujeitos. Recolhida entre outubro de 2013 e janeiro de 2015, a amostra refere-se a dados pré-terapia de pacientes que marcaram uma primeira consulta em Psicologia. A maioria (65%) afirmou não ter um diagnóstico psicopatológico, sendo "Depressão" o diagnóstico mais frequentemente reportado (26,5%) e o único reportado por mais do que um participante. A maioria dos participantes encontrava-se a tomar medicação para ajudar o seu bem-estar psicológico (76,3%). A idade dos participantes variava entre os 17 e os 85 anos, com uma média de idades de 44,1 anos. Dos 80 participantes, 61 eram do sexo feminino (75,3%) e 19 eram do sexo masculino (23,5%). A situação profissional mais frequentemente indicada foi a de trabalhador a tempo inteiro (42,5%), seguida do estatuto de desempregado (22,5%) e de reformado (18,8%). A maior parte dos participantes eram casados (43,8%), ou solteiros (23,8%). O nível de escolaridade mais frequente situava-se no intervalo entre o 7.º e o 9.º anos (27,5%), seguido do intervalo entre o 10.º e o 12.º anos (23,8%). A recolha de dados foi levada a cabo por oito assistentes de investigação, após o envio de um pedido de participação no estudo, anexo da convocatória para a primeira consulta em Psicologia.

Amostra 2

A amostra é composta por 94 pacientes que se encontravam prestes a iniciar tratamento para problemas relacionados com dependência e abuso de substâncias, tanto em ambulatório (Centro de Respostas Integradas de Évora, Centro das Taipas, Unidade de Alcoologia de Lisboa), como no contexto de comunidade terapêutica (Comunidade de Inserção Social de Esposende). A recolha de dados foi conduzida por um estudante de doutoramento, quatro estudantes de mestrado e um terapeuta, e levada a cabo imediatamente antes da primeira consulta, tendo a ordem de aplicação dos instrumentos sido aleatória. A idade média dos participantes era de 42,9 anos, e 59,1% eram do sexo masculino.

Amostra 3

Os dados recolhidos são referentes a 225 pacientes do Serviço de Saúde Mental do Hospital Universitário Nacional da Islândia, *Landspitali*. Todos os

participantes se encontravam prestes a iniciar terapia cognitivo-comportamental de grupo, dividida em duas vertentes: um grupo com a duração de 9 semanas, dirigido a pessoas com baixa autoestima; um grupo para pessoas com perturbações do humor e/ou de ansiedade, com a duração de 6 semanas. Em ambos os casos os participantes integravam sessões de terapia semanais com a duração de duas horas. Dos 225 participantes, 169 (75,1%) eram do sexo feminino, e 55 (24,9%) eram do sexo masculino, não havendo informação quanto ao sexo de um/a dos/as participantes. A idade dos participantes oscilava entre os 21 e os 70 anos, sendo a média de 38,6 anos. A amostra foi utilizada num estudo de validação e replicação do PSYCHLOPS por Héðinsson e colaboradores (2012), no qual se procurou averiguar se os resultados numa amostra islandesa replicavam os resultados obtidos com a versão inglesa.

Amostra 4

A amostra é proveniente de um contexto de saúde primária da Polónia, tendo a recolha de dados sido levada a cabo no contexto de um estudo longitudinal acerca da sensibilidade para detetar mudança do PSYCHLOPS. Os dados foram recolhidos por 35 médicos de clínica geral com formação pós-graduada em Terapia Cognitivo-Comportamental. Este procedimento resulta do funcionamento dos serviços de saúde polacos, onde as *talking therapies* são geralmente conduzidas por estes profissionais. Assim, os dados foram recolhidos pelos profissionais que conduziam uma variante de curta duração da Terapia Cognitivo-Comportamental, composta por três ou quatro sessões com a duração aproximada de 30 minutos, tendo os pacientes sido convidados a participar no estudo, após referenciação pelo seu médico para este tipo de terapia. No estudo, foram incluídos pacientes de idades compreendidas entre os 18 e os 65 anos, com sintomas psicossomáticos, de depressão, ou de ansiedade, não sendo elegíveis pacientes com história atual de perturbação psicótica, abuso de substâncias, ou com uma doença orgânica comprometedora do funcionamento cognitivo. Dos 208 participantes, 153 eram do sexo feminino (73,6%), e 44 do sexo masculino (21,2%), não havendo informação quanto ao sexo de 11 participantes. No que respeita à idade, a média é de 41,1 anos, oscilando entre os 19 e os 64 anos. Os dados recolhidos foram incluídos num estudo de avaliação da sensibilidade para detetar mudança, e de avaliação da versão intermédia do PSYCHLOPS, da autoria de Czachowski e colaboradores (2011).

Amostra 5

A administração do PSYCHLOPS do CORE-OM foi levada a cabo em contextos de cuidados de saúde primária no Reino Unido, perfazendo um total de 110 participantes. Os dados foram recolhidos durante a avaliação inicial, anterior ao início de psicoterapia e na última sessão. A amostra foi recolhida no contexto de um estudo das propriedades psicométricas do PSYCHLOPS. No total, 11 participantes tinham entre 15 e 24 anos (11%), 31 participantes estavam entre os 25 e os 34 anos (28,2%), 31 participantes encontravam-se entre os 35 e os 44 anos (28,2%), 16 participantes tinham entre 45 e 54 anos (14,5%), 14 participantes estavam entre os 55 e os 64 anos (12,7%), não existindo informação quanto à idade de sete dos participantes (6,4%). Quanto ao sexo, 75 (68,2%) eram do sexo masculino, e 30 (27,3%) eram do sexo feminino, não havendo informação quanto ao sexo de cinco participantes (4,5%). Os dados foram recolhidos no contexto de um estudo acerca das propriedades psicométricas do PSYCHLOPS, da autoria de Ashworth e colaboradores (2005).

Amostra 6

Os dados desta amostra foram recolhidos por cinco psicólogos clínicos a trabalhar em cuidados de saúde primários no Reino Unido, que administraram seis sessões de terapia de acordo com um modelo Cognitivo-Comportamental, no seguimento de referenciações feitas por médicos de clínica geral e familiar. A amostra totaliza 336 participantes, não tendo cinco deles (1,8%) completado os questionários pré-terapia. Menos de metade (41,7%) dos participantes tinha completado a terapia, sendo apenas nesses casos que existem dados pós-terapia. Dos 329 participantes que responderam, pelo menos, a um dos questionários pré-terapia, a média de idades é de 37,7 anos, não tendo dois participantes fornecido informação quanto à sua idade. Relativamente ao sexo, 124 (37,5%) eram do sexo masculino e 195 (58,9%) do sexo feminino, não existindo informação quanto ao sexo de 12 (3,6%) dos participantes. A amostra foi utilizada num estudo comparativo da sensibilidade para detetar mudança e da fiabilidade do PSYCHLOPS e do HADS de Ashworth e colaboradores (2009).

4.1.1. Distribuições da severidade do sofrimento psicológico

As distribuições referentes à severidade do sofrimento psicológico, medido pelo PSYCHLOPS, antes do início do tratamento, podem ser consultadas na Tabela 2. A média global da severidade é de 14,30, com desvio padrão de 4,08, sendo importante

notar que não foi possível obter o valor da medida compósita do PSYCHLOPS em 112 participantes (10,6%). Esta impossibilidade deveu-se se tanto à ausência de resposta nos dois primeiros itens, ou no terceiro e/ou quarto itens (8,6%), como à não administração do instrumento, por motivos desconhecidos (2%). Cerca de um décimo dos participantes identificaram apenas um problema (9,8%), tendo sido necessário, nesses casos, proceder à duplicação da cotação atribuída ao único problema reportado, para se poder calcular a medida total de severidade. O resultado médio mais elevado é referente à Amostra 4 (15,4) e o mais baixo pode ser encontrado na Amostra 5 (13,0).

Tabela 2: Severidade média do PSYCHLOPS pré-tratamento.

Amostra	*n*	*Missing*	Média	Desvio Padrão	Amplitude
Amostra 1 (POR)	80	13	15,34	4,33	5-20
Amostra 2 (POR)	94	36	13,78	4,86	0-20
Amostra 3 (ISL)	225	15	14,35	4,07	2-20
Amostra 4 (POL)	208	7	15,45	3,08	8-20
Amostra 5 (RU)	110	3	13,00	3,85	1-20
Amostra 6 (RU)	336	38	13,80	4,31	0-20
Total	1053	112	14,30	4,08	0-20

4.2. Instrumentos

Psychological Outcome Profiles (PSYCHLOPS; Ashworth et al., 2004)

Como referido anteriormente, o PSYCHLOPS é uma medida individualizada de mudança, de autopreenchimento que avalia três domínios: Problemas (duas questões), Funcionamento (uma questão) e Bem-estar (uma questão). A administração do instrumento permite obter uma medida global do sofrimento psicológico, servindo os resultados de cada domínio apenas para informação clínica, devido ao número reduzido de itens. Existindo versões distintas do PSYCHLOPS, neste estudo foram utilizadas as versões pré e pós-terapia.

Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM; Evans et al., 2000)

O CORE-OM é um questionário nomotético de autorrelato composto por 34 itens que permite obter um resultado referente ao sofrimento psicológico do paciente (Evans et al., 2000). Desenvolvido no Reino Unido para dar resposta à falta de uma medida de avaliação do progresso do paciente enquadrável no estudo da eficácia clínica, o CORE-OM beneficiou do envolvimento de prestadores de cuidados na identificação de aspetos importantes na construção de um questionário de avaliação do *outcome* (Barkham et al., 2001). Ao mesmo tempo, a revisão de medidas semelhantes pré-existentes e a consideração das opiniões de prestadores de cuidados serviram o propósito de desenvolver um instrumento curto, de cotação e preenchimento simples, e abrangente no que concerne a sintomas, problemas, e dificuldades no funcionamento apresentadas pelos pacientes (Barkam et al., 2001, Mellor-Clark et al., 1999).

Todos os itens do CORE-OM são cotados numa de escala ordinal de 5 níveis (desde 0=Nunca até 4=Sempre ou quase sempre), estando em oito desses itens a escala invertida (i.e. 0=Sempre ou quase sempre, 4=Nunca) (Evans et al., 2002; Sales, Moleiro, Evans, & Alves, 2012). Os itens estão divididos em quatro domínios diferentes, sendo possível obter uma classificação independente para cada um deles: Problemas (12 itens); Funcionamento (12 itens); Bem-estar Subjetivo (4 itens); Risco (6 itens) (Barkham et al., 2001). Existe uma versão portuguesa do instrumento, validada numa amostra não clínica de 111 participantes, tendo os valores obtidos em relação à consistência interna sido excelentes e semelhantes aos encontrados no estudo original (α = .94) (Sales et al., 2012). O estudo das propriedades psicométricas do CORE-OM levado a cabo no Reino Unido permitiu concluir que o instrumento tem boa consistência interna e fiabilidade teste-reteste (.75 e .95 respetivamente), e boa validade convergente, tendo os autores concluído haver evidências de sensibilidade para detetar mudança e de validade (Evans et al., 2002).

4.3. Procedimento

4.3.1. Procedimento de análise de dados

Os dados foram analisados com recurso à aplicação IBM® SPSS® Statistics (versão 22), ao programa Microsoft Office Excel®, e ao programa de meta-análise Comprehensive Meta-Analysis® (versão 2). Após correção de erros de digitação,

procedeu-se à exclusão de todos os participantes que não responderam ao PSYCHLOPS, ou que deixaram em branco os itens referentes ao domínio Funcionamento e/ou Bem-estar, uma vez que nestes casos não é possível obter o resultado global deste instrumento. Adicionalmente, e de acordo com as recomendações dos autores do PSYCHLOPS, nos casos em que apenas foi identificado um problema pelo participante, o cálculo do resultado global fez-se através da duplicação da cotação atribuída a esse problema (ou seja, foi atribuído ao segundo problema o valor selecionado pelo participante na escala de severidade do primeiro problema).

Procedeu-se à análise descritiva das amostras, em particular no que concerne às características sociodemográficas dos participantes. A agregação dos dados sociodemográficos provenientes das diferentes amostras foi levada a cabo mediante o cálculo da média ponderada destas variáveis.

O tratamento de *outliers* foi feito com recurso ao procedimento *winsorization*, que consiste na aproximação de observações extremas a observações mais centrais, pré-determinadas, com o propósito de evitar a sua eliminação ao estimar os parâmetros de interesse (Hawkins, 1980). Neste procedimento, os valores extremos são substituídos pelo valor da observação seguinte (ou anterior), ou seja, pelo valor mais elevado (ou mais reduzido), sem contar com a observação que se pretende substituir (Dixon, 1960). No presente estudo, o procedimento *winsorization* incidiu sobre as observações acima do percentil 98 e abaixo do percentil 2 (assumindo um modelo distribucional gaussiano), sendo a manutenção da simetria fundamental para garantir que os parâmetros estimados se encontrem tão próximos de estimações lineares quanto possível (Dixon, 1960). A identificação de *outliers* multivariados recorreu ao cálculo das *distâncias de Mahalanobis*, que se definem como o quadrado da distância padronizada entre duas observações distintas (Holgersson & Karlsson, 2012). Foi encontrado um *outlier* multivariado na amostra polaca, que foi excluído das análises posteriores.

A avaliação das propriedades psicométricas do PSYCHLOPS foi feita através da análise da fiabilidade, da validade de constructo, e da sensibilidade para detetar mudança separadamente em cada amostra. Com essa finalidade, procedeu-se ao cálculo das pontuações obtidas por cada participante nas subescalas do CORE-OM e do PSYCHLOPS, bem como ao cálculo dos *scores* de mudança de cada um dos instrumentos.

A fiabilidade foi avaliada com recurso ao cálculo do coeficiente alfa (Cronbach, 1951), que permite estimar a correlação média entre as respostas aos itens de um instrumento (Mokkink et al., 2010), avaliando assim uma das dimensões da fiabilidade,

a consistência interna (Almeida & Freire, 2003). A avaliação da fiabilidade teste-reteste não foi possível devido à ausência de medições repetidas no desenho dos estudos de onde as amostras provêm.

A validade de constructo, definida como o grau com o qual os resultados de um instrumento são consistentes com hipóteses referentes a correlações com outros instrumentos (Mokkink et al., 2010), foi avaliada mediante o cálculo das correlações entre os resultados brutos obtidos no PSYCHLOPS e no CORE-OM. A interpretação dos resultados teve em conta as hipóteses estabelecidas acerca do valor das correlações esperadas entre os instrumentos. As características do PSYCHLOPS (nomeadamente, a possibilidade da existência de indicadores causais, e o número reduzido de itens) impediram a avaliação da validade estrutural (uma propriedade do domínio da validade de constructo) mediante análise fatorial. Não obstante, procedeu-se à análise do valor das correlações entre os itens do PSYCHLOPS enquanto forma de avaliar uma parte da validade estrutural, testando-se a hipótese de que a correlação entre os dois itens pertencentes à subescala Problemas fosse maior do que as correlações com os restantes itens.

A validade convergente e discriminante (Campbell & Fiske, 1959) foi avaliada como complemento aos métodos de avaliação da validade de constructo supramencionados. Concretamente, procedeu-se ao teste de hipóteses referentes ao valor das correlações entre as subescalas do PSYCHLOPS e do CORE-OM, sendo globalmente esperadas correlações moderadas a fortes entre subescalas semelhantes dos dois instrumentos, e fracas entre subescalas conceptualmente distantes. Nesse sentido, a subescala Risco do CORE-OM serviu como base para a avaliação da validade discriminante, uma vez que é a única subescala do CORE-OM que não tem correspondência no PSYCHLOPS, o que permite antever correlações mais fracas.

No presente estudo a sensibilidade para detetar mudança foi avaliada mediante a avaliação de hipóteses referentes a correlações esperadas entre os *scores* de mudança do PSYCHLOPS e do CORE-OM.

A agregação dos dados provenientes das seis amostras foi feita de acordo com uma abordagem meta-analítica segundo um *modelo dos efeitos aleatórios*, uma vez que se assume que os parâmetros das populações de onde as diferentes amostras foram extraídas são semelhantes mas não são idênticos (Borenstein, Hedges, Higgins, & Rothstein, 2009). Este modelo permite considerar, para além da variabilidade devida ao erro de amostragem, a existência de variabilidade real entre os estudos (Borenstein et al., 2009; Hartung, Knapp, & Sinha, 2008). Embora a adequabilidade do modelo dos efeitos aleatórios face ao modelo de efeitos fixos possa ser avaliada mediante testes de homogeneidade, o cálculo da potência disponível para realizar tais testes indicou a

existência de uma baixa probabilidade de obtenção de resultados estatisticamente significativos. Por esse motivo, foram considerações teóricas que levaram à adoção do *modelo dos efeitos aleatórios* e não os resultados dos testes de homogeneidade (Hedges & Pigott, 2001; Hedges & Pigott, 2004).

Os procedimentos quantitativos utilizados para resumir os resultados das diferentes amostras (i.e., os procedimentos de meta-análise) baseiam-se nos métodos descritos por Rodriguez e Maeda (2006), e por Hedges e Vevea (1998). Relativamente ao coeficiente alfa, procedeu-se primeiro à conversão para o coeficiente de alfa transformado (Rodriguez & Maeda, 2006) e ao cálculo da sua variância (Hakstian & Whalen, 1976). Para estimar o efeito médio, foi feita uma ponderação dos valores de alfa, de cada amostra, a partir do inverso da variância, tendo-se seguido os procedimentos descritos por Hedges e Vevea (1998). O valor da componente da variância entre estudos foi obtido com recurso ao procedimento *method of moments* para estimação dos parâmetros populacionais, sendo a partir deste valor que a média ponderada, a variância, e o teste de homogeneidade são calculados no *modelo dos efeitos aleatórios* (DerSimonian & Laird, 1986; Hedges, 1983; Hedges & Vevea, 1998). Os procedimentos utilizados para obtenção dos efeitos globais das correlações, tanto as relativas aos resultados diretos dos instrumentos e aos *scores* de mudança, como as relativas aos itens do PSYCHLOPS, e aos domínios do PSYCHLOPS e do CORE-OM, foram idênticos aos descritos para o coeficiente alfa, com os devidos ajustamentos. Todos os cálculos foram feitos com recurso ao programa Comprehensive Meta-Analysis®, e ao programa Microsoft Office Excel®.

5. Resultados

5.1. Análise Psicométrica

5.1.1. Fiabilidade – Consistência interna

A especificidade dos desenhos dos diferentes estudos obrigaram a que a única propriedade do domínio da fiabilidade avaliável fosse a consistência interna, mediante o cálculo do coeficiente alfa para cada amostra. Relativamente ao CORE-OM, os valores do coeficiente alfa variam entre .93 e .95, sendo o valor mais baixo registado na primeira e na quinta amostra, e o valor mais alto referente à terceira amostra. Estes valores são considerados excelentes (Cronbach, 1951; Mâroco & Garcia-Marques, 2006). A variabilidade de valores obtidos no PSYCHLOPS é maior, uma vez que estes se situam entre .72 (na segunda amostra) e .86 (na terceira amostra). Estes valores podem ser considerados como bons (Cronbach, 1951; Mâroco & Garcia-Marques, 2006). Os valores estão representados graficamente nas Figuras 1 e 2, e podem ser consultados em anexo (Anexo 1).

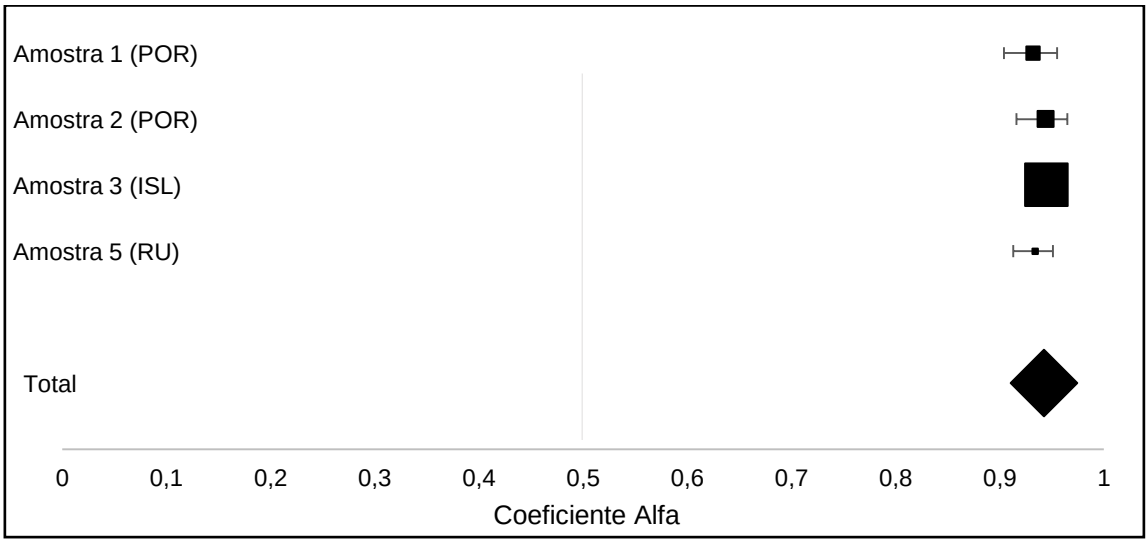

Figura 1: Valores do coeficiente alfa do CORE-OM para cada amostra e coeficiente alfa médio.

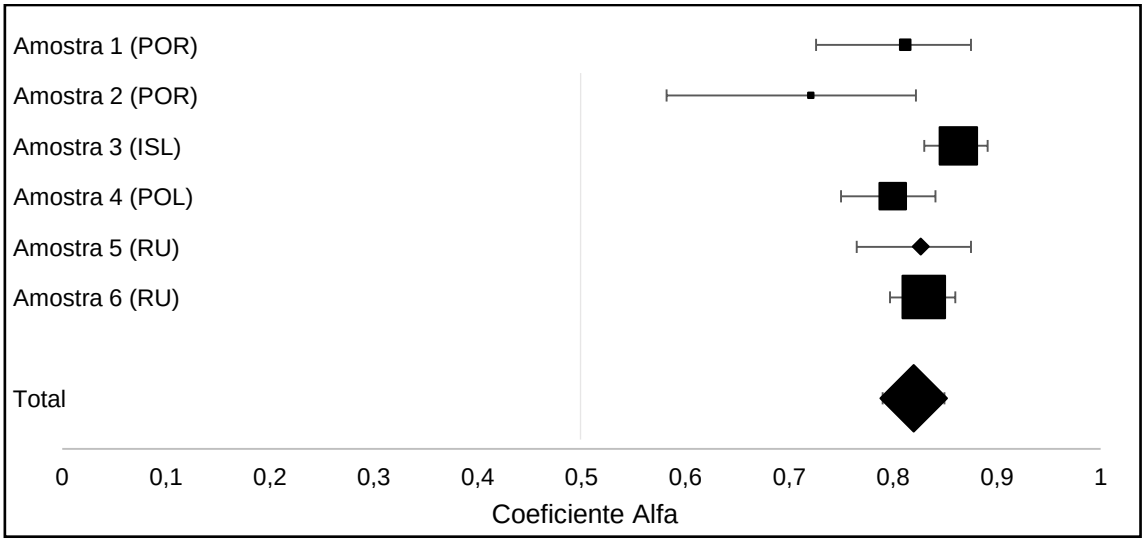

Figura 2: Valores do coeficiente alfa do PSYCHLOPS para cada amostra e coeficiente alfa médio.

O cálculo do valor médio do coeficiente alfa, integrando a informação das diferentes amostras, foi feito de acordo com uma abordagem meta-analítica, tendo a assunção de não homogeneidade entre os coeficientes suportado o recurso a uma análise de acordo com um modelo de efeitos aleatórios (Borenstein et al., 2009). A ponderação do peso de cada amostra foi feita a partir do inverso da variância, que no modelo de efeitos aleatórios considera também a variância entre estudos (Hedges & Vevea, 2008). O coeficiente alfa médio ponderado para o CORE-OM é de .94 (intervalo de confiança a 95%: .93-.95). Embora o teste de homogeneidade (Q = 1.0, *df* = 3, *p* = .801) não permita rejeitar a hipótese de que exista um efeito único subjacente, comum a todas as amostras, a potência deste teste de homogeneidade é bastante reduzida (*power* =.27 para detetar heterogeneidade elevada; *power* = .20 para detetar heterogeneidade moderada; *power* = .12 para detetar heterogeneidade reduzida). Esta baixa potência para deteção de um efeito (i.e., reduzida probabilidade de identificar diferenças reais entre os coeficientes alfa de cada amostra) torna razoável que a escolha do modelo estatístico usado na meta-análise seja feita com base em considerações teóricas referentes à não equivalência dos efeitos em estudo (Borenstein et al., 2009; Hedges & Pigott, 2001).

O valor médio ponderado do coeficiente alfa para o PSYCHLOPS é de .82 (intervalo de confiança a 95%: .79-.85). À semelhança do que ocorreu no cálculo do valor médio ponderado do coeficiente alfa para o CORE-OM, o teste de homogeneidade não permite rejeitar a hipótese nula (Q = 10.4 *df* = 5, *p* = .064). Contudo, a potência deste teste (*power* = .35 para uma heterogeneidade elevada; *power* = .25 para uma heterogeneidade moderada; *power* = .14 para uma

heterogeneidade reduzida) indica que a adoção do modelo de efeitos aleatórios pode ser adequada, pelo motivos referidos no parágrafo anterior. Globalmente, os resultados suportam as hipóteses estabelecidas quanto ao valor médio ponderado do coeficiente alfa para o CORE e para o PSYCHLOPS, bem como para a diferença entre ambos.

5.1.2. Validade de constructo

No presente estudo, as correlações entre o CORE-OM e o PSYCHLOPS variam entre r = .50 (p < .001) na Amostra 2 e r = 0,71 (p = .001) na Amostra 3. Foi adotada uma abordagem meta-analítica no cálculo da correlação média, de acordo com um modelo de efeitos aleatório, tendo-se obtido o valor r = .63 (intervalo de confiança a 95%: .53-.71). Esta correlação é moderada, o que suporta a hipótese estabelecida. Os valores das correlações podem ser consultados na Figura 3, e podem ser consultados em anexo (Anexo 2).

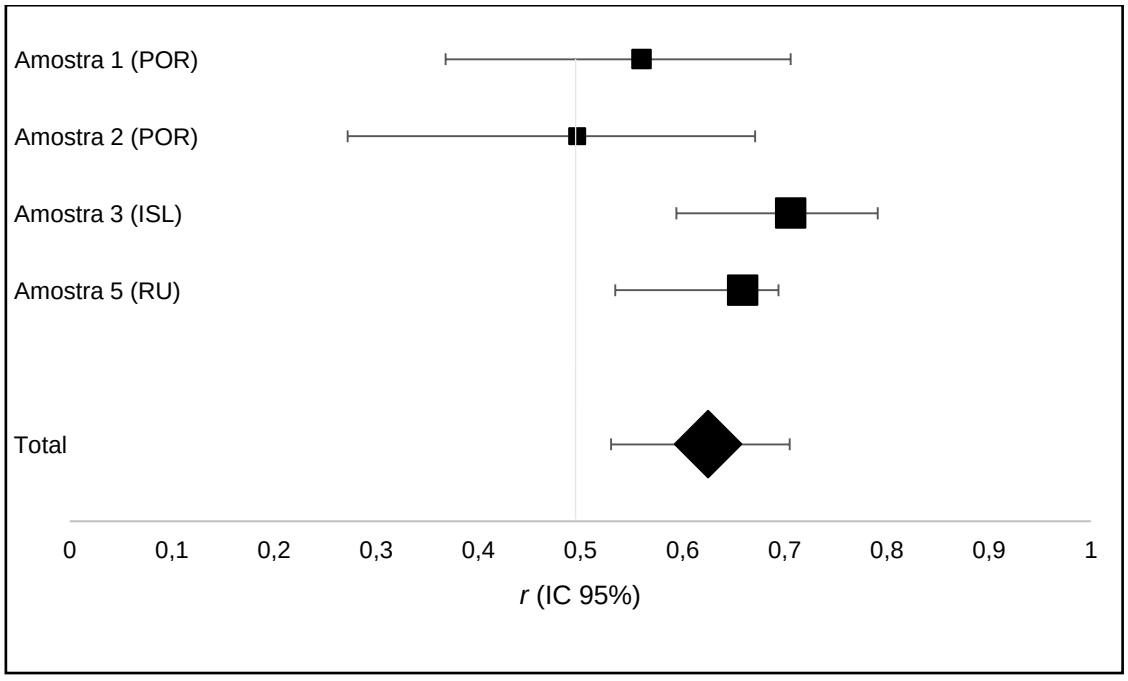

Figura 3: Correlações entre o CORE-OM e o PSYCHLOPS, por amostra, e correlação média.

5.1.2.1. Validade Convergente e Discriminante

No que respeita à convergência das medidas avaliativas dos domínios semelhantes do PSYCHLOPS e do CORE-OM, verificou-se que as correlações entre os domínios Problemas variam entre *r* = .31 (*p* < .001) na segunda amostra portuguesa, e *r* = .66 (*p* < .001) na amostra islandesa. Relativamente aos domínios Funcionamento, os valores das correlações entre PSYCHLOPS e CORE-OM variam entre *r* = .32 nas duas amostras portuguesas (*p* = 0,01 para ambas), e *r* = .49 (*p* < .001) na amostra islandesa. Nos domínios Bem-estar, as correlações oscilam entre *r* = .54 na segunda amostra portuguesa (*p* < .001) e *r* = .68 na amostra islandesa (*p* < .001). Quanto à validade discriminante, constata-se que os valores das correlações entre os resultados totais do PSYCHLOPS e a subescala Risco do CORE-OM variam entre *r* = .26 (*p* = .01) na primeira amostra britânica, e r = .39 na amostra islandesa (*p* < .001). Os valores médios das correlações referentes à validade convergente e discriminante podem ser consultados na Tabela 3. Os valores de cada amostra podem ser consultados em anexo (Anexo 3).

Tabela 3: Correlações médias (e intervalos de confiança a 95%) entre os domínios do PSYCHLOPS e do CORE-OM.

Domínios	CORE-OM Problemas	CORE-OM Funcionamento	CORE-OM Bem-estar	CORE-OM Risco
PSYCHLOPS Problemas	.51 [.35-.64]	.42 [.26-.55]	.51 [.38-.62]	.29 [.20-.37]
PSYCHLOPS Funcionamento	.44 [.31-.55]	.42 [.34-.50]	.45 [.36-.52]	.24 [.15-.33]
PSYCHLOPS Bem-estar	.62 [.50-.71]	.58 [.52-.64]	.35 [.26-.43]	.37 [.26-.46]
PSYCHLOPS Total	-	-	-	.31 [.29-.69]

5.1.2.2. Validade Estrutural

Os constrangimentos supramencionados, referentes à eventual presença de indicadores causais, impossibilitou a realização de análises fatoriais. Assim, recorreu-se à avaliação dos valores das correlações entre os itens do PSYCHLOPS para avaliação da validade estrutural. Após aplicar os procedimentos meta-analíticos, anteriormente descritos, para agregação dos valores das correlações obtidos em cada amostra, obtiveram-se os valores que podem ser consultados na Tabela 4.

Tabela 4: Correlações médias entre os itens do PSYCHLOPS (estimativas globais e intervalos de confiança 95%).

Itens	Item 1	Item 2	Item 3	Item 4
Item 1	1.00			
Item 2	.63 [.55-.70]	1.00		
Item 3	.54 [.50-.59]	.49 [.41-.57]	1.00	
Item 4	.58 [.51-.64]	.51 [.43-.58]	.50 [.41-.58]	1.00

5.1.3. Sensibilidade para detetar mudança

No presente estudo apenas foi possível calcular o valor da correlação dos *scores* de mudança em duas amostras, uma vez que apenas a amostra islandesa e a primeira amostra britânica continham dados recolhidos no final do tratamento, através da administração do CORE-OM e do PSYCHLOPS. O valor obtido na amostra islandesa é de r = .71, e na primeira amostra britânica é de r = .63. O cálculo do valor médio da correlação fez-se segundo uma abordagem meta-analítica, de acordo com um modelo de efeitos aleatórios, que permitiu obter um valor de correlação global de r = .67 (intervalo de confiança a 95%: .58-.75), um valor moderado. Os valores podem ser consultados na Figura 4.

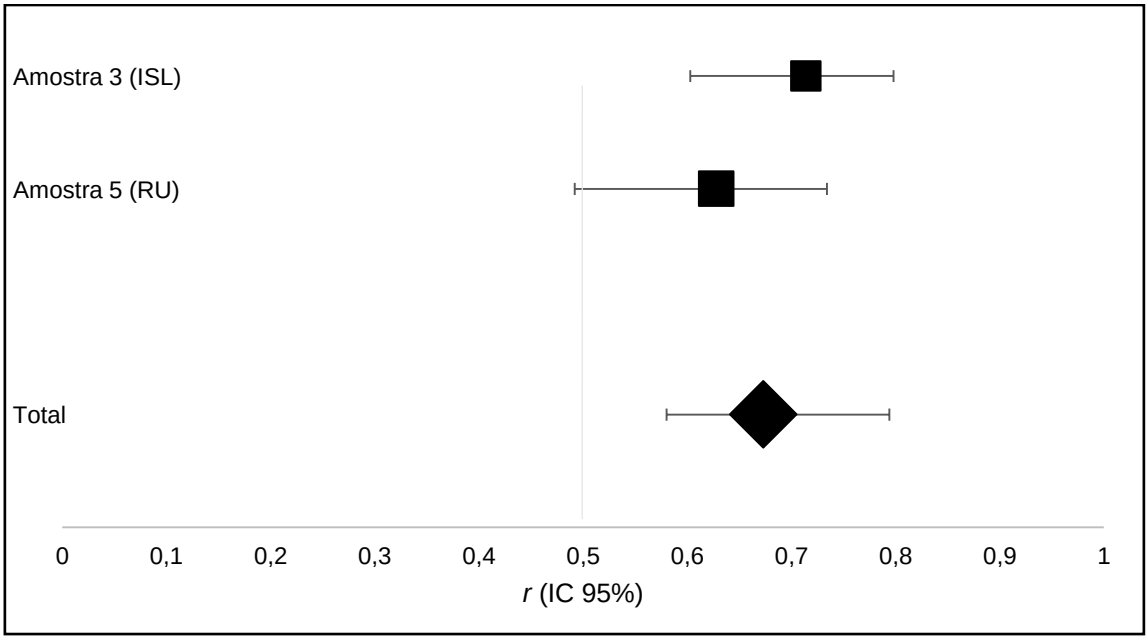

Figura 4: Correlações entre os *scores* de mudança do CORE-OM e do PSYCHLOPS, por amostra, e correlação média.

6. Discussão

O presente estudo tem como objetivo central analisar as propriedades psicométricas de uma medida individualizada de mudança, o PSYCHLOPS, assim como realizar uma revisão teórica do conhecimento referente às propriedades psicométricas de medidas nomotéticas. Pretende-se também avaliar a adequabilidade e utilidade do estudo das propriedades psicométricas na avaliação da qualidade e validade de MIM. Desta forma, espera-se contribuir, não só para o avanço da avaliação da mudança em psicoterapia, mas também para o desenvolvimento de medidas robustas, que consubstanciem a eficácia de intervenções psicoterapêuticas e que melhorem a qualidade do *feedback* obtido por terapeutas.

6.1. Fiabilidade – Consistência Interna

A avaliação das propriedades psicométricas iniciou-se pela fiabilidade, através do cálculo do coeficiente alfa (Cronbach, 1951) para avaliação da consistência interna. Globalmente, é possível concluir que o PSYCHLOPS apresenta uma boa consistência interna (alfa = .82), sendo esta estimativa baseada nas seis amostras estudadas, cujos valores de alfa não se afastaram grandemente do valor médio referido; na verdade, apesar de se registar um valor mais baixo na segunda amostra portuguesa (.72), as restantes amostras apresentam valores muito semelhantes entre si (entre .80 e .86), não havendo evidência de heterogeneidade nos valores do coeficiente alfa que ponha em causa a consistência interna do instrumento. O valor médio do coeficiente alfa do PSYCHLOPS é semelhante ao obtido no estudo da consistência interna do *Personal Questionnaire* (alfa = .80) (Elliott et al., 2016). É importante ter em conta que o CORE-OM apresentou um valor médio mais elevado para o coeficiente alfa (.94), eventualmente por reunir um maior número de itens. O valor médio do coeficiente de alfa é notável, particularmente quando se tem em conta que o PSYCHLOPS apenas contém quatro itens, e que este coeficiente é influenciado pelo número de itens (de Vet et al., 2011; Field, 2009). Contudo, é importante notar que o cálculo deste coeficiente pressupõe a verificação prévia da unidimensionalidade da escala, mediante análise fatorial ou análise de componentes principais, o que não foi feito no presente estudo. Adicionalmente, é fundamental considerar as dificuldades que a não equivalência dos itens em MIM criam à interpretação do coeficiente alfa. Quando se tem em conta que este coeficiente consiste essencialmente na estimação do nível de concordância médio de todos os testes *split-half* possíveis (Cronbach, 1951), e que, para um instrumento ter consistência interna, os itens que o compõem meçam diferentes

aspetos de um mesmo constructo, não é claro que o valor obtido numa medida individualizada de mudança possa ser interpretado à semelhança do que ocorre em medidas nomotéticas. De facto, enquanto a consistência interna de medidas nomotéticas é avaliada a partir da administração de itens iguais a sujeitos diferentes, no PSYCHLOPS (e em MIM em geral) poder-se-á argumentar que o coeficiente alfa consiste na estimação do valor médio de todos os testes *split-half* possíveis entre vários subconjuntos de quatro itens, um conjunto para cada participante. Portanto, não é possível perceber se os participantes respondem ao mesmo instrumento, ou se responderam a instrumentos diferentes, cujas respostas são semelhantes. Assim, não é claro se o valor do coeficiente alfa é interpretável, no sentido em que não parece ser possível concluir categoricamente que todos os sujeitos responderam aos mesmos itens, e portanto ao mesmo instrumento, exceto no que concerne ao número de itens e à escala de resposta.

Os valores obtidos para o coeficiente alfa também são de interpretação difícil devido à ausência de uma teoria robusta referente à forma como as pessoas respondem a MIM. Enquanto no desenvolvimento de escalas nomotéticas, compostas por indicadores de efeito, são enfatizados aspetos como as correlações entre os itens, a unidimensionalidade, e a consistência interna na seleção dos itens que irão constar do instrumento final (Diamantopoulos & Sigaw, 2006; Nunnally & Bernstein, 1994), não há literatura que permita concluir que o processo de geração de itens em medidas individualizadas seja semelhante. Isto é, não é possível saber se, ao criarem itens para uma medida individualizada de mudança, as pessoas optam por itens próximos de um mesmo constructo, e que estariam altamente correlacionados, ou se optam por incluir itens referentes a aspetos mais distantes do constructo sob estudo (no caso do PSYCHLOPS, o sofrimento psicológico), de forma a ser possível representar adequadamente todas as dimensões. Se assim for, poderia esperar-se um valor mais baixo do coeficiente alfa, na medida em que, contrariamente ao que acontece no desenvolvimento de medidas nomotéticas, as pessoas optariam por incluir itens menos correlacionados na tentativa de incluir todas as dimensões que considerassem relevantes, independentemente da existência de correlação entre os itens. A ausência de teoria e de conhecimento empírico que respondam a esta questão dificulta a avaliação da consistência interna em MIM em geral, e no PSYCHLOPS em particular, uma vez que não é possível avaliar hipóteses referentes a pressupostos processuais da geração e seleção de itens, que permitiriam inferir acerca da especificidade de MIM na avaliação de *outcome*. Desta forma, o valor do coeficiente alfa será bom, caso o processo individual de geração de itens em medidas individualizadas obedeça aos

mesmos parâmetros que norteiam a seleção e geração de itens em medidas nomotéticas.

Relativamente à consistência interna, deverá ainda mencionar-se que os procedimentos de cotação sugeridos pelos autores do PSYCHLOPS poderão inflacionar ligeiramente o valor do coeficiente de alfa. Concretamente, os procedimentos de cotação ditam que o valor do primeiro item deverá ser atribuído ao segundo item, quando a pessoa não identifica um segundo problema. Nos casos em que é necessário recorrer a este procedimento para cotar o instrumento, é artificialmente criada uma correlação perfeita entre os dois primeiros itens, o que poderá resultar no inflacionamento do coeficiente alfa. No presente estudo, foi necessário recorrer a este procedimento em perto de 10% dos casos, o que sugere algum este método de cotação poderá ter algum peso no valor de alfa; no entanto, esta hipótese não foi testada no presente estudo.

6.2. Validade de constructo

Os resultados referentes ao valor das correlações entre o PSYCHLOPS e o CORE-OM suportam as hipóteses referentes à validade de constructo, existindo por isso evidência que atesta a validade do PSYCHLOPS. O valor médio de correlação obtido (.63) é ligeiramente inferior aos valores obtidos nas amostras britânica e islandesa, e já publicados (Ashworth et al., 2005; Héðinsson et al., 2012), e está próximo do valor obtido na comparação do *Personal Questionnaire* com o CORE-OM (r = .60; Elliott et al., 2016). Contudo, à semelhança do que ocorreu na avaliação da consistência interna, a interpretação do valor desta correlação depende das considerações teóricas que justificam a relação entre o PSYCHLOPS e o CORE-OM. Se, por um lado, se considera que o PSYCHLOPS avalia um constructo semelhante, mas não igual, ao constructo avaliado pelo CORE-OM, então o valor da correlação é adequado. Porém, se a utilização de MIM é justificada pela ideia que o sofrimento psicológico é de tal forma idiossincrático que a sua medição apenas poderá ser adequadamente feita através do recurso a este tipo de medidas, então a correlação obtida será talvez excessivamente elevada uma vez que a sua magnitude não indicia diferenças marcadas nos constructos avaliados pelos dois instrumentos. Decorre dos resultados não ser possível provar, do ponto de vista psicométrico, que o PSYCHLOPS avalie aspetos particulares do sofrimento psicológico que não sejam captados também pelo CORE-OM. Na verdade, o argumento de que a existência diferenças interindividuais pode diminuir a validade de medidas nomotéticas não parece ser suportado pela análise de validade realizada; se assim fosse, não se

deveria constatar que um instrumento capaz de captar particularidades do sofrimento psicológico a que instrumentos nomotéticos não seriam sensíveis se correlacionasse fortemente com medidas proporcionadas por esses mesmos instrumentos nomotéticos. Ou seja, os resultados obtidos apenas parecem evidenciar que o PSYCHLOPS mede um constructo semelhante ao que é medido pelo CORE-OM. Assim, uma vez que do ponto de vista psicométrico não é possível provar a especificidade das medidas individuais de mudança, e que neste estudo o CORE-OM demonstrou maior consistência interna, poderá ser necessário aceitar outros aspetos como justificações mais importantes para a utilização deste género de medidas (e.g. utilidade clínica; redução do fardo do preenchimento de instrumentos nomotéticos para o paciente).

Mesmo não sendo possível avaliar a validade convergente e discriminante com recurso a instrumentos teoricamente próximos, ou distantes do PSYCHLOPS, no presente estudo procedeu-se à análise das correlações entre os domínios do PSYCHLOPS e do CORE-OM como forma de avaliar esta componente da validade de constructo. Os resultados obtidos não suportam totalmente as hipóteses colocadas. No que diz respeito à validade convergente, a consideração dos intervalos de confiança das correlações indica essencialmente que os domínios do PSYCHLOPS se correlacionam tanto com os domínios correspondentes do CORE-OM, como com os que não são correspondentes. Contudo, é relevante notar que as subescalas Problemas e Funcionamento do PSYCHLOPS apresentam correlações tão ou mais elevadas com a subescala Bem-Estar do CORE-OM do que com as subescalas correspondentes. Concomitantemente, a subescala Bem-estar do PSYCHLOPS apresenta correlações significativamente mais elevadas com as subescalas Problemas e Funcionamento do CORE-OM. Na verdade, a subescala Bem-estar do PSYCHLOPS apresenta um valor de correlação com a subescala correspondente mais baixo do que o valor da correlação com a subescala Risco do CORE-OM, teoricamente representativa de uma componente do sofrimento psicológico que não é avaliada pelo PSYCHLOPS. Assim, apesar de ter sido possível confirmar a maioria das hipóteses referentes à força das correlações entre domínios correspondentes, o facto de existirem correlações tão ou mais fortes com outros domínios levanta questões quanto à estrutura do PSYCHLOPS.

Ainda assim, há que considerar que existem divergências entre os itens que compõem os domínios dos dois instrumentos, de tal ordem que é questionável se teoricamente faz sentido considerá-los como equivalentes. Por exemplo, há itens da subescala Bem-estar do CORE-OM que estão mais próximos da subescala Problemas do que com a subescala Bem-estar do PSYCHLOPS, como o item “Senti que os meus

problemas são demais para mim". Adicionalmente, enquanto o domínio Problemas do CORE-OM é maioritariamente composto por itens referentes a sintomas (i.e., por indicadores de efeito), é razoável colocar a hipótese de que o domínio Problemas do PSYCHLOPS seja, pelo menos parcialmente, composta por indicadores causais, o que possivelmente teria efeito na correlação entre os domínios. A razoabilidade desta hipótese é suportada pela descrição do PSYCHLOPS como baseado numa sequência empírica de causalidade, onde os problemas identificados são conceptualizados como despoletando dificuldades funcionais (Ashworth et al., 2012). Ainda assim, fica por explicar por que motivo as correlações com os restantes domínios apresentam valores semelhantes, uma vez que, a título de exemplo, não existem itens diretamente ligados à dificuldade na realização de tarefas no domínio Problemas do CORE-OM que permitam compreender a existência de uma correlação tão elevada entre o domínio Funcionamento do PSYCHLOPS e o domínio Problemas do CORE-OM. Desta forma, ainda que possam atribuir-se os valores das correlações ao facto de ambos os instrumentos serem compostos por domínios que, não sendo correspondentes, avaliam o mesmo constructo, os resultados indicam que as subescalas do PSYCHLOPS não são particularmente discriminativas. No que diz respeito à validade discriminante, os resultados indicam existir evidência deste tipo de validade no que concerne à ausência de correlação com a subescala Risco do CORE-OM, mas o facto de existirem correlações entre subescalas que, apesar de surgirem nos dois instrumentos, não são equivalentes, levanta algumas dúvidas quanto à capacidade do PSYCHLOPS diferenciar de forma válida esses domínios.

A distinção entre indicadores causais e indicadores de efeito também deve ser considerada na interpretação dos resultados obtidos na avaliação da validade de constructo. Particularmente, ainda que a possibilidade de que alguns itens do PSYCHLOPS sejam indicadores causais tenha impedido a realização de análise fatorial exploratória, não foi objetivo deste estudo testar empiricamente esta hipótese, nem tal seria possível. Assim, será importante compreender, em estudos futuros, se esta é efetivamente uma questão relevante na avaliação das propriedades psicométricas de MIM, e isto por dois motivos. Em primeiro lugar, o coeficiente alfa em indicadores causais pode ser desprovido de significado, ou ser mais reduzido, devido à inexistência de correlação entre estes indicadores; contudo, tal não se verificou no presente estudo, o que não suporta a hipótese de que alguns dos itens do PSYCHLOPS sejam indicadores causais independentes entre si. Em segundo lugar, o facto de as três primeiras perguntas do PSYCHLOPS estarem divididas em duas partes (uma, na qual o sujeito tem liberdade para escolher os problemas e dificuldades que considera relevantes, outra em que tem de classificar a severidade desses

problemas e dificuldades numa escala ordinal) conduz à possibilidade de se colocar a hipótese de neste instrumento se partir da identificação de causas do sofrimento psicológico, para avaliar, de forma nomotética, a severidade dos seus efeitos. Desta forma, estar-se-ia a proceder à transformação de indicadores causais em indicadores de efeito, pelo que as dificuldades na avaliação de propriedades psicométricas das respostas de severidade, tais como a consistência interna ou a validade estrutural, não seriam postas em causa. Como ocorre em relação às dificuldades na interpretação do coeficiente alfa, seria necessário que considerações teóricas referentes ao comportamento de resposta, e ao sentido da causalidade entre os itens e o constructo, permitissem clarificar esta questão. Assim, a possibilidade de que a divisão dos itens no PSYCHLOPS permita transformar indicadores causais em indicadores de efeito é apenas uma hipótese que carece de sustentação, teórica ou empírica. Importa sublinhar que esta questão corresponde a uma particularidade do PSYCHLOPS que não é necessariamente generalizável a outras medidas individuais de mudança.

Como referido, neste estudo não foi avaliada a validade estrutural mediante análise fatorial exploratória, devido a dúvidas quanto ao sentido da causalidade nos itens do PSYCHLOPS. Ainda assim, considerou-se pertinente analisar as correlações entre os domínios do PSYCHLOPS. De acordo com a divisão do instrumento proposta pelos autores (nos domínios Problemas, Funcionamento, e Bem-estar), colocou-se a hipótese de que os dois itens pertencentes ao domínio Problemas apresentariam um valor de correlação mais elevado entre si do que com os restantes itens, por pertencerem ao mesmo domínio. Contudo, importa sublinhar que existem dúvidas quanto à razoabilidade desta hipótese, uma vez que, existindo a possibilidade de que as pessoas optem por indicar nestes itens as causas do seu sofrimento psicológico, poderá não ser expectável que a correlação entre os dois itens seja maior do que a correlação com os outros itens do instrumento. Os resultados não permitem confirmar totalmente as hipóteses referentes ao valor das correlações entre os domínios do PSYCHLOPS, na medida em que a correlação entre os dois itens do domínio Problemas não é tão elevada como o esperado, sendo próxima do valor das correlações com os outros itens. Para além disso, a interpretação dos resultados é dificultada pela existência de diferenças entre amostras (e.g., na segunda amostra a correlação entre o primeiro item do domínio Problemas e o domínio Funcionamento é a maior, acontecendo o mesmo na quarta amostra), não sendo contudo possível explicar por que motivo se observam estas diferenças entre amostras. Essencialmente, apesar de a avaliação da validade estrutural neste estudo ser incompleta, os resultados apontam para lacunas na teoria referente à estrutura do PSYCHLOPS, de tal forma que a formulação de hipóteses referentes à mesma pode

ser posta em causa. Ou seja, será necessário tornar mais clara a teoria referente à relação entre os domínios do PSYCHLOPS, para que seja possível testar hipóteses razoáveis quanto à sua estrutura (e.g., fará sentido esperar que os itens do domínio Problemas apresentem uma correlação elevada entre si, ou será mais expectável que a correlação entre o primeiro problema e o domínio Funcionamento seja a mais elevada pela existência de uma relação de causalidade entre ambos?).

6.3. Sensibilidade para detetar mudança

A análise das correlações entre os *scores* de mudança do PSYCHLOPS e do CORE-OM suporta as hipóteses referentes à sensibilidade para detetar mudança, pelo que se assumem como evidência da validade dos *scores* de mudança do PSYCHLOPS. O valor médio da correlação é semelhante ao que foi obtido por Héðinsson e colaboradores (2012), no único estudo que avaliou a sensibilidade para detetar mudança através da análise das correlações entre resultados de mudança. Contudo, o valor obtido corresponde a uma correlação forte (r = .71), o que tecnicamente não acontece no presente estudo (r = .67). Todavia, é necessário referir que a metodologia utilizada neste estudo não permite perceber se os *scores* de mudança do PSYCHLOPS são mais, ou menos, válidos do que os *scores* de mudança do CORE-OM, mas apenas que ambos os instrumentos medem mudança no mesmo constructo, ou num constructo próximo. À semelhança do que ocorreu na avaliação da validade de constructo, os resultados atestam a validade do PSYCHLOPS, mas não são totalmente consonantes com a premissa de que uma medida individualizada avalia mudanças no constructo sob estudo, de forma marcadamente diferente de medidas nomotéticas. Se assim fosse, esperar-se-ia uma correlação mais reduzida.

7. Conclusões gerais

Globalmente, o presente estudo permite compreender que o PSYCHLOPS é uma medida com boa consistência interna, apresenta resultados válidos, e é sensível à mudança em psicoterapia.

No que diz respeito à adequabilidade da avaliação psicométrica clássica para o estudo da validade de medidas individualizadas, importa sublinhar que a análise das correlações entre uma medida individualizada de mudança e uma medida nomotética não permite avaliar adequadamente a hipótese de que as medidas individualizadas são mais válidas na mensuração idiossincrática do constructo sob estudo. Assim é porque se se esperam correlações elevadas entre medidas individualizadas e medidas nomotéticas, então é razoável concluir que ambas medem fundamentalmente o mesmo constructo, ficando em segundo plano a importância das especificidades das medidas individualizadas, que poderiam ser analisadas recorrendo, por exemplo, a metodologias qualitativas, como a análise de conteúdo dos itens. Por outro lado, afirmar que, pela sua especificidade, as MIM evidenciariam correlações reduzidas com medidas nomotéticas, implica abdicar da avaliação psicométrica clássica da validade de constructo, dificultando grandemente qualquer tentativa de compreender o que está de facto a ser avaliado por essas MIM.

Ainda que seja possível argumentar que a psicometria clássica oferece soluções limitadas para a avaliação das propriedades psicométricas de MIM como o PSYCHLOPS, é também necessário considerar que, nalguns aspetos, este tipo de medidas é difícil de avaliar empiricamente. Para tal contribui a ausência de uma teoria robusta referente ao comportamento de resposta a MIM, porque a incerteza quanto à forma e ao grau com que os utilizadores tentam representar um constructo (i.e. o(s) que quer(em) incluir no instrumento) dificulta a interpretação do coeficiente alfa, e do valor das correlações com medidas nomotéticas. A isto acrescem as dificuldades ligadas à não equivalência de itens e à eventual presença de indicadores causais, já mencionadas.

As críticas dirigidas às medidas nomotéticas, nomeadamente o reduzido envolvimento do paciente na escolha de assuntos, problemas e dificuldades, que servem de base para a avaliação do seu progresso e da eficácia do tratamento recebido, serviram como ponto de partida para o presente estudo. Procurou-se compreender se o PSYCHLOPS evidencia robustez psicométrica, complementar à especificidade do conteúdo de MIM já avaliada através de metodologias qualitativas (Ashworth et al., 2007; Hunter et al., 2004; Wagner & Elliott, 2001). Nesse sentido, este estudo apresenta dados empíricos que suportam a utilização do PSYCHLOPS

como medida de mudança, caso se aceite que as dificuldades na interpretação das análises psicométricas, descritas anteriormente, não são particularmente significativas. A título de exemplo, Elliott e colaboradores (2016) desvalorizam as dificuldades ligadas à não equivalência de itens, afirmando que a subjetividade na interpretação de um item significa que, na verdade, nunca é possível afirmar que duas pessoas respondem ao mesmo item, uma posição que pode ser interpretada como pondo em causa grande parte das assunções da avaliação psicológica. Se efetivamente se optar por desvalorizar estas questões, é também necessário apresentar evidência adicional que justifique a utilização preferencial do PSYCHLOPS em detrimento do CORE-OM, uma vez que o primeiro não apresenta uma consistência interna tão elevada como o segundo, por exemplo. Ou seja, uma vez que este estudo apresenta evidência de que o PSYCHLOPS e o CORE-OM medem o mesmo constructo, e portanto mudança no mesmo constructo, é necessário apresentar argumentação sólida que permita considerar a especificidade do conteúdo das medidas individualizadas como uma característica mais importante do que propriedades psicométricas como a consistência interna. Assim é porque este estudo demonstra que o PSYCHLOPS não é, do ponto de vista da consistência interna, tão robusto como o CORE-OM, o que implica que a defesa da utilização preferencial de MIM não seja feita com recurso exclusivo a evidência psicométrica. No fundo, não será errado defender que o presente estudo ilustra as dificuldades que o envolvimento participativo de pacientes na avaliação do seu progresso e do tratamento coloca à análise psicométrica dessa avaliação, e vice-versa, pelo menos nos moldes em que é concebido nas MIM. Contudo, não é claro se tal significa que seja necessário proceder a mudanças nas metodologias de avaliação das propriedades psicométricas, ou a alterações nas MIM.

Limitações e estudos futuros

Tendo em conta que a possível presença de indicadores causais nos itens do PSYCHLOPS impossibilitou a realização de algumas análises e dificultou a interpretação dos resultados, pode argumentar-se que a principal limitação do presente estudo consiste na ausência de metodologias que permitam avaliar se a presença de indicadores causais é, efetivamente, um problema em MIM., Assim, será importante avaliar a extensão desta dificuldade em estudos futuros.

Independentemente dos resultados obtidos, os métodos utilizados não permitiriam concluir se o PSYCHLOPS é uma medida de mudança com maior ou menor validade do que o CORE-OM. Devido às características dos estudos, dos quais

as amostras utilizadas provieram, não foi possível avaliar a fiabilidade teste-reteste, o que também corresponde a uma limitação significativa.

Ainda que parte dos motivos que impediram a realização de análise fatorial exploratória se deva à possível existência de indicadores causais, o facto de três das quatro escalas do PSYCHLOPS serem compostas por um só item também torna pouco plausível a realização de técnicas de análise fatorial para avaliar a estrutura de domínios do PSYCHLOPS. Ainda assim, pode ser importante, no futuro, proceder à verificação da unidimensionalidade do PSYCHLOPS, com recurso a análise fatorial.

Neste estudo também não foi possível avaliar adequadamente os efeitos das diferentes amostras nos resultados, uma vez que os dados utilizados foram recolhidos em diferentes países, no contexto de diferentes investigações. Isto é, embora a heterogeneidade das amostras possa estar na base de algumas diferenças nos resultados, não foi possível compreender até que ponto as diferenças encontradas se devem a diferenças nas amostras ou se estão diretamente ligadas a características do instrumento sob estudo.

Por fim, o estudo da sensibilidade para detetar mudança não é generalizável, devido ao número reduzido de amostras, nas quais se procedeu à administração do PSYCHLOPS e do CORE-OM pré e pós-tratamento. Assim, em estudos futuros, deverá ser tentado o aumento do número de amostras, para que seja possível obter uma melhor estimativa da sensibilidade para detetar mudança do PSYCHLOPS.

Referências

Almeida, L.S., & Freire, T. (2003). *Metodologia de investigação em psicologia e educação*. Braga: Psiquilíbrios.

Ashworth, M., Evans, C., & Clement, S. (2009). Measuring psychological outcomes after cognitive behavior therapy in primary care: A comparison between a new patient-generated measure “PSYCHLOPS” (Psychological Outcome Profiles) and “HADS” (Hospital Anxiety and Depression Scale). *Journal of Mental Health, 18,* 167-177. http://dx.doi.org/10.1080/09638230701879144

Ashworth, M., Kordowicz, M., & Schofield, P. (2012). The PSYCHLOPS (Psychological Outcome Profiles). *Integrating Science and Practice, 2,* 36-39.

Ashworth, M., Robinson, S., Evans, C., Shepherd, M., Conolly, A., & Rowlands, G. (2007). What does an idiographic measure (PSYCHLOPS) tell us about the spectrum of psychological issues and scores on a nomothetic measure (CORE-OM)? *Primary Care and Community Psychiatry, 12*, 7-18.

Ashworth, M., Robinson, S., Godfrey, E., Shepherd, M., Evans, C., Seed, P., Parmentier, H., Tylee,A. (2005). Measuring mental health outcomes in primary care: the psychometric properties of a new patient-generated outcome measure, ‘PSYCHLOPS’ (‘psychological outcome profiles’). *Primary Care Mental Health, 3,* 261-270.

Ashworth, M., Robinson, S.I., Godfrey, E., Parmentier, H., Shepherd, M., Christey, J., Wright, K., & Matthews, V. (2005b). The experiences of therapists using a new client-centred psychometric instrument, PSYCHLOPS (Psychological Outcome Profiles). *Counselling and Psychotherapy Research, 5,* 37-45. http://dx.doi.org/10.1080/14733140512331343886

Ashworth, M., Shepherd, M., Christey, J., Matthews, V., Wright, K., Parmentier, H., Robinson, S., & Godfrey, E. (2004). A client-generated psychometric instrument: the development of ‘PSYCHLOPS’. *Counselling and Psychotherapy Research,4,* 27-31. http://dx.doi.org/10.1080/14733140412331383913

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., Benson, L., Connell, J., Audin, K., & McGrath, G. (2001). Service Profiling and Outcomes Benchmarking Using the CORE-OM: Toward Practice-Based Evidence in the Psychological Therapies. *Journal of Counsulting and Clinical Psychology. 60,* 184-196. http://dx.doi.org/10.1037/0022-006X.69.2.184

Becker, L.A. (2000). *Effect Size*. Acedido a 21 de novembro de 2015, em http://www.bwgriffin.com/gsu/courses/edur9131/content/EffectSizeBecker.pdf

Boehmer, S., & Luszczynska, A. (2006). Two kinds of items in quality of life instruments: 'Indicator and causal variables' in the EORTC QLQ-C30. *Quality of Life Research, 15,* 131-141. http://dx.doi.org/10.1007/s11136-005-8290-6

Bollen, K.A. (1989). *Structural equations with latent variables*. John Wiley & Sons, New York. http://dx.doi.org/10.1002/9781118619179

Bollen, K., & Lennox, R. (1991). Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin, 110,* 305-314. http://dx.doi.org/10.1037/0033-2909.110.2.305

Bollen, K.A., & Ting, K.F. (2000). A tetrad test for causal indicators. *Psychological Methods, 5,* 3-22. http://dx.doi.org/10.1037/1082-989X.5.1.3

Berwick, D.M (2002). A user's manual for the IOM's 'quality chasm' report. *Health Affairs, 21*, 80-90. http://dx.doi.org/10.1377/hlthaff.21.3.80

Blount, C., Evans, C., Birch, S., Warren, F., & Norton, K. (2002). The properties of self-report research measures: Beyond psychometrics. *Psychology and Psychotherapy: Theory, Research and Practice, 75,* 151-164. http://dx.doi.org/10.1348/147608302169616

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to Meta-Analysis.* West Sussex: John Wiley & Sons. http://dx.doi.org/10.1002/9780470743386

Borsboom, D., Mellenbergh, G.J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 11,* 1061-1071. http://dx.doi.org/10.1037/0033-295X.111.4.1061

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

Cortina, J.M. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology, 78*, 98-104. http://dx.doi.org/10.1037/0021-9010.78.1.98

Coulter, A. (1999). Paternalism or partnership? *British Medical Journal, 319*, 719-720. http://dx.doi.org/10.1136/bmj.319.7212.719

Crawford, M., Rovotham, D., Thana, L., Patterson, S., Weaver, T., Barber, R., Wykes, T., & Rose, D. (2011). Selecting outcome measures in mental health: the views of service users. *Journal of Mental Health, 20,* 336-346. http://dx.doi.org/10.3109/09638237.2011.577114

Cronbach, L.J. (1951). Coefficient alpha ant the internal structure of tests. *Psychometrika, 3,* 297- 334. http://dx.doi.org/10.1007/BF02310555

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302. http://dx.doi.org/10.1037/h0040957

Crosby, R.D., Kolotkin, R.L., Williams, G.R. (2003). Defining Clinically Meaningful Change in Health-Related Quality of Life. *Journal of Clinical Epidemiology, 56,* 395-407. http://dx.doi.org/10.1016/S0895-4356(03)00044-1

Czachowski, S., Seed, P., Schofield, P., Ashworth, M. (2011). Measuring Psychological Change during Cognitive Behaviour Therapy in Primary Care: A Polish Study Using 'PSYCHLOPS' (Psychological Outcome Profiles). *Plos One, 6,* 1-6. http://dx.doi.org/10.1371/journal.pone.0027378

de Vet, H.W., Terwee, C., Mokkink, L.B., & Knol, D.L. (2011). *Measurement in medicine: A practical guide*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511996214

Davy, Z., Quinn, C., Togher, F., Wilson, H., & Siriwardena, N. (2012). Investigating qualitative and quantitative validity of PSYCHLOPS: a novel Patient Reported Outcome Measure in a pilot study of primary care management of insomnia. *The International Journal of Person Centered Medicine, 2,* 845-852.

DerSimonian, R., & Laird, N. (1986). Meta-Analysis in clinical trials. *Controlled Clinical Trials, 7,* 177-188. http://dx.doi.org/10.1016/0197-2456(86)90046-2

Deshpande, P.R., Rajan, S., Sudeepthi, B.L., & Nazir, C.P. (2011). Patient-reported outcomes: A new era in clinical research. *Perspectives in Clinical Researh, 2,* 137-144. http://dx.doi.org/10.4103/2229-3485.86879

Diamantopoulos, A., Riefler, P., & Roth, K.P. (2008). Advancing formative measurement models. *Journal of Business Research, 61,* 1203-1218. http://dx.doi.org/10.1016/j.jbusres.2008.01.009

Diamantopoulos, A., & Siguaw, J.A. (2006). Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration. *British Journal of Management, 17,* 263-282. http://dx.doi.org/10.1111/j.1467-8551.2006.00500.x

Diamantopoulos, A., & Winklhofer, H. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 37,* 269-277. http://dx.doi.org/10.1509/jmkr.38.2.269.18845

Dixon, W.J. (1960). Simplified estimation from censored normal samples. *The annals of mathematical statistics, 31,* 385-391. http://dx.doi.org/10.1214/aoms/1177705900

Donovan, J.L., Frankel, S.J., & Eyles, J.D. (1993). Assessing the need for health status measures. *Journal of Epidemiology and Community Health, 47*, 158-162. http://dx.doi.org/10.1136/jech.47.2.158

Doucette, A., & Wolf, A. (2009). Questioning the measurement precision of psychotherapy research. *Psychotherapy Research Methods, 19,* 374-389. http://dx.doi.org/10.1080/10503300902894422

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures*. Psychological Methods, 5*, 155–174. http://dx.doi.org/10.1037/1082-989X.5.2.155

Eisen, S., Ranganathan, G., Seal, M., & Spiro, A. (2007). Measuring Clinically Meaningful Change Following Mental Health Treatment. *The Journal of Behavioral Health Services and Research, 34,* 272-289. http://dx.doi.org/10.1007/s11414-007-9066-2

Elfström, M.L., Evans, C., Lundgren, J., Johansson, B., Hakeberg, M., & Carlsson, S.G. (2013). Validation of the Swedish Version of the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM). *Clinical Psychology and Psychotherapy, 20,* 447-455. http://dx.doi.org/10.1002/cpp.1788

Eliasziw, M., & Donner, A. (1987). A cost-function approach to the design of reliability studies. *Statistics in Medicine, 6,* 647-655. http://dx.doi.org/10.1002/sim.4780060602

Elliott, R., Mack, C., & Shapiro, D. (1999). *Simplified Personal Questionnaire Procedure*. Retrieved from: http://experiential-researchers.org/instruments/elliott/pq procedure.html

Elliott, R., Wagner, J., Sales, C., Rodgers, B., Alves, P., & Café, M. (2016). Psychometrics of the Personal Questionnaire: A Client-Generated Outcome Measure. *Psychological Assessment, 28*, 263-278. http://dx.doi.org/10.1037/pas0000174

Entwistle, V.A., Renfrew, M.J., Yearley, S., Forrester, J., & Lamont, T. (1998). Lay perspectives: Advantages for health research. *British Medical Journal, 316*, 463–466. http://dx.doi.org/10.1136/bmj.316.7129.463

Evans, C., Ashworth, M., & Peters, M. (2010). Are problems prevalent and stable in non-clinical populations? Problems and test-retest stability of a patient-generated measure, PSYCHLOPS (Psychological Outcome Profiles), in a non-clinical student sample. *British Journal of Guidance & Counseling, 38,* 431-439. http://dx.doi.org/10.1080/03069885.2010.503701

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Towards a standardized brief outcome measure: psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry, 188,* 51-60. http://dx.doi.org/10.1192/bjp.180.1.51

Evans, C., Margison, F., & Barkham, M. (1998). The contribution of reliable and clinically significant change methods to evidence-based mental health. *Evidence Based Mental Health, 1,* 70.72. http://dx.doi.org/10.1136/ebmh.1.3.70

Faulkner, A., & Thomas, P. (2002). User-led research and evidence-based medicine. *British Journal of Psychiatry, 180*, 1-3. http://dx.doi.org/10.1192/bjp.180.1.1

Fayers, P.M., & Hand, D.J. (1997). Factor analysis, causal indicators and quality of life. *Quality of Life Research, 6*, 139-150. http://dx.doi.org/10.1023/A:1026490117121

Fayers, P.M., Hand, D.J., Bjordal, K., & Groenvold, M. (1997). Causal Indicators in Quality of Life research. *Quality of Life Research, 6,* 393-406. http://dx.doi.org/10.1023/A:1018491512095

Feinstein, A.R. (1987). *Clinimetrics*. New Haven: Yale University Press.

Field, A. (2009). *Discovering Statistics Using SPSS (and Sex and Drugs and Rock 'n' Roll)*. London: Sage Publications Ltd.

Fitzpatrick, M. (2012). Blurring Practice-Research Boundaries Using Progress Monitoring: A Personal Introduction to This Issue of Canadian Psychology. *Canadian Pschology, 2,* 75-81. http://dx.doi.org/10.1037/a0028051

Fitzpatrick R., Davey C., Buxton M.J. & Jones D.R. (1998) Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment, 2*,1–74.

Gerteis, M., Edgman-Levitan, S., Daley, J., Delbanco, T.L. (1993). *Through the patient's eyes: Understanding and promoting patient-centered care*. San Francisco: Josey-Bass.

Guyatt, G.H., Feeny, D.H., & Patrick, D.L. (1993). Measuring health-related quality of life. *Annals of Internal Medicine, 118*, 622-629. http://dx.doi.org/10.7326/0003-4819-118-8-199304150-00009

Guyatt, G.H., Kirshner, B., & Jaeschke, R. (1992). Measuring health status: what are the necessary measurement properties? *Journal of Clinical Epidemiology, 45,* 1341-1345. http://dx.doi.org/10.1016/0895-4356(92)90194-R

Hakstian, A.R., & Whalen, T.E. (1976). A *k*-sample significance test for independent alpha coefficients. *Psychometrika, 41,* 219-231. http://dx.doi.org/10.1007/BF02291840

Hartung, J., Knapp, G., & Sinha, B. (2008). *Statistical Meta-Analysis with applications*. New Jersey: John Wiley & Sons. http://dx.doi.org/10.1002/9780470386347

Hawkins, D.M. (1980). *Identification of Outliers*. London: Chapman and Hall. http://dx.doi.org/10.1007/978-94-015-3994-4

Haynes, S.N., Mumma, G.H., Pinson, C. (2009). Idiographic assessment: Conceptual and psychometric foundations of individualized behavioral assessment. *Clinical Psychology Review, 29,* 179-191. http://dx.doi.org/10.1016/j.cpr.2008.12.003

Hedges, L. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93,* 388-395. http://dx.doi.org/10.1037/0033-2909.93.2.388

Hedges, L., & Vevea, J. (1998). Fixed- and Random-Effects Models in Meta-Analysis. *Psychological Methods, 3,* 486-504. http://dx.doi.org/10.1037/1082-989X.3.4.486

Hedges, L., & Pigott, T. (2001). The Power of Statistical Tests in Meta-Analysis. *Psychological Methods, 6*, 203-217. http://dx.doi.org/10.1037/1082-989X.6.3.203

Hedges, L., & Pigott, T. (2004). The Power of Statistical Tests for Moderators in Meta-Analysis. *Psychological Methods, 9,* 426-445. http://dx.doi.org/10.1037/1082-989X.9.4.426

Héðinsson, H., Kristjánsdóttir, H., Ólason, D., & Sigurðsson, D.F. (2012). A Validation and Replication Study of the Patient-Generated Measure PSYCHLOPS on an Icelandic Clinical Population. *European Journal of Psychological Assessment, 29,* 89-95.

Holgersson, H., & Karlsson, P. (2012). Three estimators of the Mahalanobis distance in high-dimensional data. *Journal of Applied Statistics, 39,* 2713-2720. http://dx.doi.org/10.1080/02664763.2012.725464

Hunter, R., McLean, J., Peck, D., Pullen, I., Greenfield, A., McArthur, W., Quinn, C., Eaglesham, J., Hagen, S., & Norrie, J. (2004). The Scottish 700 Outcomes Study: A comparative evaluation of the Health of the Nation Outcome Scale (HoNOS), the Avon mental health measure (AVON), and an idiographic scale (OPUS) in adult mental health. *Journal of Mental Health, 13, 93-105.* http://dx.doi.org/10.1080/09638230410001654594

Hurlburt, R.T. & Knapp, T.J. (2006). Münsterberg in 1898, not Allport in 1937, Introduced the terms ‘idiographic’ and ‘nomothetic’ to American psychology. *Theory & Psychology, 16*, 287-293. http://dx.doi.org/10.1177/0959354306062541

Husted, J., Cook, R., Farewell, V., & Goldman, D. (2000). Methods for assessing responsiveness: a critical review and recommendations. *Journal of Clinical Epidemiology, 53,* 459-468. http://dx.doi.org/10.1016/S0895-4356(99)00206-1

Jacobson, N.S., & Truax, P. (1991). Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research. *Journal of Consulting and Clinical Psychology, 59,* 12-19. http://dx.doi.org/10.1037/0022-006X.59.1.12

Jarvis, C., MacKenzie, S., Podsakoff, P.A. (2003). Critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*, 579-601. http://dx.doi.org/10.1086/376806

Jenkins, P. & Turner, H. (2014). An investigation into the psychometric properties of the CORE-OM in patients with eating disorders. *Counselling and Psychotherapy Research, 14,* 102-110. http://dx.doi.org/10.1080/14733145.2013.782057

Kane, M. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement, 38,* 319-342. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kelley, S., Athay, M., Hargraves, R., Andrade, A. R., Tempesti, T., & Bickman, L. (2011). *Predicting clinician behavior based on feedback from the client and caregiver.* Paper presented at the 42nd International Meeting of the Society for Psychotherapy Research, Bern, Switzerland.

Kiresuk, T., & Sherman, R. (1968). Goal attainment scaling: a general method of evaluating comprehensive mental health programmes. *Community Mental Health, 4,* 443-453. http://dx.doi.org/10.1007/BF01530764

Lacasse, Y., Wong, E., & Guyatt, G. (1999) Individualizing questionnaires. In: C. Joyce, C.A. O'Boyle & H. McGee (Eds.). *Individual Quality of Life. Approaches to conceptualization and assessment* (pp. 87–103). Amsterdam: Harwood Academic Publishers.

Lambert, M.J., Whipple, J.L., Smart, D.W., Vermeersh, D.A., Nielsen, S.L., & Hawkins, E.J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research, 11,* 49-68. http://dx.doi.org/10.1080/713663852

Lyne, K.J., Barrett, P., Evans, C., & Barkham, M. (2006). Dimensions of variation on the CORE-OM. *British Journal of Clinical Psychology, 45,* 185-203. http://dx.doi.org/10.1348/014466505X39106

MacCallum, R., & Browne, M. (1993). The Use of Causal Indicators in Covariance Structure Models: Some Practical Issues. *Psychological Bulletin, 114*, 533-541. http://dx.doi.org/10.1037/0033-2909.114.3.533

Marôco, J., & Garcia-Marques, T. (2006). Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas? *Laboratório de Psicologia, 4,* 65-90. http://dx.doi.org/10.14417/lp.763

Mead, N., & Bower, P. (2000). Patient-centeredness: a conceptual framework and review of the empirical literature. *Social Science & Medicine, 51,* 1087-1110. http://dx.doi.org/10.1016/S0277-9536(00)00098-8

Mellor-Clark, J., Barkham, M., Connel, J., & Evans, C. (1999). Practice-based evidence and need for a standardized evaluation system: Informing the design of the CORE system. *European Journal of Psychotherapy, Counselling and Health, 3,* 357-374. http://dx.doi.org/10.1080/13642539908400818

Messick, S. (1995). Validity of Psychological Assessment – Validation of Inferences from Person's Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychology, 50,* 741-749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Mintz, J., & Kiesler, D. (1982). Individualized measures of psychotherapy outcome. In P. Kendall & J. N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 491–534). New York: Wiley.

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., & de Vet, H.C.W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63,* 737-745. http://dx.doi.org/10.1016/j.jclinepi.2010.02.006

Nordal, K.C. (2012). Outcomes measurement benefits psychology. *Monitor on Psychology, 43,* 51. http://dx.doi.org/10.1037/e734912011-019

Norman, G.R., Stratford, P., & Regehr, G. (1997). Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. *Journal of Clinical Epidemiology, 50,* 869-879. http://dx.doi.org/10.1016/S0895-4356(97)00097-8

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric Theory*. New York: McGraw-Hill, Inc.

Ogles, B.M (2013). Measuring Change in Psychotherapy. In M.J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (pp. 134-166). New Jersey: John Wiley & Sons, Inc.

Overington, L., & Ionita, G. (2012). Progress Monitoring Measures: A Brief Guide. *Canadian Psychology, 2,* 82-92. http://dx.doi.org/10.1037/a0028017

Palmieri, G., Evans, C., Hansen, V., Brancaleoni, G., Ferrari, S., Porcelli, P., Reitano, F., & Rigatelli, M. (2009). Validation of the Italian version of the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM). *Clinical Psychology and Psychotherapy, 16*, 444-449. http://dx.doi.org/10.1002/cpp.646

Pascal, G.R., & Zax, M. (1956). Psychotherapeutics: Success or failure. *Journal of Consulting Psychology, 20,* 325-331. http://dx.doi.org/10.1037/h0040582

Paterson, C. (1996). Measuring outcome in primary care: A patient-generated measure, MYMOP, compared to the SF-36 health survey. *British Medical Journal, 312,* 1016-1020. http://dx.doi.org/10.1136/bmj.312.7037.1016

Patrick, D.L., Erickson, P. (1993). *Health status and health policy*. Oxford: Oxford University Press.

Petersen, K., Hounsgaard, L., Borg, T., & Nielsen, C.V. (2012). User involvement in mental health rehabilitation: a struggle for self-determination and recognition. *Scandinavian Journal of Occupational Therapy, 19*, 59-67. http://dx.doi.org/10.3109/11038128.2011.556196

Polit, D.F. (2014). Getting serious about test-retest reliability: a critique of retest research and some recommendations. *Quality of Life Research, 23,* 1713-1720. http://dx.doi.org/10.1007/s11136-014-0632-9

Polit, D.F., & Yang, F. (2014). *Measurement and the measurement of change: A primer for health professionals*. Philadelphia: Lippincott Williams & Wilkins.

Prous, M., Salvanés, F., & Ortells, L. (2008). Responsiveness of Outcome Measures. *Reumatología Clínica (English Edition), 6,* 240-247. http://dx.doi.org/10.1016/s2173-5743(08)70197-7

Reese, R.J., Usher, E.L., Bowman, D.C., Norsworthy, L.A., Halstead, J.L., Rowlands, S.R., & Chisholm, R.R. (2009). Using client feedback in psychotherapy training: An analysis of its influence on supervision and counselor self-efficacy. *Training and Education in Professional Psychology, 3,* 157-168. http://dx.doi.org/10.1037/a0015673

Roberts, G., Davenport, S., Holloway, F., & Tattan, T. (2006). *Enabling recovery: The principles and practice of rehabilitation psychiatry*. London: Cromwell Press.

Robinson, O.C. (2011). The Idiographic/Nomothetic Dichotomy: Tracing Historical Origins of Contemporary Confusions. *History and Philosophy of Psychology*, *13*, 32–39.

Robinson, S., Ashworth, M., Shepherd, M., & Evans, C. (2006). In their own words: a narrative-based classification of clients' problems on an idiographic outcome measure for talking therapy in primary care. *Primary Care Mental Health, 4,* 165-173.

Rodriguez, M.C., & Maeda, Y. (2006). Meta-Analysis of Coefficient Alpha. *Psychological Methods, 11,* 306-322. http://dx.doi.org/10.1037/1082-989X.11.3.306

Rothwell, P.M., McDowell, Z., Wong, C.K., & Dorman, P.J. (1997). Doctors and patients *don't* agree: Cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. *British Medical Journal, 314,* 1580–1583. http://dx.doi.org/10.1136/bmj.314.7094.1580

Rottger, M., Rubel, J., & Lutz, W. (2011). What do therapists do with feedback? Results of a German feedback study. Paper presented at the 42nd International Meeting of the Society for Psychotherapy Research, Bern, Switzerland.

Sales, C., & Alves, P. (2012). Individualized Patient-Progress Systems: Why We Need To Move Towards a Personalized Evaluation of Psychological Treatments. *Canadian Psychology, 2,* 115-121. http://dx.doi.org/10.1037/a0028053

Sales, C., Gonçalves, S., Fragoeiro, A., Noronha, S., & Elliott, R. (2007). Psychotherapists Openness to Routine Naturalistic Idiographic Research? *Mental Health and Learning Disabilities Research and Practice, 4* , 145-161. http://dx.doi.org/10.5920/mhldrp.2007.42145

Sales, C., Moleiro, C., Evans, C., Alves, P. (2012). Versão Portuguesa do CORE-OM: tradução, adaptação e estudo preliminar das suas propriedades psicométricas. *Revista de Psiquiatria Clínica, 39,* 54-59. http://dx.doi.org/10.1590/S0101-60832012000200003

Skre, I., Friborg, O., Elgarøy, S., Evans, C., Myklebust, L., Lillevoll, K., Sørgaard, K., & Hansen, V. (2013). The factor structure and psychometric properties of the Cinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM) in Norwegian clinical and non-clinical samples. *BMC Psychiatry, 13,* 99-112. http://dx.doi.org/10.1186/1471-244X-13-99

Sorenson, R., Gorsuch, R., & Mintz, J. (1985). Moving targets: Patients' changing complaints during psychotherapy. *Journal of Consulting and Clinical Psychology, 53*, 49-54. http://dx.doi.org/10.1037/0022-006X.53.1.49

Storm, M., & Edwards, A. (2013). Models of User Involvement in the Mental Health Context: Intentions and Implementation Challenges. *Psychiatry Quarterly, 84,* 313-327. http://dx.doi.org/10.1007/s11126-012-9247-x

Stratford, P.W., Binkley, J., Solomon, P., Finch, E., Gill, C., & Moreland, J. (1996). Defining the minimum level of detectable change for the Roland-Morris Questionnaire. *Physical Therapy, 76,* 359-365.

Strauss. M. E., & Smith, G.T. (2009). Construct Validity: Advances in Theory and Methodology. *Annual Review of Clinical Psychology, 5,* 1-25. http://dx.doi.org/10.1146/annurev.clinpsy.032408.153639

Streiner, D.L., & Norman, G.R. (1995). *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford: Oxford University Press. http://dx.doi.org/10.1097/00004356-199112000-00017

Terwee, C. (2014). Responsiveness to Change. In A. Michalos (Ed.). *Encyclopedia of quality of life and well-being research*. Dordrecht, Países Baixos: Springer. http://dx.doi.org/10.1007/978-94-007-0753-5_2512

Terwee, C., Bot, S., de Boer, M., van der Windt, D., Knol, D., Dekker, J., Bouter, L., de Vet, H. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60,* 34-42. http://dx.doi.org/10.1016/j.jclinepi.2006.03.012

Terwee, C., Dekker, F., Wiersinga, W., Prummel, M., & Bossuyt, P. (2003). On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Quality of Life Research, 12,* 349-362. http://dx.doi.org/10.1023/A:1023499322593

Turner-Stokes, L. (2009). Goal attainment scaling (GAS) in rehabilitation: a practical guide. *Clinical Rehabilitation, 23,* 362-370. http://dx.doi.org/10.1177/0269215508101742

van der Linde, J.A., van Kampen, D.A., van Beers, L.W., van Deurzen, D.P., Terwee, C., & Willems, W.J. (2015). The Oxford Shoulder Instability Score; validation in Dutch and first-time assessment of its smallest detectable change. *Journal of Orthopaedic Surgery and Research, 10*, 1-8. http://dx.doi.org/10.1186/s13018-015-0286-5

Wagner, J., & Elliott, R. (2001). *The Simplified Personal Questionnaire.* Manuscript submitted for publication. Department of Psychology, University of Toledo, Toledo, OH.

Ware, J., Brook, R.H., Davies, A.R., & Lohr, K.N. (1981) Choosing measures of health status for individuals in general populations. *American Journal of Public Health, 71,* 620-625. http://dx.doi.org/10.2105/AJPH.71.6.620

Ware, J., Kosinski, M., & Keller, S.D. (1996). A 12-item Short Form Health Survey. Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34,* 220-233. http://dx.doi.org/10.1097/00005650-199603000-00003

Williams, J.I., Naylor, C.D. (1992). How should health status measures be assessed? Cautionary notes on procrustean frameworks. *Journal of Clinical Epidemiology, 45,* 1347-1351. http://dx.doi.org/10.1016/0895-4356(92)90195-S

Zill, J.M., Scholl, I., Härter, M., & Dirmaier, J. (2015). Which Dimensions of Patient-Centeredness Matter? – Results of a Web-Based Expert Delphi Survey. *Plos One, 10,* 1-15. http://dx.doi.org/10.1371/journal.pone.0141978

Anexos

Anexo 1: Valores do coeficiente alfa, do CORE-OM e do PSYCHLOPS, para cada amostra

Amostra	CORE-OM	PSYCHLOPS
Amostra1 (POR)	.93	.81
Amostra 2 (POR)	.94	.72
Amostra 3 (ISL)	.95	.86
Amostra 4 (POL)	-	.80
Amostra 5 (RU)	.93	.83
Amostra 6 (RU)	-	.83

Anexo 2: Correlações entre o CORE-OM e o PSYCHLOPS, por amostra, e correlação média

Amostra	Correlação [Intervalo de confiança]	*n*
Amostra 1 (POR)	.56 [.37-.71]	66
Amostra 2 (POR)	.50 [.27-.67]	57
Amostra 3 (ISL)	.71 [.60-.79]	103
Amostra 5 (RU)	.66 [.53-.76]	104
Total	.63 [.53-.71]	330

Anexo 3: Correlações entre os domínios do PSYCHLOPS e do CORE-OM, por amostra

Amostra 1 (POR)

Domínios	CORE-OM Problemas	CORE-OM Funcionamento	CORE-OM Bem-estar	CORE-OM Risco
PSYCHLOPS Problemas	.42	.29	.53	.28
PSYCHLOPS Funcionamento	.32	.327	.44	.18
PSYCHLOPS Bem-estar	.59	.57	.65	.39
PSYCHLOPS Total	-	-	-	.35

Amostra 2 (POR)

Domínios	CORE-OM Problemas	CORE-OM Funcionamento	CORE-OM Bem-estar	CORE-OM Risco
PSYCHLOPS Problemas	.31	.22	.30	.30
PSYCHLOPS Funcionamento	.38	.33	.34	.17
PSYCHLOPS Bem-estar	.46	.46	.54	.41
PSYCHLOPS Total	-	-	-	.37

Amostra 3 (ISL)

Domínios	CORE-OM Problemas	CORE-OM Funcionamento	CORE-OM Bem-estar	CORE-OM Risco
PSYCHLOPS Problemas	.65	.54	.61	.33
PSYCHLOPS Funcionamento	.51	.44	.48	.28
PSYCHLOPS Bem-estar	.72	.62	.68	.42
PSYCHLOPS Total	-	-	-	.39

Amostra 5 (RU)

Domínios	CORE-OM Problemas	CORE-OM Funcionamento	CORE-OM Bem-estar	CORE-OM Risco
PSYCHLOPS Problemas	.54	.48	.49	.20
PSYCHLOPS Funcionamento	.46	.48	.43	.26
PSYCHLOPS Bem-estar	.62	.58	.67	.21
PSYCHLOPS Total	-	-	-	.36