Development of an Efficient

and Broadly Applicable Measure of Case Conceptualization Quality

by

Kevin Capobianco

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Psychology

Waterloo, Ontario, Canada, 2022

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Giorgio Tasca, Ph.D. Department of Psychology, University of Ottawa
Supervisor	Jonathan Oakman, Ph.D. Department of Psychology, University of Waterloo
Internal Members	Christine Purdon, Ph.D. Department of Psychology, University of Waterloo
	Walter Mittelstaedt, Ph.D. Department of Psychology, University of Waterloo
Internal-External Member	Mark Ferro, Ph.D. School of Public Health Sciences, University of Waterloo
Non-voting Member	Marjory Phillips, Ph.D. Department of Psychology, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Case conceptualizations are seen as important to the provision of effective and efficient psychological interventions. In creating a case conceptualization, a clinician summarizes information obtained during a course of psychotherapy into an explanatory model of the causes and maintaining factors for a client's areas of difficulty. In turn, the conceptualization is meant to aid the clinician in better empathizing with their client, it provides a shared framework for understanding the problem, and suggests the most relevant treatment targets to maximize the benefits of the psychotherapeutic work.

Many of these proposed benefits, including positive impacts on treatment outcomes, have not been robustly examined. One factor which may be contributing to this paucity of research is the need for an efficient, psychometrically sound, and broadly applicable method of evaluating case conceptualization quality. Previously developed methods for evaluating case conceptualization quality lack breadth in the types of conceptualizations to which they can be applied (e.g. - only a specific therapeutic orientation or a specific presenting problem), appear somewhat cumbersome to use, require access to very specific resources or materials, and/or have poor or inadequately explored psychometric soundness. These same features may limit how well such measures can be applied to other useful contexts in clinical psychology, such as clinical training and supervision.

To address these issues, a broadly applicable and easily accessible method of evaluating case conceptualization quality was generated and evaluated across a pilot and three subsequent studies. Initial evidence for the scale's high internal consistency, moderate inter-rater reliability, good retest reliability, and construct validity was obtained. The results were also indicative of a need for some modifications to coder training and the coding scheme itself to address areas where inter-rater reliability was somewhat low.

The results from these initial applications of this new measure are discussed in the context of future use by researchers and supervisors in clinical psychology and in the context of how to improve and further validate the measure. Given the purported importance of case conceptualizations to psychotherapy and the unanswered questions about their actual impacts, the results of this research suggest the Broadly Applicable Conceptualizations Quality Scale has the potential to become a useful tool in this field of research and in clinical practice.

Acknowledgements

My first and deepest thanks go to Dr. Jonathan Oakman for his consistent encouragement, humour, patience, critical thinking, and enthusiasm throughout my graduate studies. I also want to thank Dr. Walter Mittelstaedt and Dr. Marjory Philips for helping me significantly refine and improve this research during the planning and writing stages, and for the support and excitement they showed towards this project in general. I also greatly appreciate Dr. Christine Purdon for adding her expertise to the committee. The Psychological Intervention Research Team and the research assistants who helped me plan and conduct these studies also have my many thanks. Adam, your excellent acting skills and interest in psychology were vital.

Beth, Katie, Martyn, and (other) Kevin: calling you friends has been the greatest pleasure of this whole thing. Nick, Julia, and Jerrica: graduate school would have been much drearier without you three, thanks for making it a fun adventure. To Mackenzie, Steffen, Mikaela, Melissa, John, Matt(s), and all my friends, thanks for letting me prattle on about psychology so much and for re-energizing me throughout this journey.

To my parents (including Scott and Rosalind) and sister, you've sent so much support and in so many ways I don't think I'll ever be able to thank you enough. Lauren, you carried so much for our family over this process and you've been a true partner through both the exciting and disheartening times. You have been and will always be amazing. Claire, "I love you equally", you make my life better and brighter every day.

Table of Contents

List of Figures

List of Tables

General Introduction

Thesis Statement

There is a need for an efficient, psychometrically sound, and broadly applicable method of evaluating case conceptualization quality. Most tools currently available for this purpose have poor or limited psychometric soundness, are limited in terms of the types of conceptualizations to which they can be applied, and/or may be impractically long for some purposes (Bucci et al., 2016). This research project was focused on creating a measure to address these limitations, with the hope that it could eventually provide researchers with a robust and adaptable tool for answering longstanding questions around the role of case conceptualizations in psychotherapy; for example, how their quality relates to treatment outcomes. The measure developed in this project may also be useful in clinical supervision and in the training of future psychologists by providing a simple method for supervisors and supervisees to evaluate their case conceptualizations. In this document findings are presented from a pilot and three subsequent studies that tested how this new measure, the Broadly Applicable Conceptualization Quality Scale (BACQS), functioned in several contexts.

What is a Case Conceptualization?

Case conceptualizations are clinical tools that are widely considered to be vital in the provision of psychological services such as psychological assessment and psychotherapy (Beck, 1995; Kanter et al., 2009; Needleman, 1999; Persons, 1989). A general definition for case conceptualizations would be that they are the product of sorting through client information, deciding what is relevant, and then interpreting and summarizing this information into a useful set of hypotheses about what has caused and what is maintaining a client's difficulties (Eells, 2011; Johnstone et al., 2011). It is worth noting that case conceptualization is roughly synonymous with terms such as case formulation, or simply conceptualization or formulation, and that these terms will be used interchangeably in this paper.

Two main functions of case conceptualizations follow from the basic definition above and have often been proposed by clinicians writing about their usefulness. The first is that a case conceptualization reduces an often-complex problem (client difficulties and their causes/maintaining factors) to become more accessible, simple, and easily understood (Benjamin, 2003; Kuyken et al., 2008; 2011); this is thought to result from the clinician selecting or prioritizing the most important information gleaned so far and forming clear hypotheses about how this information links together causally.

The second proposed benefit is that a case conceptualization provides a clinician with guidance for making treatment decisions, usually in order to maximize the efficiency and effectiveness of interventions, and to keep interventions from wandering aimlessly (Hill & O'Brien, 2004; Kuyken et al., 2011; Meichenbaum, 2014). This function stems from the first, as many see a better understanding of the client's difficulties as allowing for better treatment decisions to be made. This may involve a psychotherapist considering what the conceptualization indicates are the most important factors driving the client's difficulties and addressing those in therapy first or most extensively (Johnstone et al., 2011).

Researchers have faced significantly difficulty creating a more comprehensive, prescriptive, or detailed definition of case conceptualization than the one provided above (Ridley et al., 2017). This difficulty is problematic as it highlights the gap between how important case conceptualizations are seen to be in clinical practice versus how well they are understood. Related to this, vague definitions also often hinder research into a topic of interest by contributing to miscommunications between research teams, increasing the difficulty of replicating research, and making it more difficult for researchers to ensure that the constructs being assessed across different experimental designs are the same (in essence, the issues that arise from poor operational definitions).

One barrier to creating a more detailed and comprehensive definition may be that different psychotherapy orientations emphasize different structures or content. For example, the same presenting problem may be understood to have different important driving mechanisms, or a different emphasis on the past versus present, when approached from a psychodynamic versus cognitive-behavioural tradition. Another barrier may be that different therapy modalities may necessitate more or less detail or different foci for conceptualizations. For example, across individual, couple, family, or group therapies the amount of detail that can be gathered for any one client is likely to vary, as may the focus on interpersonal dynamics and relationship factors. Various presenting problems may also dictate what types of information may need to be included to produce an effective conceptualization. An example here would be that psychotic disorders may more often require a therapist to include biological factors and cognitive impacts in their conceptualization versus a conceptualization focused on a needle phobia.

Factors such as a client’s age, ethnicity, gender, sexual orientation, and personal histories may also necessitate unique tailoring of conceptualizations. A client who reports feeling anxious about dating might feel that way due to social anxiety, family norms about dating before marriage, potential discrimination against same-sex relationships, challenges around dating after becoming widowed, experiences of abuse in past romantic relationships, or some combination of all those. For each of these possibilities the degree to which a conceptualization “should” consider cultural components, traumatic experiences, cognitive distortions, or age-related norms is likely to vary. Independently, each of the factors above represents a potential difference in structure, content, or focus across conceptualizations. Together, their interactive effects exponentiate the difficulty of finding a precise and yet also “"one size fits all” definition for case conceptualizations, particularly if the focus is on very specific aspects of conceptualization content or structure.

Flitcroft et al. (2007) conducted a study relevant to this problem of achieving a more detailed consensus definition. They provided 23 therapists with 86 statements considered relevant to a cognitive-behavioural therapy (CBT) formulation of depression, then asked each participant to Q-sort these statements. This involved participants sorting cards with statements on them into piles based on their perceived importance. The result of their research was a three-factor solution, suggesting that across their participants three groups of therapists could be identified with differing priorities for what to conceptualize. One group prioritized "here and now" features of the case, another prioritized "function and process" elements, and the final emphasized client trait elements. As the authors discuss, even in the cognitive-behavioural tradition, and even focused on a single presenting problem, therapists can differ in how they engage in case conceptualization; varying in what they consider to be the most important elements on which to focus.

One reaction to the difficulty of finding a clear and prescriptive definition for case conceptualization in the face of all the diversity to be found in their forms and content has been to adopt a narrower focus. Examples of this include researchers developing methods for evaluating case conceptualizations specific to certain disorders from certain psychotherapeutic orientations, or case conceptualizations solely in a specific population (see Bucci et al., 2016).

Despite the difficulties in obtaining a more comprehensive definition, case conceptualizations remain emphasized across a wide range of geographic regions and a wide range of services in clinical psychology, such as assessment, consultation, and intervention (Bieling & Kuyken, 2003; College of Psychologists of Ontario, 2019; Collie et al., 2008; Hessen et al., 2018; Johnstone et al., 2011; Meier, 2003). They are also considered important to other mental health professions such as psychiatry and mental health nursing (Crowe et al., 2008; Fernando et al., 2013; Toews, 1993). For this research project our focus was on case conceptualizations in clinical psychology and for the provision of psychotherapy, particularly psychotherapy with adults. The BACQS may prove to be useful in other domains, age groups, or across other professions, but assessing every aspect of such a large range of potential applications was beyond the limits of this research project.

Why are Case Conceptualizations Considered Important in Psychotherapy?

Beyond the clinical wisdom suggesting that case conceptualizations serve as important guides for the direction of psychotherapy and make complex client information more accessible to psychotherapists, other benefits of utilizing them have often been proposed. For example, clinicians appear to believe that a solid case conceptualization generated collaboratively with a client can enhance client engagement in treatment as client and clinician come to better understand the problems being addressed (Ridley et al., 2017). Others argue that conceptualizations can enhance a client's hope for change, reduce shame and stigma by explaining the nuances of problem persistence, and enhance clinician empathy and the working alliance by increasing a shared understanding of the problem (Johnstone, 2011; Nezu et al., 2015; Samstag et al., 2004). Given that many of these factors (e.g. the strength of the working alliance, clinician empathy, and client hope) are either believed to be, or have been demonstrated to be related to treatment outcomes (Constantino et al., 2012; Del Re et al., 2012, 2021; Elliot et al., 2018; Flückiger et al., 2021; Irving et al., 2004; Truax et al., 1966), and given the oft proposed argument that a better understanding of the client's problems should enhance treatment efficiency, it is not surprising case conceptualizations are also believed to contribute to therapy outcomes (Benjamin, 2003; Johnstone et al., 2011; Kuyken et al., 2011). Similar arguments have spurred clinicians to increasingly use mental health diagnoses to select and guide their interventions. However, some have argued that diagnoses (which are often included as part of a case conceptualization) lack explanatory power (they describe a problem, not what is maintaining it), do not include enough context to serve as adequate guides for treatment, and that case conceptualizations serve this function better (Restifo, 2011).

Research on Case Conceptualizations and Therapy Outcomes

Despite the theoretical promise, there is limited research evaluating the claim that case conceptualizations contribute to psychotherapy outcomes. Several early studies applied a structured psychodynamic framework to case conceptualizations and investigated possible relationships with outcomes (Barber & Crits-Christoph, 1993; Crits-Christoph et al., 1988). The most often discussed framework in this literature is known as the Core Conflictual Relationship Theme (CCRT; Luborsky, 1977) which tasks clinicians and researchers with examining client provided stories of difficult interpersonal situations and identifying unconscious desires the client had for the interaction, how the client's approach was responded to by the other person, and finally how the client reacted in turn. When clinicians produced and utilized CCRT conceptualizations in their interventions that matched those generated by "CCRT experts" (generated from transcripts of the sessions) this predicted better therapy outcomes (Crits-

Christoph et al., 1988). This has been interpreted as a point of evidence in favour of the case conceptualization-outcome relationship.

In the CBT tradition case conceptualizations are emphasized as key aspects of providing effective interventions (Beck, 1995). This is held to be true whether a CBT therapist operates from a more tailored and idiographic approach (such as Persons' Case Formulation Guided Psychotherapy; Persons, 2006) or from evidence-based treatment manuals, which often emphasize the importance of strategically adapting interventions to best meet a client's needs (Wilson, 1996). Thus, it is not surprising that more recently researchers have been exploring the role and impact of case conceptualizations in CBT outcomes as well.

The small body of research examining links between treatment outcomes and case conceptualizations in CBT has mostly consisted of head-to-head comparisons of treatment outcomes when using tailored or manualized interventions or examining effect sizes of tailored and conceptualization-guided interventions and comparing these to typical effect sizes of manualized interventions. Findings from these studies have been inconsistent, with some suggesting an outcome enhancing effect of case conceptualizations (Ghaderi, 2006), others showing that stricter adherence to treatment manuals is superior (Schulte et al., 1992), others showing that both approaches have equivalent effects (Chadwick et al., 2003; Emmelkamp et al., 1994; Persons et al., 2006), and that immediate outcomes may be similar but maintenance is improved in conceptualization guided interventions (Jacobson et al., 1989). The limited body of research, mixed set of results, and methodological issues (such as small samples or differing presenting problems across studies) together have resulted in a sense that the question of the importance of case conceptualizations to outcomes in CBT is still relatively unanswered (Bieling & Kuyken, 2003; Easden & Kazantzis, 2017).

Beyond treatment outcomes, other research has examined how clients respond to the therapist sharing their CBT case conceptualization with them. For example, Chadwick et al. (2003) found that for some clients in a hospital setting for treatment of psychosis, discussions about their case conceptualizations were somewhat disheartening and upsetting. Other researchers have found strong emotional reactions following the sharing of formulations in a small sample of “difficult to treat” patients (Evans & Parry, 1996). Somewhat contrary to these, in a later study of patients with obsessive compulsive disorder treated with CBT, the working alliance appeared to improve substantially following the assessment phase of therapy and the discussion of the case conceptualization (Nattrass et al., 2015). These findings could suggest that depending on the context, case conceptualizations may have a positive or negative impact on client hope and the working alliance between therapist and client, which may in turn have indirect effects on outcome. Although a sparse body of evidence to draw from, these points highlight the potential importance of creating case conceptualizations that do not dishearten, stigmatize, or damage the therapist-client relationship.

Assessing Case Conceptualization Quality

Although the research above offers some preliminary insights into how case conceptualizations may impact treatment outcomes, more research has been directed at developing methods of assessing case conceptualization quality and the level of skill therapists demonstrate while developing case conceptualizations. Interesting findings have arisen from such research which highlight potential issues in how case conceptualizations are developed and applied in clinical practice. These include findings that many clinicians struggle to produce conceptualizations of “adequate” quality, at least when evaluated by well-regarded and experienced cognitive behavioural therapists (Haarhoff et al., 2011; Kuyken et al., 2005), and that psychotherapists can achieve high rates of agreement for surface elements of a case, but struggle to reach adequate agreement on more inferential aspects of their conceptualizations (Dudley et al., 2010; Kuyken et al., 2005; Persons et al., 1995; Persons & Bertagnolli, 1999).

Though much of the research into case conceptualizations has been from either a psychodynamic or cognitive behavioural perspective, approximately one fifth of psychologists who responded to a survey sent to American Psychological Association Division 12 (Clinical Psychology) members identified their therapeutic approaches as integrative or eclectic. Furthermore, approximately 14% identified with orientations aside from integrative, cognitive, behavioural, or psychodynamic (Norcross & Karpiak, 2012). Touching again on the consensus definition issue from earlier, several authors have gone against the trend of focusing on very specific forms of conceptualization, instead suggesting that there may be more commonalities than differences across conceptualizations and that a broader perspective on them that can span orientations is preferable (Butler, 1998; Goldfried, 1995). At the very least, the mismatch between the predominant focus on CBT and psychodynamic conceptualizations and the proportion of psychologists who do not operate from these highlights the value of research tools that follow a more trans-theoretical operational definition for case conceptualizations.

Strengths and Weaknesses of Existing Measures of Case Conceptualization Quality

Bucci et al. (2016) conducted an extensive search of the available case conceptualization literature and identified eight conceptualization quality measures which met their inclusion criteria of being published in English, containing a measure of psychological case conceptualization, and intended for use with mental health or forensic populations. Some measures they identified have been designed for use in real practice with current cases while others are meant to evaluate conceptualizations generated in response to predetermined case vignettes (sometimes in video form) presented to clinicians. Some of the measures focused more on content and quality of conceptualizations, while others focused more on the process used in arriving at a conceptualization.

The authors evaluated the measures according to four main criteria: ease of administration, generalizability (the narrowness or breadth of contexts in which the measure could be used), whether and how the measures assess conceptualization reliability and validity, and the psychometric qualities of the measures themselves. They go on to discuss the shortfalls and trade-offs in the measures on these four criteria, how these may explain the overall paucity of research in the field of case conceptualization research, and the relevance of these points to future directions for the field.

Few of the methods reviewed by Bucci et al. (2016) appear particularly practical for use outside research contexts, and perhaps even in some research contexts, due to a variety of barriers to ease of administration. Several methods require access to unique stimuli, such as specifically developed video simulations of psychotherapy sessions (Dudley et al. 2010), or unpublished scoring manuals. Other measures are highly detailed, translating into a significant time investment – both for training coders and actual coding of conceptualizations (Eells et al., 1998). Yet other measures have technological or practical barriers such as requiring coders to watch either live or recorded psychotherapy sessions (Padesky et al., 2011).

Bucci et al. (2016) also found that many of the measures trade off broad applicability for a more niche application. Examples of this include measures designed solely for assessing case conceptualizations of clients experiencing obsessive-compulsive disorder (Zivor et al., 2013), or a measure assessing cognitive behavioural conceptualizations generated from a video vignette of a client experiencing delusions (Dudley et al., 2010). Furthermore, six of the eight measures focus specifically on cognitive-behavioural conceptualizations. Although likely to fit certain researchers' needs very nicely, the authors suggest it may be difficult for others to adapt such specific measures for their own purposes. This narrow applicability also limits the degree to which such measures could be deployed by actual psychotherapists or clinical supervisors, as many practicing or training psychologists are faced with a diversity of client diagnoses, comorbidity is common (Van Loo et al., 2013; Zimmerman et al., 2008), and as noted above many therapists approach psychotherapy from perspectives other than CBT.

The reviewers also note that several of the case conceptualization quality measures have not demonstrated particularly encouraging psychometric properties, for example having low inter-rater reliability, no published evaluation of test-retest reliability, or lacking adequate tests of convergent, discriminant, or predictive validity. Compounding this is the fact that many of these measures have only been validated on somewhat homogeneous and small samples of conceptualizations. The shaky psychometric foundations upon which several of these measures rest may be contributing to the difficulties researchers appear to be having in addressing outstanding questions about case conceptualizations in psychotherapy; perhaps the field is having difficulty moving past the barrier of creating a well validated and reliable measure.

Although the main thrust of their review appears to be that the extant case conceptualization measures are limited by the significant issues noted above, three are noted by the authors as being the most promising or the most widely tested and used to date. These three are the Case Formulation Content Coding Method (CFCCM: Eells et al., 1998), the Collaborative Case Conceptualization Rating Scale (CCC-RS: Padesky et al., 2011), and the Case Formulation Quality Checklist (CFQC; McMurran et al., 2012). However, these most promising measures still come with noteworthy limitations.

For example, the CFCCM has a strength in that it appears to have been designed to be applicable across various psychotherapeutic orientations and has demonstrated very high inter-rater reliability. The CFCCM also appears comprehensive in that it evaluates the relative presence or absence of over 40 different categories of possible information, rates a number of specific features in the conceptualization for quality, and provides several ratings of quality for aspects of the conceptualization as a whole.

This comprehensiveness is not itself a negative, and again in some contexts could match a researcher's needs very well. However, the reviewers note this comprehensiveness may also translate into significant investments in time and effort and may not be ideal for routine clinical practices. Additionally, when considering this comprehensiveness in the context of findings that some content areas appear very infrequently utilized by therapists in conceptualizing their cases (an interesting, and somewhat troubling finding in itself: Eells et al., 1998; Capobianco, 2015) the CFCCM may be somewhat unnecessarily cumbersome.

The CCC-RS is discussed in the review as the measure which has demonstrated the highest reliability and validity in the context of evaluating live or recorded CBT sessions. In particular, this measure demonstrated good internal consistency, inter-rater reliability, and some convergent validity with a measure of CBT therapist skill in a study evaluating the psychometrics of the measure (Kuyken et al., 2016), and even demonstrated some relationship to therapy outcomes in a doctoral dissertation utilizing the scale (Gower, 2011). Additionally, this measure's creators have helpfully published the coding manual online and encourage other interested researchers to collaborate with their team in the development of this measure.

However, as with the CFCCM, the CCC-RS also requires a significant investment in resources and time, impacting the ease of administration of the measure. First, the measure requires the resources to record or observe live therapy sessions as coders are tasked with watching sessions and rating therapists on 16 different items falling into four subscales. Some of these items address complex and nuanced psychotherapy processes that may be difficult for coders to evaluate without significant training or experience with psychotherapy. For example, items which require a rater to evaluate whether the most appropriate CBT model has been selected as a base for conceptualizing the client's difficulties, or which ask the rater to evaluate whether the therapist has linked a client's goals, the session agenda, and the conceptualization in seamless ways. The measure is also characterized by a limited scope and low generalizability; it has been explicitly designed to evaluate CBT conceptualizations. Additionally, this measure has an emphasis on the process of developing a case conceptualization by asking coders to rate the level of collaboration present, whether the conceptualization is iteratively tested and adjusted over time, and whether the conceptualization deepens over time. While these process factors are very interesting and have promise in terms of relating to therapy outcomes, this measure may not precisely answer questions about the quality of a conceptualization in itself.

The CFQC (McMurran et al., 2012) differs from the CFCCM and CCC-RS in that it is noted for high ease of use. The measure utilizes a scoring template which helps guide coders and has a clear organization and structure. The scale also appears to have adequate psychometric properties in several domains such as inter-rater reliability, test-retest reliability (after one week), and a high internal consistency. In terms of limits, the authors of the review suggest the measure may not adequately emphasize the integrating, summarizing, and sense-making aspects of case conceptualization. Additionally, to our knowledge the measure has so far only been validated on a small sample of forensic case vignettes, and so the evidence of psychometric soundness may not hold in other contexts. In terms of generalizability, the measure was developed for use by professionals working in the forensic system and with offenders and it is unclear how it would fare when utilized in other settings.

Modified versions of both the CCC-RS and the CFCCM were used together in previous research evaluating case conceptualizations in a psychology training clinic setting (Capobianco, 2015). In this research, a sample of psychotherapy reports were obtained, the case conceptualizations in them were coded for quality, and these ratings were correlated with indicators of treatment progress and the strength of the therapeutic alliance. To generate the ratings of conceptualization quality modified versions of the CCC-RS and CFCCM were applied by a coding team. Some of the findings and reflections from executing that research project are relevant to the points raised above on the strengths and weaknesses of existing measures. This includes that we also found the coding schemes we utilized were somewhat cumbersome to apply due to their length, particularly as some content categories of the CFCCM were regularly absent from almost all conceptualizations. We also found that applying aspects of the modified CCC-RS to non-CBT conceptualizations was difficult, again reinforcing the value of a broadly applicable measure.

Additionally, that study highlighted the need to obtain a broader sample of conceptualizations (from beyond just one psychotherapy clinic) to avoid developing a tool of only local utility. Finally, we found that many of the reports focused more on diagnostic impressions rather than a detailed case conceptualization or included incomplete conceptualizations (as the reports they were drawn from were more focused on articulating broad treatment plans and/or treatment progress). This highlighted the need to find methods of obtaining more complete and detailed conceptualizations in future research.

Common Features of Previous Measures

As previously discussed, the variation in content and structure of case conceptualizations (influenced by factors such as clinician training, presenting problems, client characteristics, etc.) has made a consensus on what they should contain, or how they should be structured, difficult to achieve. However, there does appear to be some agreement on broader points of what might be useful markers of conceptualization quality. This agreement is reflected in common elements and items included across several of the case conceptualization evaluation schemes.

An easily identified common element across several case conceptualization quality coding schemes is a rating to evaluate the overall quality of the conceptualization. This item usually tasks raters with considering some combination of other more specific elements included in that scheme. Overall quality items allow for quick and intuitive judgments to be made by raters and also represent a holistic sense of the strength and usefulness of the conceptualization. The CFCCM, the Quality of Cognitive Formulation Rating Scale (Fothergill & Kuyken, 2002), CFQC (McMurran et al., 2012), and Rating the Quality of Case Formulation for Obsessive-Compulsive Disorder (Zivor et al., 2013), are examples of conceptualization quality schemes that include overall scores.

A second common element is that several quality measures include items which assess the degree to which ideas have been integrated and organized, and the presence of a narrative (versus a series of independent facts) in a conceptualization. These items ask raters to score how well the information in the conceptualization has been synthesized into a meaningful account, how ordered and categorized the information in the conceptualization is, and how well pieces of information are tied together to create a larger narrative. In the CCC-RS this idea is represented by the items of the Levels of Conceptualization subscale (Padesky et al., 2011). This subscale emphasizes that over time case conceptualizations should become deeper and more explanatory versus being simply descriptive. In this way a higher quality case conceptualization process moves the conceptualization from describing features without much integration to integrating client information into a more meaningful structure.

The CFQC and the CFCCM have items with a similar focus. The CFQC's item titled "generativity" asks raters to judge how much the conceptualization goes beyond simply being a description or statement of facts into making predictions. It also has an item on narrative, asking raters to judge how much the conceptualization tells a "coherent, ordered, and meaningful story", and an item on explanatory breadth (which is also relevant to the next common feature described below) which involves rating the degree to which a variety of information has been tied together. Finally in the item on "diachronicity" raters evaluate how well information from the past, present, and future are tied together. The CFCCM item "Degree of Inference" asks to what degree the formulation goes beyond descriptive information offered by the patient, with higher scores representing more hypothetical considerations from the clinician.

The third common feature identified across the available conceptualization schemes focuses on the breadth and diversity of information in a conceptualization. In some cases this emphasis is demonstrated in items assessing whether the conceptualization as a whole has a variety of ideas from different perspectives or categories. In others, this focus is seen more broadly and in the structure of a measure itself. For example, some coding schemes assess a conceptualization for the presence or absence of a variety of different pieces of information. Also relevant to this element of conceptualization measures are instances where openness to new information and effective management of complexity and detail are assessed (as these also appear linked to the conceptualization's breadth and diversity of information).

Some measures include items which address breadth and integration simultaneously (to what degree the conceptualization has a variety of information and also links this together), while others separate out items on breadth. Examples of the former include the explanatory breadth item of the CFQC and the complexity item of the CFCCM. Examples of the latter can be seen in the CCC-RS levels of conceptualization subscale. For example, this subscale has several items that focus on uncovering more information throughout sessions, the clinician having a curious mindset during sessions, the clinician asking detailed questions, and the clinician being open to unexpected client responses.

As mentioned above, other measures reflect an interest in breadth by assessing the presence or absence of several types or categories of information (such as whether the conceptualization considers thoughts, feelings, actions, core beliefs, client background, history, or includes each of the “five Ps” – presenting problem, predisposing factors, precipitating factors, protective factors, perpetuating factors, etc. (Haarhoff et al., 2011; Page et al., 2008). The variety of content categories of the CFCCM is also another example of this. The scale assesses conceptualizations for the presence of specific types of information representing biological, social, and psychological domains. Breadth is typically seen as a positive feature provided that it is balanced with parsimony and the trimming of extraneous information (Sim et al., 2005). The importance of breadth is also echoed in arguments that conceptualizations should be person-specific, not problem specific, should consider multiple models and perspectives, and should consider broader contextual factors of a case (Johnstone et al., 2011).

Empirical basis in psychological science and evidence of logical thinking is a fourth common feature of conceptualization measures. Several measures ask raters to evaluate whether the conceptualization is consistent with a currently supported nomothetic model of psychopathology, or whether the conceptualization is logically consistent and does not self-contradict. For example, the Case Formulation Quality Checklist evaluates conceptualizations for theoretical coherence -the degree to which the conceptualization is in line with established psychological research into human psychopathology and behaviour (for this measure, forensic psychological theories in particular). Additionally, the measure assesses conceptualizations for internal coherence, or a lack of logical contradictions. Both these features appear to be indicative of the perceived importance of careful reasoning and basing conceptualizations in evidence and fact. Other measures (e.g. the CCC-RS) approach this idea by emphasizing a skeptical and scientific mindset while conceptualizing. This measure emphasizes the testing of conceptualizations for accuracy, making predictions from the conceptualization, and maintaining an openness to evidence disconfirming the conceptualization. Items such as these tap into core clinical psychological ethical principles around providing responsible care (Canadian Psychological Association, 2017), broader values inherent in the scientist-practitioner model of clinical psychologists (Frank, 1984), and in the CBT approach of linking client problems with broader and well validated models of mental health disorders (Persons, 2006).

Several measures assess whether a conceptualization has identified a psychological mechanism driving or maintaining the client’s difficulties, whether the conceptualization helps to highlight important aspects of the client’s case and helps prioritize treatment targets. This appears to be a fifth common element across measures. These items arguably tie closely with the main functions conceptualizations serve that were highlighted previously: enhancing and guiding treatment. This idea is represented in the item of the CFCCM on the identification of a mechanism causing the client’s difficulties, or in the CCC-RS item on the conceptualization guiding the intervention by selecting the most key maintenance processes to be addressed, or in the "action-oriented” item of the CFQC which focuses on how well the formulation prioritizes information and helps plan treatment.

Finally, even though it is only clearly present in a few quality rating schemes, a sixth feature which appears important to conceptualization quality is the degree to which the conceptualization focuses on client strengths and resiliency or has a positive and encouraging tone. As mentioned above, in some circumstances a conceptualization shared with a client might be disheartening (Chadwick et al., 2003; Evans & Parry, 1996), and other research hints at links between strengths and resiliency in conceptualizations and positive treatment outcomes (Welfare et al., 2013). There are also broader findings emphasizing the importance of clinician and client hope in predicting positive outcomes (Coppock et al., 2010; Irving et al., 2004). Indeed, Milovanov (2017) found that in psychotherapy (and alternative medicine interventions), the greater the degree to which a client believed in, and had hope stemming from, the treatment rationale (in essence the argument for how a case conceptualization links to treatment outcome) the more positively they perceived their treatment outcomes to be. Altogether, it appears that the presence of a positive and hopeful tone, and the degree to which strength and resiliency have been incorporated should be considered in any conceptualization quality measure seeking to predict or contribute to therapy success.

Examples of items relating to strengths and resiliency are found in the items of the CCC-RS subscale aptly titled “Strengths and Resilience”. The CCC-RS manual describes this subscale as evaluating the degree to which a therapist helps draw out and conceptualize client strengths and examples of resilience, and how these are used to enhance the therapy and increase client engagement. The second can be seen in the CFCCM content categories for client strengths and resiliency, which include adaptive aspects of the client, positive social support, or strong motivation for treatment. Including elements such as these is likely to help bolster client hope and confidence in change, provide some practical tools for leveraging positive changes, and also emphasize the clinician's own hope for their client to experience positive change.

To summarize, through examining the available case conceptualization quality measures these six common elements were identified as having some consensus and rationale as markers of conceptualization quality. Higher quality conceptualizations should have an overall face-value quality; should integrate client information to produce a meaningful, structured, and organized understanding; should be broad and comprehensive without straying into unnecessary complexity; should suggest a priority target for treatment by highlighting key factors contributing to the client's difficulties; should include client strengths and resiliencies in order to foster hope and reduce stigma and shame; and should be logical and based in established psychological theory. These also appear to correspond well to the proposed benefits and functions of utilizing case conceptualizations in psychotherapy.

Psychological Mindedness and Case Conceptualization

There appears to be a significant conceptual overlap between the act of case conceptualization and thinking with psychological mindedness. Psychological mindedness is a trait-like construct (in that it is presumed to be relatively stable over time) which represents the degree to which a person is capable of attending to and using psychological information to make sense of their own and other's behaviors and experiences (Beitel et al., 2005; Daw & Joseph, 2010). Psychological mindedness can also be thought of as the ability to extract psychological information from situations and the ability to make sense of that information by noting patterns or relationships (Daw & Joseph, 2010). Psychological mindedness is also demonstrated when a person is open to new ideas in order to better understand themselves or others, when a person has an awareness of feelings and emotions, and has a reflective approach to understanding people (Beitel et al., 2004; Farber, 1985; McCallum & Piper, 1996a; Trudeau & Reich, 1995). Psychoanalytic psychotherapists have long considered *client* psychological mindedness to be an important predictor of success in psychoanalytic therapy (Bachrach & Leaff, 1978). For our purposes; however, it may be more important to consider the role of *therapist* psychological mindedness, and how this manifests in the level of psychological mindedness *demonstrated* in a case conceptualization.

The relevance of psychological mindedness to clinical psychologists may be seen quite early in a clinician's career. It has been posited as one factor leading some to choose careers in this field (Farber et al., 2005); those who are interested in understanding how psychological processes impact behaviour are presumed to be predisposed to the profession. Students in psychology have also been found to have higher levels of psychological mindedness than students in natural science programs or other social sciences (Trudeau & Reich, 1995; Westen et al., 1991) which adds further support to there being some relationship between psychological mindedness and interest in psychology.

However, this relationship may not simply be a matter of a trait predisposing individuals to studying psychology. It may also be that this construct is a skill which can be developed and improved. In psychiatry for example, some argue that increasing the psychological mindedness of residents should be emphasized during their training (Cogburn, et al., 2022), and improvements in psychological mindedness was found to be possible through deliberate training provided to nurses (Saito et al., 2017). More relevant yet, is some research indicating that psychological mindedness amongst a sample of clinical psychology graduate students does appear to increase throughout their training (O'Brien, 2001). These latter points about psychological mindedness may highlight the relevance of viewing it as a form of competency in the field of clinical psychology, particularly if this competency results in improvements to case conceptualizations (a possibility which is returned to later in this section).

In Westen et al.'s (1991) study (mentioned above), the method of determining participant psychological mindedness is of particular interest to this research project. Participant responses to a variety of tasks (being asked to describe themselves, provide stories for two thematic apperception task cards, reflect on those stories and how they might demonstrate something about themselves, explain the rise in teen pregnancy, and explain why some have difficulty with finishing their dissertations) were scored on criteria meant to assess psychological mindedness. Lower scores were given for glib responses or those with a focus on physical or concrete traits or behaviors, and higher scores were given when there was a focus on sophisticated psychological processes and complex personal traits.

Other researchers have used the concept of psychological mindedness to explore explanations of mental health difficulties. McCallum and Piper (1990) generated a procedure for evaluating psychological mindedness by presenting role-played clinical scenarios to participants through video recordings, and then asking them to describe what was troubling the clients shown. Similar to the ratings given in Westen et al.'s (1991) study, responses were scored on a scale meant to assess psychological mindedness. Low scores were given when only a single, simplistic internal experience of the client was identified, and the highest scores when the participant identified complex, sometimes unconscious, psychological processes and related these to the difficulties presented in the video. The researchers obtained good inter-rater reliabilities across several studies for this scale, and found scores correlated with self-report measures of psychological mindedness.

Although not initially framed as rating psychological mindedness, another study approached the examination of explanations for mental health difficulties in a similar manner. The complexity of explanations for mental health difficulties (what appear to essentially be case conceptualizations) generated by psychologists, general practitioners, and lay persons were evaluated, and the psychologists were found to have those with the highest complexity (Cape et al., 2008). The specifics of the coding scheme used to evaluate "explanation complexity" appear to correspond very closely with how psychological mindedness was evaluated in the studies above. Low ratings of complexity were given to explanations of mental health difficulties that were limited to global statements of feelings, diagnoses, or circumstances, whereas high levels involved linking internal psychological processes with other more concrete elements. In discussing their results, the authors even hypothesized that the higher complexity of explanations provided by psychologists could be attributed to higher psychological mindedness in that group.

As seen above, case conceptualization quality and psychological mindedness appear to share some conceptual overlap. Indeed, more recently, Hartley and colleagues (2015) have found some empirical support for a relationship between the two. They found that clinicians with higher psychological mindedness produced case conceptualizations that more closely corresponded to an expert's conceptualization, and which were rated as higher in quality. In that study psychological mindedness was not assessed through self-report, but through coding participant generated responses.

Given that participant generated responses, particularly those related to explanations of mental health difficulties, have previously been coded for demonstrated psychological mindedness also suggested the possibility of assessing the level of psychological mindedness evident in a case conceptualization. Evaluating the demonstrated psychological mindedness in a conceptualization could be a way of capturing the depth, abstraction, openness to psychological complexity, and emphasis on psychological processes that appears to be a unique aspect of thinking like a psychologist and which may be underemphasized in other conceptualization quality measures, or at least which could be missed by the previous shared features identified.

Research Objectives

The broader goal of this doctoral research project was to develop a brief and accessible measure of case conceptualization quality which can be applied to a wide range of conceptualizations and which demonstrates good psychometric performance in contexts analogous to those where it could eventually be applied. To this end, an initial pilot and three subsequent studies were conducted. The first focused on demonstrating the Broadly Applicable Conceptualization Quality Scale's (BACQS's) psychometric performance on a larger sample of conceptualizations from a set of participants that may be roughly analogous to junior clinical psychology graduate students (which may be one possible application of the BACQS in the future). The second focused on demonstrating that the BACQS could be applied to conceptualizations beyond those in written form, in an effort to both show broader applicability and as an analogy to the drawn conceptualizations which are sometimes used in clinical supervision settings or in sessions with clients. The third study focused on applying the scale to conceptualizations from a range of different therapeutic orientations and for various presenting problems, again to demonstrate broad applicability. This last study also allowed for a closer examination of how the scale would perform on more ecologically valid conceptualizations, as they had actually guided courses of psychotherapy and had been generated by real clinicians.

Before exploring the methods and results of these studies the following section outlines the steps that went into the scale's creation and how these relate to its content and face validity. The items and structure of the BACQS itself is then outlined, followed by the training used to prepare the coders who went on to utilize it in the studies presented.

Content Validity and Face Validity

The selection and development of the items and coding manual for the BACQS were informed by the research reviewed in the general introduction. Seven items were generated (outlined in more detail in the next section) to capture the best available definition for case conceptualization (a tool that organizes the information from a psychotherapy case that helps guide treatment), how conceptualizations are meant to benefit therapy, and the main functions they are intended to serve. This provided ways of differentiating levels of quality in a conceptualization based on: how well a conceptualization fits with their definition, how well a conceptualization fits with their purported purpose, and how well a conceptualization could be expected to provide the benefits many argue they can.

Additionally, six of the seven items reflect commonalities across previous case conceptualization quality measures. By reviewing those measures and incorporating those common features/ideas the BACQS could remain brief while still assessing case conceptualizations for the features that appeared most agreed upon as important markers of quality. The remaining seventh item focused on how psychologists appear to demonstrate particularly high psychological mindedness, and how this construct appears closely related to the task of psychological case conceptualization. Altogether, these may be good reasons to suspect the BACQS has high content validity - the measure reflects much of the current literature relevant to case conceptualizations and the current arguments for how to evaluate their quality.

In terms of face validity, after generating the items and the coding manual describing them in more detail a series of consultations were held where the BACQS was presented to registered clinical psychologists and clinical psychology graduate students with diverse areas of clinical expertise. After reviewing the BACQS all agreed the measure appeared promising, did not appear to be missing any major dimensions for assessing quality (given the goal of having a broadly applicable measure), and all endorsed the opinion that it could prove useful in the contexts of clinical supervision and research. Given this positive feedback from experienced clinicians and researchers the BACQS could be considered to have demonstrated initial face validity.

The Broadly Applicable Conceptualization Quality Scale

The BACQS assesses case conceptualization quality through seven items: 1) overall quality impression, 2) psychological mindedness and depth, 3) integration, 4) differentiation/breadth, 5) core mechanism identified, 6) strengths-resiliency focus, and 7) theoretical and logical grounding. Scores on each item range from one to five; where one represents a very weak, poor, or severely inadequate quality for that dimension, a score of three is meant to represent adequacy, and a score of five represents outstanding quality, skillful execution, and very high expected utility. The full scale and more detailed scoring instructions for each item can be seen in Appendix 1 of this paper, but a brief outline of each item is provided below.

General quality impression: This item is obtained after an initial examination of the conceptualization without the rater "over thinking" their rating. The rating is meant to reflect an overall evaluation of the expected usefulness of the conceptualization based on the basic definition that conceptualizations should enhance a clinician's understanding of a client, should simplify a complex problem, and should guide and enhance a course of psychotherapy.

Psychological mindedness and depth: This item is scored based on the depth of thinking evident in the case conceptualization, particularly in terms of how effectively psychological constructs and intra-psychic factors are considered to explain the client's difficulties. For example, a low score would reflect a surface level explanation that oversimplifies the client's problems (i.e. "the client is lazy") or which refers solely to external factors ("the client got fired from work so they are sad"). Higher scores reflect an understanding of more complex and nuanced psychological factors working to cause the client's difficulties. This dimension is meant to be applicable across a variety of therapeutic orientations and so while the item gives examples for "complex and nuanced factors" (noting mixed-emotions, complex cycles of thoughts-feelings-behaviours, internal conflict, etc.) it does not seek to provide a comprehensive list.

Integration: This dimension reflects the degree to which the conceptualization is tied together into a meaningful whole versus appearing to be a list of standalone ideas. To count as more highly integrated, the links should not only represent what appear to be client provided explanations of concrete facts. For example, if a client expressed that "after I moved to the new city I found my anxiety increased", and the conceptualization included the idea that moving between cities makes the client anxious this does not really provide an explanation for what is happening at a useful level. Instead, the clinician would have to link the information provided by the client to some other idea, perhaps hypothesizing that the client has difficulty entering into new social situations and forming new relationships, which left them more isolated and bored after the move, resulting in more time ruminating and greater anxiety. The highest levels of this dimension reflect a conceptualization where most ideas presented are linked with at least some other part of the conceptualization and where an overall narrative, structure, or ordered clustering of ideas is clearly present. Part of the rationale for this item is that "experts" in various domains, including in psychology, have typically been found to organize their ideas with more clustering and ordering, and so this may also put this item in line with developmental trajectories of expert level skill (Bordage & Lemieux, 1991; Cummings et al., 1990; Gong et al., 2015).

Differentiation/breadth: This item evaluates the conceptualization for comprehensiveness of information and on attempts of the author to take more than one perspective on the case. For example, a high scoring conceptualization would leave an impression that most possibly important types of information have been considered, and that the creator of the conceptualization is not overly fixated on just one main idea. As such, having a great deal of information on just one aspect of the case while ignoring other potentially important perspectives would not result in a very high score.

Core mechanism identified: This item evaluates whether a conceptualization successfully helps the therapist prioritize information in order to guide the course of therapy. A low scoring conceptualization would not appear to have any central features or ideas, with every idea being presented with equal weight. A high scoring conceptualization would note a central psychological mechanism, or a few, which are driving a large portion of the client's difficulties. Given that the main utility of having a central mechanism described is so that it can be targeted for some form of therapeutic change, this item requires that the central mechanism can be addressed in psychotherapy (for example, a certain gene may be an important driver of some mental health issues but it cannot be changed by psychologists) and that if this mechanism was resolved it would be expected to significantly improve the client's difficulties. As an example to illustrate these points, an effective conceptualization might highlight how behavioural deactivation, sleep issues, and negative beliefs about self-worth each may be contributing to a client's depression while emphasizing one of those as the most important mechanism ("the client needs to challenge their negative beliefs about their worth as they have repeatedly avoided engaging in behavioural activation due to thoughts that they don't deserve to feel better").

Strengths-resiliency focus: Given an increasing emphasis on the importance of including client strengths in treatment and conceptualization, and the known importance of hope in therapy outcomes, this item evaluates whether the conceptualization leans more towards disheartening and stigmatizing a client, versus bolstering confidence and easing shame. Low scores represent a conceptualization that harshly blames a client or uses words or ideas that could be unnecessarily hurtful. High scores are for conceptualizations which have no significant faux pas such as those, and which incorporate client strengths and resiliency, or which help to reduce blame and shame, and which describe the client with a sense of positive regard and hope for the future.

Theoretical and logical grounding: This item reflects the value of relying on established theories in psychotherapeutic traditions when conceptualizing a case. The lowest scores reflect a conceptualization that lacks even basic logical consistency and which may include pseudoscientific ideas. Moderate scores demonstrate some familiarity with theories of mental health and well-being in psychotherapeutic traditions and contain no obvious faults in logical consistency. The highest scores reflect a skillful, carefully considered, and logical integration of client information with broader theories relevant to psychotherapy and scientific understanding of human well-being and psychopathology.

Coder Training

Undergraduate level research assistants were trained to apply an earlier version of the BACQS which did not include item seven (as at the time it was unclear how well these coders would fare in identifying adherence to established theories in psychotherapy). The coders were selected based on their interest in taking part in this research project. Undergraduate research assistants are important members of many psychology research labs and are often delegated tasks such as coding. As such, assessing the performance of the BACQS with a coding team largely comprised of undergraduate level coders appeared useful, particularly with the goal of keeping the measure easy to use and resource-light (versus requiring master's level or higher coders).

Training involved four one-hour meetings. The first stage involved providing foundational knowledge about case conceptualizations, psychotherapy, psychotherapeutic orientations, and how these relate to the BACQS while also reviewing principles of coding in psychological research (coding independently, aiming to code each item without being influenced by halo effects, avoiding drift by referring to the coding manual regularly). The research assistants were also provided with several articles to review from the field of case conceptualization research and the author made himself available to the coding team to answer any questions related to the topics above. Next, the BACQS was reviewed as a group in more detail. Each item was discussed along with examples for how low, medium, or high scores might manifest in a conceptualization. The coding team then independently practiced coding ten fictitious conceptualizations. Following this, scoring disagreements were noted and resolved through discussion and by referring to the coding scheme. Once this was complete the coders appeared familiar enough with the BACQS to move on to our pilot testing.

Pilot Study: Performance of the BACQS on Fabricated Conceptualizations

A preliminary assessment of the BACQS was needed in advance of committing more resources and time to a larger scale study. To serve this need the BACQS was applied to a small sample of fictitious conceptualizations and the scale's inter-rater reliability, internal consistency, and inter-item correlations were assessed. Our primary aim was to evaluate whether the BACQS and our coders could perform well enough to move forward to the next stages of this research project.

Aims and Hypotheses

The main aim of this pilot study was to demonstrate that the coder training and the revisions to the BACQS coding scheme had readied the coding team for independent coding of conceptualizations while still obtaining high inter-rater reliability. Items of the BACQS could also be tested to ensure they were achieving at least moderate inter-item correlations and good internal consistency.

Finally, evidence for test-retest reliability was sought, to help determine whether the coding scheme allowed for consistent scoring on repeat applications. However, in this and the subsequent studies a full and robust assessment of the scale’s test-retest reliability was not possible due to practical constraints. For example, it was not feasible to ask the full team of volunteer coders to recode sets of conceptualizations a second time. It was also not feasible to have the authors of the case conceptualizations produce a second conceptualization based off the same materials/cases at a later time, and then have those conceptualizations recoded. Had both been possible, this would have more effectively satisfied the conditions required for a true examination of the BACQS’s test-retest reliability. Given that the method used across studies was for one coder (myself) to recode the same conceptualizations at two times and then correlate the scores from my original and second coding, these more limited tests of consistency are referred to as “recode reliability” in the remainder of the document.

Our main hypotheses were:

1) The BACQS would demonstrate adequate inter-rater reliability for the (then) 6 items and the scale total score.

2) The BACQS would demonstrate high recode reliability.

3) The BACQS would demonstrate high internal consistency and the (then) 6 items would relate to each other in a coherent pattern suggestive of a scale assessing overall conceptualization quality.

Methods

Procedure

Twenty fictitious case conceptualizations were written by the author, ranging from 66-311 words in length (*M* = 208, SD = 69.91) (see Appendix 2 for several examples). The cases were designed to represent an assortment of presenting problems (e.g., anxiety, depression, personality disorders, eating disorders), and were designed with the goal of having a high variability in quality. This was done to ensure coders were equipped to score conceptualizations across the whole range of the scale. Following their preparation, five trained raters (four undergraduate research assistants and the author of this thesis) independently evaluated all 20 case conceptualizations using the BACQS.

Data Cleaning

All coders returned complete scores for all 20 conceptualizations. Scores were examined for impossible values, and none were identified.

Data Analyses

A series of intraclass correlation coefficients (ICCs) were calculated to determine inter-rater reliability for each item of the BACQS. The scores from each item for each rater were also summed to obtain five scale total scores and these were entered into an ICC as well. Intraclass correlation coefficients are a commonly utilized statistic when researchers are evaluating inter-rater reliability when there are more than two raters (Shrout & Fleiss, 1979). Several forms of ICC can be utilized depending on whether raters remain the same across the coded stimuli and depending on whether the scores obtained must be exactly the same across raters (absolute agreement ICC: a more conservative method), or merely correlated with other rater's scores (consistency ICC; more liberal). The consistency ICC is more appropriate when having some raters be biased to higher or lower scores on average is not an issue and agreement on relative levels of the coded values is sufficient (Koo & Li, 2016). In this pilot study, the more stringent ICC method of absolute agreement was utilized. One reason for this choice is that the BACQS includes descriptions for the qualities and features at each score (1-5) which should allow coders to agree on the exact scores. Additionally, it was reasoned that if the BACQS succeeded under the more stringent criteria of the absolute agreement ICCs it gave more support that the scale was ready for use in further studies.

Another important point about ICCs is their similarity to measures of internal consistency, such as Cronbach's Alpha, in that adding raters will improve the reliability obtained in a way that is similar to how additional items tend to improve the internal consistency of a measure. For this reason, when ICC results are reported it is useful to include statistics for the reliability across all coders (which will generally improve with additional raters) and the "reliability" that can be expected for any one coder, on average. Respectively these are known as average measures and single measures ICCs.

In terms of understanding the results of ICCs, higher values represent a higher rate of agreement and a greater proportion of "signal to noise" captured in the coding across raters. General interpretive guidelines have been established such that scores below .5 are considered "poor", scores between .5 - .75 are "moderate", those between .75 - .90 are "good", and above .90 are "excellent" (Koo & Li, 2016; Shrout & Lane, 2012).

After examining the inter-rater reliability of the BACQS the scale's structure and internal consistency were also analyzed. For this initial analysis a unifactorial structure was assumed, as other case conceptualization quality measures have tended to use this model (see Bucci et al., 2016), and the scale was assessed for internal consistency using Cronbach's alpha. To calculate this an average of each item's scores from across the raters was computed and these were used. Similar to ICCs, guidelines have been proposed for minimal acceptable values of Cronbach’s alpha. These vary by test purpose, with higher minimums for tests with higher stakes applications. For basic research purposes a value of .80 is often seen as adequate while a value of .90 or higher may be required for clinical applications (see Streiner, 2003 for a discussion). Corrected item-total correlations were also analyzed to examine how strongly each item was correlating to the sum of remaining items. As a final step in examining the scale's internal structure inter-item correlations were also examined.

The final analysis conducted was on the scale's recode reliability. One coder's (the authors) initial total scores for the 20 conceptualizations were compared to total scores obtained after recoding the sample six months later. In interpreting the recode reliability results, standard interpretive ranges for the strengths of correlation coefficients were consulted (Schober et al., 2018) and a value of .70 or higher was selected as a cut-off for considering a correlation strong.

Results

Inter-Rater Reliability

The ICC values for each item and the 95% confidence intervals around those values are reported in Table 1.

Table 1

Absolute Agreement Inter-rater reliability (Pilot Study)

Item	Single Measures ICC (95% CI)	Five Coder Average Measures ICC (95% CI)
1) Overall Quality	.69 (.48-.85)	.92 (.82-.97)
2) Psychological Mindedness	.70 (.53-.85)	.92 (.85-.97)
3) Integration	.66 (.49-.82)	.91 (.82-.96)
4) Differentiation/Breadth	.58 (.38-.77)	.87 (.75-.94)
5) Core Mechanism	.58 (.38-.77)	.87 (.76-.94)
6) Strengths/Resiliency	.64 (.45-.81)	.90 (.80-.95)
BACQS total score	.80 (.67-.90)	.95 (.91-.98)

Recode Reliability

One coder returned to the sample and recoded the entire set of conceptualizations after a period of six months. The long gap between original scoring of conceptualizations and the subsequent recoding (in this pilot and the later studies) was intended to allow for the original associations between conceptualizations and their BACQS scores to be forgotten by the coder to the greatest possible degree. This would allow for the most robust test of whether the BACQS can guide coders to similar scores without the benefit/impact of remembering previous applications. Forgetting of these types of associations typically follows an exponential curve where the largest amount of forgetting occurs rapidly, followed by increasingly slower forgetting of residual information over time. Significant forgetting, ~90% of material learned, can occur over the course of even one month (Murre & Dros, 2015), and as such a gap of between 4-6 months was expected to be sufficient for these purposes. The correlation between the original BACQS total scores and their recoded values obtained was *r* (*1*8) = .81, *p* < .001.

Internal Consistency and Item Inter-Correlations

Cronbach's alpha across the (then) six items of the BACQS was .96, indicating a very high internal consistency. Excluding item 1, the overall quality item (which one could assume would be the most strongly related to the others), still resulted in a high Cronbach's alpha of .92. Corrected item-total correlations provide a way of determining how strongly an item of a scale correlates with the total (or average) score of the other items, excluding itself. The lowest corrected item-total correlation was found for the strengths and resiliency focus item, and was *r* =.69, while the highest was for the overall score item, at *r* = .97. The remaining items all had corrected item-total correlations higher than *r* = .83. The inter-item correlations for the BACQS are presented in Table 2 below.

Table 2

Inter-item Correlations (Pilot Study)

	1	2	3	4	5
Item 1: Overall Quality...	-				
Item 2: Psych. minded...	.90***	-			
Item 3: Integration	.89***	.93***	-		
Item 4: Differentiation...	.90***	.79***	.75***	-	
Item 5: Core mechanism...	.91***	.88***	.87***	.77***	-
Item 6: Strengths/resil...	.73***	.65**	.65**	.61**	.59**

note: *** p < .001, ** p < .01, * p < .05

Discussion

Summary of Pilot Results

The results of the pilot study generally supported our hypotheses and the readiness of the BACQS for use in subsequent studies. The inter-rater reliability analyses (ICC values) suggested that at the level of individual items raters achieved at least moderate levels of agreement/reliability when considering one "average" rater, though in looking at the confidence intervals some items strayed into the poor range. However, the full-scale score performed well enough to achieve a good level of reliability even considering a single rater's scores, with the confidence interval remaining in the moderate range at the lower end. The results were even more positive when considering the reliability of the whole team of five raters. There, the item reliabilities fell in the good or excellent ranges, even when considering lower limits of the confidence intervals. The scale total score's reliability appeared particularly robust, remaining in the excellent range across the entire range of the confidence interval.

In considering the internal consistency results, where a very high Cronbach's alpha was obtained and generally high inter-item and item-total correlations were demonstrated, generating a total score for the BACQS appeared appropriate. The pattern of inter-item and item-total correlations suggested a high degree of overlap between items, and (without having a large enough sample to conduct a factor analysis) also suggested that the scale was assessing an overarching construct (ideally the quality of case conceptualizations).

From one point of view, the high item-total and inter-item correlations which we obtained are strengths of the scale, suggesting that each item is contributing to the measurement of a latent construct; however, from other perspectives these very high correlations could be seen as a potential issue. If scale items are too highly correlated they may be providing little non-redundant information. Very high correlations would also contrast with the theory underlying the selected items, that the scale is comprised of relatively distinguishable and independent components of conceptualization quality.

The strong recode reliability obtained in this study is also worth noting. This result appears to indicate the scale can guide coders to consistent scores over a period of six months. This consistency is another demonstration that scores obtained from the BACQS capture non-random information. This initial result was particularly promising given that few existing case conceptualization quality measures have assessed consistency in scores, though a more complete examination of the coding scheme’s test-retest reliability would still be important in the future (Bucci et al., 2016).

Finally, we can consider that the promising results obtained across the various reliability estimates were obtained while the majority of the coding team was comprised of undergraduate coders. It appears reasonable to suspect that higher reliability could be obtained with expert raters, and therefore that adequate single-rater reliability could be obtained under those more ideal conditions.

Limitations

There are several significant limitations in this pilot study. The sample of conceptualizations was written by someone with interest in demonstrating that the BACQS performs well, which tends to not be best scientific practice. Although no deliberate efforts were made to create conceptualizations that were "easy" to code, the intentionally high variance in quality (deliberately including some very low quality and some very high quality) may have served to inflate the reliabilities obtained as conceptualizations at the extremes may have been easier to identify and agree upon. The length of these conceptualizations also tended to be somewhat shorter on average than the conceptualizations obtained in the later studies, which could also have made this sample easier to code.

Additionally, having only one writer for the conceptualizations likely resulted in a more homogenous sample in terms of writing style, vocabulary, and clinical and theoretical perspectives. This could have contributed to making the coding process somewhat easier (less taxing than having to deal with varied styles, perspectives). In addition, the small sample of 20 conceptualizations limits the confidence that can be placed in the results.

Although this pilot provided some evidence for the inter-rater reliability, recode reliability, and internal consistency of the BACQS, methods for assessing the scale's construct validity (such as convergent or divergent validity) were not included. As such, despite indications that coders were generally agreeing on scores, and that a rater’s scores could remain consistent over time, whether these scores were reflecting conceptualization quality could not be determined.

Conclusions and Next Steps

While preliminary, the results of the analyses were generally supportive of the hypotheses proposed. The BACQS scores were demonstrated to be sufficiently reliable across coders, though this was primarily true when considering the five raters' scores together or a rater's total score. The BACQS also demonstrated promising recode reliability, high internal consistency, and generally high inter-item correlations. As discussed above, the very high internal consistency and inter-item correlations was a potential problem, perhaps indicating scoring was subject to halo biases or that items of the BACQS were somewhat redundant. However, given the many limits of this study it was unclear whether this issue would persist in future studies.

In the following studies efforts were made to address some of the limits of this pilot study. The same tests of the scale's reliability were to be conducted on a larger sample of conceptualizations which were generated by more than one person and in response to something more like an actual clinical scenario (not just created wholesale for the purpose of being coded). Also sought were ways of providing evidence of the scale's construct validity. In moving forward, the decision was also made to add back the theoretical and logical grounding item. This was due to the success the coders had with the current items which indicated that this potentially more challenging item could also be coded reliably.

Study 1: Applying the BACQS to Written Case Conceptualizations of a Mock Psychotherapy Case

In this study the performance of the scale in terms of inter-rater reliability, recode reliability, and internal consistency was reassessed on a larger sample of conceptualizations which were generated more naturalistically. Support for the scale's construct validity was also sought out by examining how BACQS scores correlated with other variables. These included some variables which were expected to correlate at least moderately with BACQS scores, serving as tests of convergent validity, and variables which were not expected to significantly correlate with the scale's scores (but which could plausibly have been confounded with conceptualization quality), serving as tests of divergent validity. To this end, an in-lab analogue experiment involving undergraduate participants was conducted, from which the data for both Study 1 and Study 2 were obtained.

Analogue studies have long been a useful way of obtaining a larger sample of data while avoiding the difficulties of recruiting less numerous or less available participants (McNeil & Hayes, 2014). The higher number of anticipated participants (recruiting undergraduates versus clinical psychologists) was a main driver for choosing an analogue research design. This choice comes at the cost of lower external validity, in that the results of the analogue study may not necessarily generalize to the real-world context it is simulating. For example, it was expected that our analogue participants might tend to produce conceptualizations of lower quality than those which would likely be obtained from practicing clinical psychologists or graduate students in clinical psychology. However, there are examples where undergraduate participants (when provided with some training) have been effective in providing psychotherapy-like interventions (Pascual-Leone et al., 2014) which suggested that perhaps they may have been able to generate reasonably realistic conceptualizations in this study and serve as useful analogues. Including undergraduate participants also allowed for some useful comparisons across subgroups (across different majors of study) in the study sample, which served as additional tests of the BACQS's validity (this is returned to below).

In the context of psychotherapy research, recruiting undergraduate participants also has the advantage that this is the pool from which graduate students in clinical psychology are selected. As such, the ability to reliably evaluate case conceptualizations generated by undergraduates (who are not so different in age or training from junior graduate students in clinical psychology) may give some insight into how the BACQS could be applied in graduate training and supervision contexts.

Past researchers interested in case conceptualizations have used case vignettes in the form of videos with actors portraying clients, audio recordings of sessions, or short written descriptions of cases presented to their participants who then produce a case conceptualization (Kuyken et al., 2005). Given these precedents, a similar approach was used in this study. Participants were shown a video recording vignette where an actor portrayed a fictional client during their first psychotherapy session.

A benefit of using vignettes is that they are standardized stimuli. This controls for some of the variation in the conceptualizations being produced. This can make it easier to identify useful statistical relationships that might otherwise be obscured by less consistent stimuli and the additional noise this can generate. The use of vignettes or recordings can also allow researchers more control in the nature of the information available to participants, for example in this study the vignette was designed to contain enough detail and information that a high-quality conceptualization could theoretically be created. The case and the content of the session was also created with the goal of it being relatively typical, which may allow for some generalization of these results to other typical cases and their conceptualizations.

Additionally, the combination of a case vignette and the in-lab nature of the of the study made it easier to obtain a complete and current conceptualization for later coding. Through the instructions given to participants they could be directed to provide their whole understanding of the case presented to them without this being filtered through a psychotherapy report (where more tentative/inferential or less flattering aspects of a conceptualization may not be included) or via retrospective reporting from therapists (where information may be forgotten, changed, or omitted).

The control over the case presented in this study also allowed for a method of testing aspects of the BACQS' performance, specifically in terms of finding support for the scale's validity. During the creation of the fictional case, which is described in the methods section more thoroughly, notes were taken on the key points about the client and their situation which were to be presented through the video vignette. This allowed for the research team to create a "master conceptualization" which could be compared to conceptualizations of the study participants. This allowed for a test of the criterion validity of the BACQS scores, tempered by the recognition that our master conceptualization is not necessarily the only valuable way of conceptualizing the case.

As mentioned above, including undergraduates allowed for some useful between-subgroup comparisons across the sample. These comparisons were motivated by several studies which have identified differences between psychology and non-psychology students on: mental health literacy (Miles et al., 2020); interest in pursuing mental health careers and level of empathy (Harton & Lyons, 2003); and psychological mindedness (Trudeau & Reich, 1995).

Given these findings, it appeared reasonable to suspect that if BACQS scores from psychology students were found to be higher this might indicate that the BACQS evaluates something related to psychology-specific training and knowledge, as it logically should. Related to this, it was suspected that individuals with interest in and familiarity with psychotherapy and mental health difficulties would also generate higher quality conceptualizations.

The proposed link between case conceptualizations and empathy was also considered. It has been suggested that higher quality case conceptualizations should contribute to more empathy from a clinician (Johnstone et al., 2011). As such, attempts were made to evaluate the empathy a participant felt towards the simulated client to test if this related to the quality of their conceptualization. Additionally, a person's trait level of empathy could plausibly impact the quality of their case conceptualizations. Many definitions of empathy emphasize caring for another person's experiences and being able to imagine these more accurately through perspective taking (Elliot et al., 2011), which both appear relevant to case conceptualizing. Additionally, the broader Big Five trait of Agreeableness shares conceptual overlap with empathy and has been found to correlate with it (Bamford & Davidson, 2019), those higher in agreeableness tend to be kinder, more sympathetic, more pro-social, more considerate of others, and interpersonally warm (Habashi et al., 2016). It appeared plausible that this trait could also relate to interest in and ability to conceptualize the difficulties of others.

Case conceptualization has sometimes been referred to as a cognitive task (Eells, 2011; Ridley et al., 2017), which may be reflected in how much the information processing, summarizing, causal link-forming, and hypothesis testing aspects of this process are emphasized in the literature. As a clinical tool that requires information processing it was suspected that the quality of case conceptualization might relate to several personality traits known to impact how individuals engage in cognitive tasks. For example, that an individual's need for cognition might play a role. Need for cognition is a personality trait which has been described as a person's interest for and enjoyment of tasks which require thinking (Cacioppo & Petty, 1982). Those higher in this trait engage in more effortful cognitive activities and show greater effort while engaging in these tasks (Cacioppo et al., 1996; Verplanken et al., 1992). They also appear to utilize more persuasive and less biased judgments, and hold less dogmatic beliefs (Cacioppo & Petty, 1982; Shestowsky et al., 1998). Each of those points appears to be relevant to a person's ability to produce high quality case conceptualizations.

Need for cognition is related to, but is considered distinct from, the big five personality trait of openness, and to measures of intelligence (Fleishhauer et al., 2010). Given this, openness was also considered as possibly playing a role in conceptualization quality. Trait openness is related a person's curiosity, imagination, openness to emotions, and tendency to think abstractly (Conelly et al., 2014), all of which appeared likely to contribute to conceptualization quality. Some research has found psychologists, and psychology majors, have higher levels of openness and curiosity (Chapman et al., 2009; Vedel, 2016). Some qualitative research into the traits of "master therapists" found that they tend to report high openness to confusing or disconfirming information from their clients (Jennings & Skovholt, 2016), which could translate into more accurate and well thought out conceptualizations.

Mental health stigma also appeared promising as a variable through which the scale's validity could be supported. Whereas openness was expected to predict higher quality conceptualizations it appeared likely that a participant who held stigmatizing beliefs towards mental health difficulties would produce lower quality conceptualizations. Individuals who hold such stigmatizing beliefs often attribute a person's mental health difficulties to moral failings or character flaws in stereotyped ways, and/or can believe those with mental health difficulties are dangerous or incompetent (Fox et al., 2018), they also have been found to show less empathy and greater social distancing from those with mental health difficulties (Feldman & Crandall, 2007; Hing et al., 2016). As such, more stigmatizing and stereotyped attitudes towards mental health difficulties should result in lower quality conceptualizations, as they would likely resort to stereotyped and simplistic explanations, be less invested in understanding the experiences of someone with mental health difficulties, and be less likely to have empathy for their struggles.

Students admitted into clinical psychology programs are often selected at least partly on the basis of their academic achievement, which shares some overlap with measures of intelligence (Roth et al., 2015), which as mentioned above also tends to correlate with openness and need for cognition. It was considered likely that students with higher academic achievement would likely have more knowledge to draw upon in conceptualizing, may be more effective at attending to and selecting relevant information, may tend to form more useful links and inferences based on the information they consider, and may be better able to clearly express their ideas. For these reasons academic performance was also considered as likely to correlate with BAQCS scores, and as another possible method of obtaining some support for the construct validity of the scale.

In considering the traits and variables above, it appeared plausible that several may also be linked to a participant's writing skills, for example: intelligence, academic achievement, and need for cognition. Writing skill could also be impacted by English fluency, which at a multicultural university with many international students, such as the University of Waterloo, may vary across participants. While some correlation between a high-quality conceptualization and the quality of writing might be expected, it appeared important to test that this overlap was not too high, to help demonstrate that coders were not using a heuristic where writing quality was influencing their evaluations the conceptualizations obtained. Similarly, the possibility that coders could conflate length of conceptualizations with quality was also considered. Part of the BACQS explicitly involves evaluating breadth and differentiation, and even for the other items more detail would often be necessary for higher scores. That coders might begin to evaluate conceptualizations too strongly on the basis of length appeared a plausible concern.

Study Aims and Hypotheses

The main aims of this study were to:

1) Evaluate the inter-rater reliability of the BACQS on a sample of written conceptualizations

2) Evaluate the recode reliability of the BACQS

3) Evaluate the internal consistency of the BACQS and the inter-item correlations to assess the scale's internal structure

4) Assess how the BACQS scores varied and correlated with other variables in order to establish signs of construct validity

Our main hypotheses were:

1) The BACQS would demonstrate adequate (.50 or above) inter-rater reliability for all seven items and the scale total score.

2) The BACQS would demonstrate high recode reliability.

3) The BACQS would demonstrate high internal consistency and the seven items would relate to each other in a coherent pattern suggestive of a scale assessing overall conceptualization quality.

4) The BACQS would demonstrate criterion validity when assessed via a correlation between BACQS scores and a count of how many key points from a "master conceptualization" participant written conceptualizations contained.

5) The BACQS would demonstrate initial evidence of convergent and divergent validity through associations between the total scale and other variables. Specifically, the BACQS scores would:

5a) be positively correlated with participant GPA

5b) be positively correlated with student interest in and familiarity with mental health and psychotherapy.

5c) be positively associated with need for cognition

5d) be positively correlated with participant empathy (at both a trait level and towards the fictional client specifically)

5e) be positively correlated with the personality traits of open-mindedness and agreeableness

5f) be negatively correlated with mental health stigma

5g) not be strongly associated with the length of conceptualizations

5h) not be strongly associated with the writing skill of a participant, estimated via grade level of the written text of the conceptualizations

5i) not vary by participant first language, also as a proxy for writing skill

5j) be higher in participants who are currently studying psychology

Methods

Participants

The study involved volunteer undergraduate participants recruited from the University of Waterloo's online psychology research participant pool. Participants self-selected into the study after reviewing a basic outline of the study details, which also indicated that they must be able to read, write, and follow a conversation fluently in English and have normal or corrected to normal vision. Participants in the pool were enrolled in at least one psychology course at the time of participation but not all were enrolled in psychology as their major of study. Participants received research participation course credits.

A total of 80 participants took part in the study. Eight participants either did not finish the study or were identified as having not understood important aspects of study instructions, and so were excluded. Of the remaining 72 participants 86.1% were female and the average age of the sample was 20 years (*SD* = 2.86, range of 18-33). A majority of the participants were not enrolled in psychology as their major (58.3%). The other fields participants were enrolled in included: arts & business, mathematics, chemistry, biology, software engineering, therapeutic recreation, social development studies, anthropology, biomedical sciences, legal studies, and public health. Participants were diverse in cultural/ethnic backgrounds. For simplicity we have collapsed these into the following categories: Caucasian (45%), East Asian (23.8%), South Asian (17.5%), African (6.3), Middle Eastern (5%), and Hispanic (2.4%). English was identified as the first language for 73.6% of participants.

Procedure

The procedure involved participants individually attending a 1.5 hour long in-lab study involving four major components. The first component of the study had participants provide demographic information and complete several self-report questionnaires. This first component of the study took approximately 20 minutes to complete.

The second component involved participants watching and listening to a recording of a simulated first session of psychotherapy (akin to an intake session) with a client named “Larry". The recording lasted 30 minutes, displayed only the simulated client on screen, but contained audio from both Larry and his "therapist". Participants were given lined paper and writing utensils so they could take notes during the video, should they desire. The following instruction was given to participants before the video was played:

> *You are now going to be shown a video of a fake psychotherapy session. The person playing the role of the client is an actor, but the therapist has actual training in conducting psychotherapy. The client and video were designed to realistically simulate a first session. Your task now is to watch the video as if you were a psychologist or psychotherapist trying to understand the client and their problems, identify what needs to change or what is causing the problems, and plan a treatment. You may take notes on paper throughout the video to use later. You may not pause or rewind. The video is approximately 30 minutes long. If you need to, please use the washroom before beginning the video.*

Larry's case was designed to portray a first year University of Waterloo undergraduate student experiencing academic difficulties stemming from social anxiety, low mood, and problematic alcohol use. Several significant events in Larry's life were discussed during the interview, as well as several more recent situations related to his difficulties. Efforts were made to ensure that the interview also covered information regarding Larry's case that could be considered relevant for a variety of therapeutic orientations (i.e. - indications about his personality, specific symptoms, negative automatic thoughts, interpersonal style, maladaptive behavioural patterns, level of emotional awareness and avoidance, level of self-awareness, unconscious resentments, unconscious interpersonal needs, core beliefs, etc.). Some strengths of Larry's were also represented or discussed during the video, such as him having a few acquaintances, past periods of academic success, and a cooperative approach in his session. Larry was portrayed by a male Caucasian psychology undergraduate research assistant with experience in acting who also helped generate the case. The therapist was portrayed by the author of this thesis.

The third section of the study involved the participants producing a conceptualization of Larry's difficulties and representing these in a conceptualization diagram/concept map, and a typed version (which is the focus of this study while the concept map is the main focus of Study 2 and is explained in more detail there). For both the concept map and the written explanation the instructions directed participants to express their understanding of Larry's difficulties and their causes in a way that it could guide a course of psychotherapy. For the written conceptualization they were asked to write in full sentences. They were instructed that not all information from the video had to be included and that they could also include information not explicitly stated in the video. Participants could reference their notes while creating their conceptualizations and had 15 minutes to complete each version.

In the final portion of the study participants responded to data quality and effort check questions. They also answered questions on their empathy for Larry in particular, and questions assessing how strongly they hold stereotyped and stigmatizing beliefs around mental health difficulties.

Following the completion of data collection three trained coders applied the BACQS to the written conceptualizations of Larry's difficulties obtained from the 72 participants who were retained from the original sample of 80. The coders were blind to the participants' responses to the study questionnaires, and each independently coded through all of the available written conceptualizations. One coder additionally returned to the conceptualizations and made a series of dichotomous present/absent judgments regarding whether each of 30 key points of information from the "master conceptualization" were included in a participant's conceptualization.

Measures

Participants completed a collection of self-report measures which addressed three main objectives: (1) obtaining demographic information to help characterize and sub-group the sample, (2) checking data quality around participant attention and effort, (3) measuring relevant participant personality traits and experiences. The full versions of measures and items utilized in this study are presented in Appendix 3.

Demographic Questions: Participants were asked to indicate their age, number of terms in post-secondary studies, ethnicity/culture, sex, first language, and program of study.

Number of Key Features: This variable was generated by the author of the thesis returning to the written conceptualizations months after finishing all the coding using the BACQS and evaluating how many of the 30 key features from the "master conceptualization" which guided the video vignette were present in each participant conceptualization. The full list of key features can be found in Appendix 4.

Grade Point Average (GPA): Participants in the University of Waterloo's research participation pool have the option of completing a separate online questionnaire (for additional research credits) which can be accessed by any approved researcher to utilize alongside the data collected from their own studies. From this optional questionnaire we were able to obtain the current grade point average for some participants.

Mental Health Skill and Interest Items*:* In the first section of the study three items were presented to gauge a participant's self-reported interest in and familiarity with mental health issues and psychotherapy.

The Big Five Inventory - 2- Short (Soto & John, 2017): This brief (30-item) measure assesses the big five personality traits and contains short scales for each (openness, conscientiousness, extraversion, agreeableness, neuroticism). The primary scales of interest were agreeableness and openness. The scale was selected based on past evidence of good psychometric performance. For example, having scale internal consistencies above α = .70 (range: .73 -.82) when applied to a sample of university students, high correlations between the scales from the 30-item measure and a 60-item version (which itself has demonstrated reliability and validity, all correlations above *r* = .90), and moderate concordance between self and peer ratings using the measure.

Interpersonal Reactivity Index (Davis, 1980): This widely used and longstanding 28-item self-report measure assesses trait level empathy, broadly defined by the scale's creators as a reaction one person has when observing another person's experiences. The measure views empathy as a multifactor construct and includes four subscales - perspective taking, fantasy, empathic concern, and personal distress. The perspective taking subscale represents the ability and propensity to adopt the points of view of others. The fantasy subscale measures the tendency of a person to feel and experience the feelings and actions of fictitious characters. The empathic concern subscale assesses to what degree a respondent experiences feelings of sympathy and concern for unfortunate others. The personal distress subscale represents a predisposition to have personal anxiety and unease in difficult interpersonal settings.

Need for Cognition Inventory (Cacioppo et al., 1984): This 18-item self-report questionnaire assesses trait level interest in engaging with cognitive tasks. Individuals who score higher on this scale demonstrate more interest in tasks which require thinking deeply, thinking abstractly, solving complex problems, and understanding why something works the way it does. In our study the scale had excellent internal consistency (Cronbach's alpha = .86).

Attention, Effort, and Confound Check Items: Several items were presented to participants in the final portion of the study to help ensure data validity and that participant effort and attention were adequate. Participants were asked about their level of effort, distraction, how accurate their written conceptualizations were of their internal understanding, whether they knew the actor portraying Larry, and how real the simulated psychotherapy session appeared to them.

Empathy for Larry: Participants responded to nine items at the end of the study regarding their empathy for Larry in particular, imagining how they might feel if he was a real person. These items were generated after reviewing several empathy measures, including the IRI, Toronto Empathy Questionnaire (Spreng et al., 2009), and the Questionnaire of Cognitive and Affective Empathy (Reniers et al., 2011). The nine items emphasized the degree to which participants held a warm, caring, non-judgmental, and curious stance towards Larry.

Select Items from the Attitudes to Severe Mental Illness Scale (taken from the ASMI; Madianos et al., 2012). In the final portion of the study participants were also asked six questions from a measure of severe mental health stigma. These items were selected as they loaded highly onto one of the original scale's four factors, representing stereotyped thinking about individuals with mental health difficulties. The six items, rather than the whole scale, were selected due to time constraints in the study.

***Writing Quality*:** As a proxy for writing quality the grade level of the text was calculated. The grade level of writing is often used to evaluate how simple versus challenging samples of text are to read and the typical level of education someone would need to find the text readable. Of relevance here, there is some evidence more difficult text is perceived as more trustworthy or believable to readers, regardless of the accuracy of information being conveyed, possibly as readers assume more difficult text relates to subject expertise (Withall & Sagi, 2021). The most common method of evaluating grade level of writing is by the Flesh-Kincaid Grade level formula, which is calculated by examining the ratio of syllables to words and words to sentences, with more syllables per word and more words per sentence resulting in a higher grade level (Kincaid et al.,1975).

***Conceptualization Length*:** Conceptualization length was calculated by conducting a word count for each conceptualization.

Missing Data, Data Quality, and Data Cleaning

Aside from the eight participants excluded due to failure to complete the study or failure to understand important study instructions, the remaining dataset was quite complete (with one exception noted below). Two participants missed responding to the interest and familiarity with mental health items, one missed items on the personal distress subscale of the IRI, and one missed items on the empathy for Larry and stigma scales. All coders returned full data for all conceptualizations. To preserve as much statistical power as possible, and given a very small amount of apparently random missing data, pairwise deletion was utilized across later analyses.

The exception to the relatively complete dataset was for the GPAs of participants. As noted above, this data was obtained from an optional questionnaire which participants could complete for extra research participation credits through the online system undergraduates use to track and register for psychology research. This questionnaire is not a part of any particular research study and is available to all researchers using the system to recruit participants into their studies. A total of 41 participants had completed this questionnaire and answered the question on their current GPA.

Data were examined for impossible values and none were identified. Data quality was assessed via responses to the items assessing level of effort during the study, personal familiarity with the actor portraying Larry, level of distraction while watching the video of Larry's session, and how accurate written conceptualizations were to a participant's internal understanding of Larry's difficulties. Participants were removed if their scores for any question were above or below "neutral" (depending on which direction indicated a potential data quality problem). Five participants were identified as having potentially problematic data due to their responses to the data quality items. The scores obtained from coding their conceptualizations were still included in reliability and internal consistency analyses but were excluded from analyses where BACQS scores were analyzed for relationships to other variables and participant traits.

The believability of the video as a simulation of an actual therapy session was ultimately not utilized for screening as following the same rule as above would have resulted in removing an additional 14 participants. The wording of this item appears overly stringent in retrospect ("If I hadn't been told ahead of time I might have thought this was a real video of a therapy session), and a better choice may have been to "the video appeared to be a plausible simulation of psychotherapy" or something similar. The scores on this item did not correlate with participant's self-rated familiarity with psychotherapy which also lends support to not utilizing them to screen participants.

Suspicious/extreme multivariate/bivariate outliers were excluded from analyses. To identify these outliers, DF betas were calculated for all bivariate/multivariate analyses and participants were removed from specific analyses if their DF beta scores fell more than three inter-quartile ranges away from the mean DF beta score. For interested readers, we report the original pattern of results found with the total sample in Appendix 5, Table 20, and in our results section we note where excluding the extreme influencer cases had a large effect.

Data Analyses

A series of intraclass correlation coefficients (ICCs) were calculated to determine inter-rater reliability for each item of the BACQS as well as the inter-rater reliability of the scale total score (summing all of a rater's item scores together). The ICCs were interpreted using the same guidelines of: .50 - poor reliability, .50 to .75 - moderate reliability, .75 to .90 - good reliability, and values above .90 - excellent reliability.

A two-way random effects model was appropriate for all ICC analyses in this study and as in the pilot study we report absolute agreement model results (where scores are expected to be agreed on between raters) but additionally added the consistency model results (where scores between raters are merely expected to correlate) to characterize how the scale performed in this less stringent approach. Next, the internal consistency of the BACQS was calculated using Cronbach's alpha.

The scale's items were then correlated with each other to determine the strength of the relationships between the various pairs of items. This allowed for a crude examination of the pattern of relationships in the items of the scale. Correlations were then conducted between the BACQS total scores and the other study variables of interest.

A regression analysis was also conducted predicting BACQS scores. Given this study's modest sample size the number of predictor variables in the regression needed to be reduced. A principal components analysis (PCA) was performed to identify meaningful groupings of variables so these could be combined, reducing the total number in the regression. The big five personality variables, the mental health skill and interest scale, the empathy for Larry scale, the subscales of the IRI, the need for cognition scale, the grade level of the writing, and the number of words in a conceptualization were entered into the PCA with a varimax rotation. In both the PCA and the regression, the mental health stigma scale and participant GPA were not utilized. The stigma scale was not included due to the poor internal consistency of the scale and participant GPA was not included due to the missing data which would further restrict the sample size and statistical power.

Five components were obtained through the PCA. Variables which moderately or strongly loaded onto each component (a magnitude above .40) were then standardized, multiplied by their loading magnitude, and summed together in order to create component scores/marker variables. The first component included three of the IRI subscales (not the personal distress subscale), the empathy for Larry scale, and the agreeableness scale of the big five measure. This component appeared to represent a non-distressed and warm empathy. The second component included the personal distress subscale of the IRI, the neuroticism subscale of the big five inventory, and also a negative loading for conscientiousness. This component could have represented dysregulation in emotion and organization or a more intense and affective aspect of empathy. The third component included need for cognition, the openness subscale of the big five inventory, and a negative loading of the personal distress subscale. This component appears to represent a more emotionally detached cognitive openness and curiosity. It is interesting that a common theme in psychotherapy literature is the necessity for psychotherapists to modulate their own emotional reactions, including to a client's distress (Negd et al., 2011; Kim & Han, 2018), and this finding that personal distress did not associate closely with other aspects of empathy, and in fact negatively associated with openness and need for cognition may support this idea. The fourth component included the Flesch-Kincaid grade level of writing, the mental health skill and interest scale, and a negative loading for extraversion. This component was somewhat harder to grasp, but could represent an introspective nature, consisting of greater skill with writing, more familiarity with mental health concepts, and a tendency towards introversion. The final component had one variable loading onto it, the number of words in the conceptualization.

The final set of analyses conducted to assess construct validity were between-group t-tests. First, whether the BACQS total scores differed across psychology and non-psychology students was tested, then whether BACQS scores differed across English as first language or not English as first language participants.

Results

Internal Consistency of Measures

Given that the familiarity and skill with mental health items had a shared focus their inter-item correlations were explored and were found to range from .40 -.47, *p* < .001. These items were also found to have good internal consistency when assessed together (Cronbach's alpha = .71). As such, these items were summed together in later analyses.

Each trait scale of the Big Five Inventory reached adequate or higher levels of internal consistency. The lowest was for the agreeableness scale, Cronbach's Alpha = .68, and the highest was for neuroticism, Cronbach's Alpha = .83. Though we were most interested in agreeableness and openness, we also examined the correlations between the BACQS and the remaining three Big Five traits (extraversion, conscientiousness, neuroticism) as they were available.

The Interpersonal Reactivity Index also had good internal consistency for the total scale, Cronbach's alpha = .78, and for each subscale (between .75 - .78). However, this appeared mostly due to a high correlation between three of the subscales, with which the personal distress subscale did not correlate strongly. The personal distress subscale also did not correlate with agreeableness as assessed in the Big Five Inventory - S - 2 (whereas the others did). This suggested that this subscale may represent a distinct construct, less tied to the other forms of empathy. Given this, the IRI full scale was not utilized.

The empathy for Larry items were amalgamated into a single score for analyses as they all significantly inter-correlated and had high internal consistency as a short scale (Cronbach's Alpha .89). These items correlated significantly with all the subscales of the IRI except for the personal distress subscale (the remaining correlations ranged between *r* = .36 and .53, p < .001).

In our sample the set of mental health stigma items had a low internal consistency, below typical cut-off for use, at Cronbach's alpha = .60. We nonetheless summed these items together into a short scale and did utilize this measure in some analyses. Given the low internal consistency of this scale the results of those analyses should be interpreted with caution.

Descriptive Statistics

In the table below the descriptive statistics for the study variables (aside from the scores obtained from applying the BACQS) are presented.

Table 3

Descriptive Statistics for Study 1 Variables

	n	Mean *(SD)*	Range	Skewness	Kurtosis
Number of Key Features	68	10.57 (3.61)	2-19	.12	-.51
Length of conceptualizations	68	237.59 (84.85)	87-557	1.15	2.22
Flesch-Kincaid Grade Level	68	10.73 (2.53)	6.60-20.80	1.62	4.87
Participant GPA	38	78.16 (5.80)	70-90	.27	-.61
Need For Cognition	68	62.81 (10.41)	29-82	-.57	.49
Interpersonal Reactivity Index	67	100.03 (11.28)	69-123	-.35	-.14
Fantasy	68	26.15 (5.19)	16-35	-.28	-.80
Perspective Taking	68	26.06 (4.53)	12-35	-.61	.89
Empathic Concern	68	28.13 (4.04)	18-35	-.50	-.09
Personal Distress	67	19.73 (4.97)	8-33	.30	.06
Short Big Five Inventory					
Openness	68	23.14 (4.11)	14-30	-.23	-.84
Conscientiousness	68	19.59 (4.46)	7-28	-.29	.03
Extraversion	68	19.62 (5.13)	7-29	-.45	-.75
Agreeableness	68	23.07 (3.88)	11-30	-.55	.44
Neuroticism	68	17.78 (5.84)	8-30	.28	-.83

Stigma Items Total	67	11.72 (4.06)	6-23	1.03	.61
Mental Health Skill and Interest	66	14.37 (3.89)	4-20	-.46	-.42
Empathy for Larry	67	50.27 (9.11)	20-63	-1.38	2.14

The table below presents the descriptive statistics for the scores obtained from the three coders applying the BACQS to the sample of conceptualizations.

Table 4

Descriptive Statistics for Written Conceptualization BACQS Scores

	n	Mean *(SD)*	Range	Skewness	Kurtosis
BACQS total coder 1	72	17.93 (4.68)	9-30	.36	-.31
BACQS total coder 2	72	19.99 (3.32)	12-27	.13	-.64
BACQS total coder 3	72	18.04 (5.33)	8-34	.47	-.12
All Coders Average					
Item 1: General Quality...	72	2.37 (.70)	1-4	.25	-.69
Item 2: Psych. Mindedness	72	2.57 (.72)	1-4.33	.19	-.64
Item 3: Integration	72	2.79 (.79)	1.33-4.67	.09	-.70
Item 4: Differentiation	72	2.72 (.60)	1.67-4.00	.10	-.47
Item 5: Core Mechanism	72	2.67 (.75)	1.33-4.33	.16	-.54
Item 6: Strengths and resil...	72	2.85 (.44)	1.67-4.00	-.68	1.57
Item 7: Theoretical and...	72	2.69 (.67)	1.33-4.33	.41	-.21
BACQS total score	72	18.65 (3.85)	10.67-28.33	.19	-.54

Inter-Rater Reliability

Below, the intraclass correlation coefficients on the BACQS items and scale total score are reported. The total score was generated by summing a coder's scores for each item.

Table 5

BACQS Inter-rater Reliability Estimates (Three Raters, Written Conceptualizations)

	Absolute Agreement ICC (95% CI)	Consistency ICC (95% CI)
Item 1: General Quality Impression		
Average measures	.63 (.36-.78)	.72 (.59-.82)
Single measures	.37 (.16-.55)	.46 (.32-.60)
Item 2: Psychological Mindedness		
Average measures	.77 (.65-.85)	.77 (.66-.85)
Single measures	.52 (.39-.65)	.53 (.39-.65)
Item 3: Integration		
Average measures	.73 (.60-.82)	.74 (.62-.83)
Single measures	.47 (.33-.60)	.49 (.35-.62)
Item 4: Differentiation/Breadth		
Average measures	.66 (.50-.78)	.68 (.53-.79)
Single measures	.39 (.25-.54)	.42 (.27-.56)
Item 5: Core Mechanism Identified		
Average measures	.67 (.48-.80)	.73 (.60-.82)
Single measures	.41(.23-.57)	.47 (.33-.60)
Item 6: Strengths and Resiliency Focus		
Average measures	.65 (.47-.78)	.69 (.54-.80)
Single measures	.39 (.23-.54)	.43 (.28-.57)
Item 7: Theoretical and Logical Grounding		
Average measures	.70 (.55-.80)	.69 (.55-.80)
Single measures	.43(.29-.57)	.43 (.29-.57)
Total Score		
Average measures	.79 (.68-.87)	.81 (.72-.88)
Single measures	.56 (.41-.68)	.59 (.46-.70)

Recode Reliability

Four months after coding the sample of written conceptualizations the study author recoded each conceptualization again. The correlation between original and retest total scores was *r* = .79, p < .001. As a more fine-grained look at the rates of agreement across the original and retest coding individual values from each item across the whole sample of conceptualizations were examined. Fifty-four percent of values did not change from the first to the second coding and 44% changed by only one point. Only 2% of scores changed by more than one point.

BACQS Internal consistency and Item Inter-Correlations

The internal consistency of the BACQS was within the excellent range (Cronbach's alpha = .92). Corrected item-total correlations fell above .85 for all items except the Differentiation/Breadth and Strengths and Resiliency Focus items which were *r* = .57 and *r* = .19 respectively. When compared to the others, the Strengths and Resiliency Focus item had a smaller standard deviation and higher kurtosis (See Table 3 for descriptive statistics), and 36 (50%) of conceptualizations received a score of 3 on this item (which for this item indicates a neutral tone with neither positive or negative content) from all raters. Item correlations are reported below in Table 6.

Table 6

BACQS Inter-Item Correlations (Written Conceptualizations)

	1	2	3	4	5	6
Item 1: Overall qual...	-					
Item 2: Psych minded.	.90***	-				
Item 3: Integration	.83***	.85***	-			
Item 4: Different...	.64***	.64***	.48***	-		
Item 5: Core mech...	.87***	.82***	.76***	.49***	-	
Item 6: Strengths...	.25	.17	.16	-.01	.20	-
Item 7: Theoretical...	.83***	.83***	.80***	.49***	.82***	.23*

note: *** p < .001, ** p < .01, * p < .05

Construct validity: Correlations

In the following table correlations between the BACQS total score and other scale totals and relevant variables are reported.

Table 7

BAQCS Correlations to Study Scales and Variables (Written Conceptualizations)

Variable	*n*	*r*
Criterion Validity Variable		
Number of Key Features	68	.65**
Convergent Validity Variables		
Participant GPA	37	.50**
Mental health skill and interest	66	.33**
Need for Cognition	68	.27*
Empathy for Larry	63	32*†
Interpersonal Reactivity Index		
Fantasy subscale	68	.34*

Perspective taking subscale	68	-.07
Empathic concern subscale	66	.06
Personal distress subscale	66	-.12
Openness	67	.31*†
Agreeableness	66	-.06
Mental health stigma items	64	-.20
Discriminant Validity Variables		
Conceptualization word count	62	.17‡
Flesch-Kincaid grade level	64	.18‡
Other Big Five Traits		
Conscientiousness	67	-.16
Extraversion	68	-.28*
Neuroticism	68	.07

note: *** p < .001, ** p < .01, * p < .05
† = this relationship was non-significant before excluding bivariate outliers
‡ = this relationship was significant before excluding bivariate outliers

Construct Validity: Regression

The regression model with the five variables obtained from the PCA predicting BACQS scores was significant, $R^2 = .22$, *SE* = .3.48, *F* (5, 49) = 3.84, *p* = .005. The number of words component, the "introspection" component, and the detached openness component were all significant predictors, while the warm empathy and affective empathy/distress components were not. Table 8 below contains a more detailed breakdown of the regression results.

Table 8

Regression Predicting BACQS Total Scores (Written Conceptualizations)

Variable	*β*	*t*
Warm Empathy	.03	0.26
Affective Empathy/Distress	.15	0.85
Detached Openness	.43**	2.70
Introspection	.31*	2.01
Number of Words	.31*	2.43

Note: * = significant at the *p*< .05 level, ** = significant at the *p* < .01 level.

Construct Validity: Between Sub-Group Analyses

An independent samples t-test was conducted to assess between group differences in BACQS scores relevant to the scales' construct validity. In particular, the interest here was if participants in psychology performed better on this task than students from other fields of study. The results indicated that students in psychology produced conceptualizations that scored significantly higher on the BACQS (*M* = 20.14, *SD* = 4.07) compared to students in non-psychology fields of study (*M* = 17.77, *SD* = 3.39), *t* (66) = 2.60, *p* = .01.

Another t-test was conducted to assess whether first language impacted BACQS scores as another proxy to evaluate whether writing skill (a potential confound) was related to BACQS scores. A significant difference was found between these groups, with the conceptualizations of English as first language participants scoring higher than the ESL group (*M* = 19.37, *SD* = 3.84 versus *M* = 17.00, *SD* = 3.29), *t* (66) = 2.37, *p* = .02.

Discussion

Through this study the performance of the BACQS was tested on a larger sample of more naturalistically obtained case conceptualizations, focusing on inter-rater reliability, recode reliability, and the scale's internal consistency. Evidence of the scale's construct validity was also sought by examining relationships between the scale total score and other variables, and by evaluating whether scores of conceptualization quality varied across subgroups in our sample.

Did the BACQS Demonstrate Adequate Inter-Rater Reliability for All Seven Items and The Scale Total Score?

The inter-rater reliability achieved after coding this sample of written conceptualizations was lower than what was obtained in the pilot study. At the level of individual items and the single average rater, most of the ICC values fell in the poor range, below .50, irrespective of whether absolute or consistency agreement models were utilized. The exceptions to this were the scale total and the item on psychological mindedness, which fell in the moderate range; however, considering the confidence intervals, both these could still fall in the poor range. The results of the average measures ICCs, assessing the performance of the coding team as a whole, were more positive. Item 2 (psychological mindedness) had an ICC value in the good range while the remaining items were in the moderate range. The confidence intervals around these values generally remained above .50 at the lower bound for the absolute agreement ICCs, and consistently did for the consistency ICCs. The total score ICC value fell solidly within the good range even considering the lower end of the confidence interval. These results can be taken as support for the adequate inter-rater reliability of the BACQS, at least when using three coder's scores.

The somewhat lower inter-rater reliability in this study versus in the pilot does not merely appear to be the result of having two fewer coders (compared to the pilot study), as there was a drop the single measures reliability as well. This suggests that perhaps this sample of conceptualizations was more challenging to code compared to the pilot study conceptualizations, or that coding suffered from a lack of practice on a subset of the sample. It may also be that coders would have benefitted from a chance to code several of the conceptualizations from this sample together and then discuss coding issues before working through the main sample independently. The approach of utilizing a practice sample has been utilized in other examples of case conceptualization research involving coding to apparently positive effect (Eells et al., 1998)

Did the BACQS Demonstrate High Recode Reliability?

Although it would have been more ideal for all three coders to have re-coded the conceptualizations this was unfeasible due to the time demands it would have placed on our volunteer coding team. As such, the estimate of recode reliability comes from a single coder, the writer. Being the only graduate level coder, this may reflect the performance of well-trained coders in future uses of the BACQS. The correlation between the original scores and those obtained four months later was quite high and supports good recode reliability even over a relatively long span of time. When looking at where scores did shift from the original to recoded values, the changes were almost all within one point of the original scores.

Did the BACQS Demonstrate High Internal Consistency and Inter-Item Correlations Suggestive of a Scale Assessing Overall Conceptualization Quality?

The BACQS demonstrated an excellent degree of internal consistency, similar to that which was obtained in the pilot study. Strong item-total and inter-item correlations were also found across most items, with the exception of Item 6 (the ‘strengths’ item). One possible explanation for this item's poorer correlation to the rest of the scale is that there appeared to be less variance in scores for this item. It may be that if more participants had incorporated strength and resiliency into their conceptualizations, or inversely included more negative content, a stronger relationship between this item and the others would have been evident. Interestingly, this finding that most participants did not include strengths into their conceptualizations echoes previous research, including the writer's Master's research, suggesting that strengths are not often incorporated into case conceptualizations (Capobianco, 2015). It may be that the study instructions contributed to this result, as the task assigned to participants was to understand the simulated client and his difficulties, which is a more problem focused framing. Perhaps more neutral wording could be utilized in the future to avoid leaning participants away from incorporating strengths.

Did the BACQS Demonstrate Criterion, Convergent, and Divergent Validity Through Associations with Other Variables?

The BACQS's criterion and construct validity was tested through examining the associations the scale's total score had with other variables. When tested via correlations this did not account for issues of shared variance and suppression, while the regression analysis required the number of variables entered to be reduced through a PCA and relied upon a smaller sample size due to missing data and the exclusion of extremely influential cases. Overall, there was support for the criterion validity of the scale and some support for the construct validity as well, though the magnitude of the relationships between the BACQS scores and convergent validity variables was typically small to moderate. Also, the relationship between empathy and the

BACQS scores was less straightforward than was hypothesized as three of the four subscales of the IRI had only small associations with case conceptualization quality.

It may be that features of the case and/or the fact that it was an artificial client help to explain this result. Larry was designed and portrayed as being emotionally reserved and somewhat alexithymic and did not display or articulate much emotion aside from nervousness in the session recording. It could be that a participant's general level of empathic concern for others may have been less activated by Larry's restricted range of emotions. It may also be that as participants were aware Larry was being portrayed by an actor they felt less moved by his experiences, less interested in imagining his perspectives, and felt less personal distress at his challenges. This possibility may be plausible given previous research linking client alexithymia with therapist feelings of boredom or frustration and a weaker therapeutic alliance (see Quilty et al., 2017). At the same time, a participant's ability to engage with fictional characters' experiences (represented through the fantasy subscale) may have been more relevant given Larry was not a real person and so this subscale contributed more to their success at case conceptualization. For similar reasons, this may be why we did not observe a relationship between participant agreeableness and the quality of conceptualizations.

Support was also not found for an inverse relationship between participant levels of mental health stigma and BACQS scores. There are several possible reasons for this. First, this measure was originally intended to assess mental health stigma around severe mental health issues, such as psychosis. Attitudes towards severe mental illness may simply be less relevant in the context of understanding a more common difficulty such as social anxiety. Second, we only utilized a small portion of the original scale and the items selected had a low internal consistency in this study which likely attenuated the strength of any relationship that may have been present.

Third, the distribution of participant scores on this measure was quite skewed with most participants scoring low on mental health stigma, which could also have impacted the detection of a relationship to scores on the BACQS.

In terms of discriminant validity through a lack of correlations with potential confounds, the BACQS performed relatively well on our two modest tests, at least when examining the correlations. Neither the length of conceptualizations nor the grade level of writing correlated significantly with the quality ratings obtained from applying the BACQS. At the very least it appears unlikely that the complexity of a participant's writing or how much was written strongly influenced coder scores.

For the regression predicting the BACQS total scores using the marker variables generated through the PCA three of the five components were identified as predicting the BACQS scores. These were a participant's "detached curiosity and openness", their "introspectiveness", and the length of their writing. These results would be in line with arguments that case conceptualization should be viewed as a cognitive task (Eells, 2011; Fernando et al., 2013; Goodyear, 1997), as well as research illustrating the impact of information processing abilities and heuristics on clinical decision making in general (Featherston et al., 2020; Whelehan et al., 2020), and in the process of case conceptualization specifically (Falvey et al., 2005; Welfare & Borders, 2010).

It did not appear that the "affective empathy/distress", or "warm empathy" components were predictive of BACQS scores after accounting for shared variance with the three components which did. Additionally, length of conceptualizations was found to be a significant predictor. These findings (that empathy was not, and that length of conceptualization was, predictive of conceptualization quality) went against our hypotheses, though it is worth noting the strength of the relationship between BACQS scores and conceptualization length was relatively small. Indeed, in retrospect, this relationship might be expected given that higher scores on the BACQS may require somewhat more elaborated conceptualizations.

In the between-group analyses the BACQS also received some support for its construct validity as students in psychology were able to produce higher quality case conceptualizations compared to students in other fields of study. It appears plausible that the field-specific knowledge students in psychology gain may contribute to their ability to understand and conceptualize mental health difficulties more successfully. Alternatively, traits or prior experiences that lead someone to study psychology may be the main contributors to this difference.

In analyzing differences in BACQS scores across English as first or second language participants we did find a difference. This result could be due to differences across the two groups, for example the writing skill of the participants, which would not be desirable. Ideally, the BACQS should be sensitive to the quality of ideas in a conceptualization versus the quality of how they are expressed. However, this difference in scores could also be due to cultural differences in how mental health difficulties are understood, as familiarity with western perspectives on mental health have been shown to vary across cultures (Altweck et al., 2015). Or, it may reflect more difficulty non-English as first language speakers had retaining or working with information from the video session due to language barriers.

Considering the regression analysis, the between group comparisons, and the correlations between variables to the BACQS scores, a generally promising picture of the construct validity of the scale emerges. With varying degrees of robustness in the findings, there are signs that participants who had higher academic achievement, greater interest and familiarity with mental health concepts, higher openness and need for cognition, and who provided longer conceptualizations tended to have higher scoring conceptualizations. We can also see signs that some aspects of empathy, a participant's tendency to engage with fictional characters and their level of empathy for Larry specifically, may be related to conceptualization quality. It also appears possible, and as a somewhat unexpected finding, that a cluster of variables representing a participant's tendency towards introversion, interest in mental health concepts, and ability to express themselves with more sophisticated writing influences ratings of case conceptualization quality.

In a series of studies by Gollwitzer and Bargh (2018) introversion was found to predict lay understanding of, and ability to accurately predict, social psychological behaviour. In their paper the authors discuss that introverts may be less likely to follow motivated patterns of thinking that are self-promoting, self-deceptive, or generally positively skewed in their social reasoning due to their more internal (rather than socially oriented) focus. In return, introverts may be more likely to generate realistic predictions about the social behaviour of others. Interestingly, and to some degree paralleling this research, the authors also found evidence that higher academic achievement, intelligence, and cognitive curiosity also predicted greater skill in understanding social psychological behaviour. They also discuss the possibility, and some preliminary evidence from their research, that some of these same traits may be relevant to a person's ability to understand individuals (person perception) rather than social groups. Much of this appears consistent with our findings, and also may help explain our findings around the role of introversion in case-conceptualization ability.

In reflecting on the results from this study, there also appears to be a link between the variables which were found to relate to case conceptualization quality and how many clinical psychology programs evaluate potential applicants. Applicants to clinical psychology programs in Canada are often selected in part based on their past academic performance (GPA), their performance on standardized tests which (among other things) assess vocabulary, writing ability, and familiarity with psychology knowledge (e.g. the Graduate Records Exam and Psychology Graduate Records Exam), and also based on an applicant's research and volunteer experiences in areas related to mental health. Applicants are also typically expected to be empathetic and interesting in the wellbeing of others, and to be curious and inquisitive. This overlap between this study's findings and these criteria may be another reason to view the BACQS's performance as promising.

Limitations and Future Directions

This study had several noteworthy limitations. Using an analogue population allowed for a larger sample of participants to be recruited in comparison to recruiting local clinical psychologists, however the study sample remains somewhat small. This issue is further compounded by some missing data (participant GPA) and the exclusion of cases with extreme influence from some analyses. Additionally, while the stronger conceptualizations obtained in this study provide a reasonable analog to those of junior clinical psychology students (which is a situation where the BACQS could eventually find use), others appear much poorer than what could be reasonably expected from that population, and quite different than what should be expected from independent clinical psychologists. That said, in Appendix 6 some examples of the high and low scoring conceptualizations from this study are shown (with participant permission) and the higher scoring conceptualizations appear quite sophisticated and insightful, and at face value appear to overlap with how clinical psychologists might understand Larry's difficulties.

Self-selection into the study was another limitation as it likely influenced the characteristics of our sample. For example, the level of mental health stigma in this sample appeared skewed and may be lower than what could be found in the general population. This sample may have also tended to be higher on the big five traits of openness and agreeableness, possibly reflected in the higher mean scores for those scales versus the others on the big five measure. This may have limited our ability to detect some relationships and could have influenced our results.

Vignette studies have also been recognized to have some limitations in case conceptualization research. For example, the information is presented unilaterally at a single time without a chance for a more organic evolution of the conceptualization (Kuyken, et al., 2008). In these situations the creator of a conceptualization also cannot pose their own questions or influence the information to which they have access. The unilateral presentation of information in this study could have limited how much participant traits influenced the quality of their conceptualizations, as they could not direct the questions being asked of Larry and instead had to use the information the therapist considered worth asking about. It is also possible that the questions being asked of Larry in the interview suggested something about the therapist's own conceptualization which may have reduced the variability in conceptualization quality of the participants as it could have guided their understanding. This also is a way the study procedure does not generalize well to real world situations, where case conceptualizations are often seen as evolving over time to fit with new information and changes in understanding.

Features of Larry's case may have also impacted the study outcomes in ways this study cannot easily address or explore. For example, empathy and perspective taking appears to be modulated by the level of similarity between a person and the target of their empathy in terms of values, emotional states, and previous experiences (Atzil-Slonim et al., 2019; Batson et al., 1996; Chung & Bemak, 2002). Larry being a socially anxious university student may have been particularly easy for some participants to understand based on their own experiences or similar traits, or even based on their relationships to individuals who experience social anxiety.

This could have influenced the pattern of relationships between the traits we assessed and BACQS scores by allowing some participants to draw upon their personal knowledge/experiences versus their ability to think about mental health difficulties in general while conceptualizing. This may be another possible explanation for why more introverted participants appeared to be more successful at conceptualizing Larry. They may have understood his social anxiety as similar to their introverted natures; though, it might be expected that a stronger correlation between the BACQS scores and neuroticism would have been observed if success in conceptualizing related to a participant's similarity to Larry.

Together, these points about this study's limitations suggest useful modifications should a similar study be conducted in the future. It may be beneficial to prepare several varied cases for participants to conceptualize. It may also be useful to assess how similar participants feel their own experiences are to those cases and to see if this relates to the quality of their case conceptualizations. If the resources are available, it may be also be useful to have the participants interview a mock client in real time or have a digital system that reveals specific information to participants based on questions they ask. This may increase the degree to which a participant’s own characteristics have an impact on the case conceptualizations obtained at the end of this process.

Finally, the study has some weaknesses in the measures utilized. Having more of the coding team conduct recode coding and the coding of the number of key points from the "master conceptualization" in participant conceptualizations would have made these tests more robust. It may also have been useful to include fewer but sounder measures in the study as the instances where shortened or newly created measures were used limited the confidence that can be placed in their validity. In the case of the stigma scale, we can also see this negatively impacted the reliability of the scale. Related to this idea, while need for cognition appears relevant to case conceptualization it may have been more useful to include a measure of psychological mindedness, given how important this trait was argued to be in case conceptualizing.

Despite these limitations, through this study support was found for the inter-rater and recode reliability of the BACQS, the scale's high in internal consistency, and even some modest signs of construct validity were obtained. It was also interesting to find that participant empathy was not universally tied to the ratings of conceptualization quality as obtained from applying the BACQS. This could lend support to the idea that case conceptualization is a more cognitively mediated process, though more work is needed to demonstrate the validity of the scale before interpreting this result with much confidence. Moving forward into Study 2, efforts were made to further extend these promising results by examining the performance of the BACQS on a clinically relevant yet understudied form of conceptualization that also adds to the scale's broad applicability.

Study 2: Applying the BACQS to Conceptualization Diagrams of a Mock Psychotherapy Case

The performance of the BACQS on the written conceptualizations from Study 1 was a positive first step in validating the measure. Written conceptualizations appeared to be coded with adequate reliability (when considering the coding team as whole and the overall scale score), the scale had strong internal consistency, and some evidence of construct and criterion validity was obtained. However, one of the main goals in creating the BACQS was for it to be applicable across a wide variety of contexts and types of conceptualizations. This study focused on demonstrating that the coding scheme could be applied to visually represented, rather than written, case conceptualizations. This form of case conceptualization is relevant to clinical and supervision settings and appears understudied in contrast to their frequent use and potential to influence courses of psychotherapy.

Case conceptualization diagrams are visual representations of case conceptualizations that are often used by clinicians in their sessions with clients and in meetings between supervisors and their trainees in clinical psychology training programs. A case conceptualization diagram typically is drawn by hand using boxes or bubbles that contain main ideas and lines or arrows that link these boxes or bubbles and show relationships of various kinds. With clients they are often used as a visual reference around which a shared understanding of the presenting problem can be built and elaborated on (Kuyken et al., 2008), particularly in early sessions of psychotherapy. They can then also serve to illustrate how and where psychological interventions can be applied to address the factors maintaining the problem.

Often these diagrams start from nomothetic models for various presenting mental health difficulties that are informed by psychological research. These nomothetic models are then often adapted, tailored, modified, or expanded upon to fit with a client's unique circumstances, comorbid mental health difficulties, life experiences, and other characteristics. Many evidence-based psychological interventions utilize case conceptualization diagrams in this way. For example, in Multisystemic Therapy (MST; Henggeler et al., 2002) diagrams known as fit circles are created to highlight drivers of problematic behaviours and to make decisions around treatment priorities. When shared with a client, worked upon collaboratively, and used to guide interventions conceptualization diagrams appear likely to have some impact on the success of a course of psychotherapy, and so being able to evaluate their quality appears important.

Case conceptualization diagrams also find use in supervision contexts. Clinical psychology programs emphasize supervision as a central method for helping trainee clinicians reach clinical competencies, hone their skills, and prepare for independent practice (Falender & Shafranske, 2014). In meetings, clinical supervisors and their supervisees work together to make treatment and diagnostic decisions which often interact with efforts to create and refine a case conceptualization. While not always used, case conceptualization diagrams are a common way of going about conceptualizing cases with trainees. Indeed, guidelines have been published outlining some effective ways a supervisor can navigate this process with their trainees (Liese & Esterline, 2015). As such, it appears supervisors could eventually utilize the BACQS in these situations to help evaluate a trainee’s skills in conceptualizing and expressing that conceptualization.

Concept maps are ways of visually demonstrating the relationships between various concepts when exploring a complex topic and were originally developed as a method of assessing children's knowledge of scientific concepts (Novak, 1990). Their structure is nearly identical to conceptualization diagrams, with the addition of descriptive phrases added to the linking arrows to help clarify the nature of relationships between different concepts. In post-secondary academic settings there is a growing interest in the use of concept maps (De Simone, 2007) possibly due to some researchers finding evidence that their use in learning settings aids knowledge consolidation (Jaafarpour et al., 2016; Ortega-Tudela et al., 2019) or, as others have suggested that concept maps could be effective ways of testing how well students are grasping the relationships in the topic being learned (Llinás et al., 2018; Nicoll, 2001). These relate to the points made above about conceptualization diagrams and were further reasons to suspect that conceptualization diagrams and their quality may impact outcomes in psychotherapy, and that evaluating their quality utilizing the BACQS would be a useful application of the scale.

Aims and Hypotheses

The main aims of this study were to:

1) Evaluate the inter-rater reliability of the BACQS on a sample of case conceptualization diagrams

2) Evaluate the recode reliability of the BACQS on this sample of case conceptualization diagrams

3) Evaluate the internal consistency of the BACQS and the inter-item correlations to assess whether the scale's internal structure remains consistent with an overall measure of conceptualization quality when applied to this form of conceptualization

4) Assess how the BACQS scores relate to other variables to search for evidence of construct validity

Our main hypotheses were:

1) The BACQS would demonstrate adequate inter-rater reliability for all seven items and the scale total score.

2) The BACQS would demonstrate good recode reliability.

3) The BACQS would demonstrate high internal consistency and the seven items would relate to each other in a coherent pattern suggestive of a scale assessing overall conceptualization quality.

4) The BACQS would demonstrate initial evidence of convergent and discriminant validity through associations between the total scale and other variables. Specifically, BACQS scores would:

> 4a) be positively correlated with participant GPA
>
> 4b) be positively correlated with student interest in and familiarity with mental health and psychotherapy.
>
> 4c) be positively associated with need for cognition
>
> 4d) be positively associated with participant empathy
>
> 4e) be positively correlated with the personality traits of open-mindedness and agreeableness
>
> 4f) be positively correlated with BACQS scores from the matching written conceptualizations obtained in Study 1
>
> 4g) be negatively correlated with mental health stigma

4h) be positively correlated with the BACQS scores obtained from the written conceptualizations

4i) not be strongly associated with the complexity of the conceptualization diagram

4j) be higher in participants who are currently studying psychology

Methods

Participants

The sample of participants in this study is the same as those from Study 1.

Procedure

The procedure for this study is also the same as in Study 1, though as the focus of this study is the case conceptualization diagrams rather than the written conceptualizations, two variables which were not highlighted in Study 1 are detailed in the measures section, and just below the instructions participants received in advance of creating their conceptualization diagram is presented.

Immediately after watching the video recording of the simulated session participants were instructed on what a concept map is and how they are created (we did not use the term conceptualization diagram with the participants, instead indicating they would make a concept map of Larry's difficulties as these appear roughly analogous). Participants were instructed that:

> *Concept maps are visual representations of a topic that show relationships between different components, sub-topics, or parts of the main topic. Concept maps are made of short phrases or individual words (concepts) inside of boxes or bubbles. Each box or bubble contains one concept. Lines and arrows are used to connect boxes or bubbles. A*

> *word next to a linking line explains the relationship between the two concepts. In a concept map all bubbles/boxes should connect to the main body of the concept map through at least one line, no part of the concept map should be completely detached from the rest of the concept map.*

Participants were also told that the number of bubbles/boxes can vary across concept maps, and that the shape and structure of concept maps vary as well. Several example concept maps (of unrelated topics) were presented to illustrate these points. Participants were then asked to create a concept map of Larry's difficulties (on a piece of 11 x 17-inch paper) so that the map explained the causes of the difficulties and could guide a course of psychotherapy. As with the written conceptualization task, participants could review their notes while creating the conceptualization diagram, did not have to include all the information in the video, and they could include information not explicitly stated in the video.

Measures

The majority of the items and measures from Study 1 were also utilized in this study. However, where the length of the written conceptualizations (word count), and the grade level of the writing in the written conceptualizations were used in Study 1, a different method for quantifying the superficial complexity and “length” of the conceptualization diagrams was required here (for a similar test of divergent validity). The analog for these that appeared most appropriate was obtained by counting how many bubbles/boxes and linking lines were contained in a conceptualization diagram (referred to later as the "number of elements").

Missing Data and Data Quality

The same eight participants who were excluded from Study 1 due to misunderstanding important aspects of study instructions or failure to reach the end of the study were also excluded from this study. One participant was also excluded as they had produced their conceptualization diagram on the regular sized paper meant for note taking, and not the larger paper provided. Participants were also excluded for poor data quality in a very similar fashion to Study 1. However, participants were not excluded based on whether their written conceptualizations were inaccurate to their internal understanding of Larry, but instead if they reported their conceptualization diagrams were not accurate to their internal understanding of Larry's difficulties. Altogether, six participants were excluded due to potentially problematic data. The scores obtained from coding their conceptualizations were still included in reliability and internal consistency analyses but were excluded from the other analyses. Just as in Study 1, coders returned complete scores for all the available case conceptualization diagrams and very few data points (0.58%) were missing from participant self-report measures (aside from the self-reported GPA). Where participant data was missing, we used pairwise deletion from analyses.

Data analyses

Very similar analyses were conducted in this study to those in Study 1. For the correlations between the BACQS scores and other variables extreme bivariate outliers, and in the regression multivariate outliers (more than three inter-quartile ranges away from the mean DF beta score), were again removed from those analyses. Results without the outliers removed can be found in Appendix 7. In the regression analysis we retained the components from Study 1, though we replaced the length of conceptualizations component with a standardized variable for the number of elements. GPA was again excluded due to the missing data.

Results

Descriptive Statistics

The descriptive statistics for the study variables, aside from coded BACQS scores, are presented below.

Table 9

Descriptive Statistics for Study 2 Variables

	n	Mean *(SD)*	Range	Skewness	Kurtosis
Participant GPA	36	78.16 (6.13)	70-90	.32	-.69
Number of Elements	66	41.23 (17.04)	18-89	.86	.41
Need For Cognition	66	62.97 (10.56)	29-82	-.57	.47
Interpersonal Reactivity Index	65	99.74 (11.16)	69-123	-.36	.03
Fantasy	66	26.02 (5.14)	16-35	-.31	-.82
Perspective Taking	66	26.11 (4.60)	12-35	-.62	.79
Empathic Concern	66	28.00 (4.00)	18-35	-.51	-.08
Personal Distress	65	19.66 (4.91)	8-33	.38	.26
Short Big Five Inventory					
Openness	66	23.26 (4.08)	14-30	-.31	-.76
Conscientiousness	66	19.62 (4.39)	7-28	-.22	.01
Extraversion	66	19.47 (5.19)	7-29	-.41	-.81
Agreeableness	66	22.98 (3.93)	11-30	-.50	.33
Neuroticism	66	18.00 (5.80)	8-30	.25	-.84
Stigma Items Total	65	11.63 (3.83)	6-23	.95	.52
Mental Health Skill and Interest	64	14.36 (3.37)	4-20	-.47	-.25
Empathy for Larry	65	50.44(9.41)	20-63	-1.37	2.04

Descriptive Statistics for BACQS Scores

Below, the descriptive statistics for BACQS scores obtained from coding the conceptualization diagrams are presented.

Table 10

Descriptive Statistics for Case Conceptualization Diagram BACQS Scores

	n	Mean *(SD)*	Range	Skewness	Kurtosis
BACQS total coder 1	71	19.66 (4.28)	10-29	.03	-.63
BACQS total coder 2	71	20.82 (5.50)	10-32	.21	-.70
BACQS total coder 3	71	20.10 (4.87)	10-32	-.03	-.45
All coders average					
Item 1: General Quality...	71	2.64 (.69)	1.00-4.00	.25	-.69
Item 2: Psych. Mindedness	71	2.83 (.74)	1.00-4.33	.19	-.64
Item 3: Integration	71	3.22 (.70)	1.33-4.67	.09	-.70
Item 4: Differentiation	71	2.90 (.72)	1.67-4.00	.10	-.47
Item 5: Core Mechanism	71	2.73 (.75)	1.33-4.33	.16	-.54
Item 6: Strengths...	71	2.88 (.29)	1.67-4.00	-.68	1.57
Item 7: Theoretical and...	71	2.99 (.62)	1.33-4.33	.41	-.21
BACQS total score	71	20.19 (3.94)	11.00-27.67	.19	-.54

Inter-Rater Reliability

Table 11

BACQS Inter-rater Reliability Estimates (Three Raters, Case Conceptualization Diagrams)

	Absolute Agreement ICC (95% CI)	Consistency ICC (95% CI)
Item 1: General Quality Impression		
Average measures	.61 (.42-.74)	.62 (.43-.75)
Single measures	.34 (.19-.49)	.35 (.20-.50)
Item 2: Psychological Mindedness		
Average measures	.71 (.57-.81)	.72 (.59-.82)
Single measures	.45 (.30-.59)	.47 (.33-.60)
Item 3: Integration		
Average measures	.56 (.34-.71)	.59 (.39-.73)
Single measures	.29 (.15-.45)	.33 (.18-.48)
Item 4: Differentiation/Breadth		
Average measures	.67 (.51-.78)	.67 (.51-.78)
Single measures	.40 (.26-.55)	.40 (.26-.55)
Item 5: Core Mechanism Identified		
Average measures	.50 (.26-.67)	.51 (.27-.68)
Single measures	.25(.11-.41)	.25 (.11-.41)
Item 6: Strengths and Resiliency Focus		
Average measures	.29 (-.02-.52)	.31 (-.01-.55)
Single measures	.12 (-.01-.27)	.13 (-.02-.29)
Item 7: Theoretical and Logical Grounding		
Average measures	.69(.54-.80)	.69 (.55-.80)
Single measures	.42(.28-.56)	.43 (.29-.57)
Total Score		
Average measures	.72 (.59-.82)	.73 (.59-.82)
Single measures	.46 (.32-.60)	.47 (.30-.60)

Recode Reliability

Five months after coding the conceptualization diagrams a random selection of 20 were recoded by the author. The correlation between original and retest scores was *r* = .76, p < .001.

Internal Consistency and Item Inter-Correlations

The internal consistency of the BACQS scores when applied to the conceptualization diagrams was very similar to the results from study one, in that it was also in the excellent range (Cronbach's alpha = .94). Similar to Study 1 a lower range in scores was found for item 6, with 81% of rater's scores being a 3, which may explain why this item's corrected item-total correlation remained somewhat low (.45). The inter-item correlations are presented below, and appear similar to the results from Study 1, though item 6 had a stronger pattern of correlations to the other items.

Table 12

BACQS Item correlations (Case Conceptualization Diagrams)

	1	2	3	4	5	6
Item 1: Overall qual...	-					
Item 2: Psych minded.	.87***	-				
Item 3: Integration	.84***	.80***	-			
Item 4: Different...	.79***	.70***	.58***	-		
Item 5: Core mech...	.84***	.85***	.71***	.62***	-	
Item 6: Strengths...	.42***	.44***	.48***	.21	.41**	-
Item 7: Theoretical...	.88***	.88***	.79***	.63***	.85***	.47***

note: *** p < .001, ** p < .01, * p < .05

Construct Validity

Table 13

BACQS Correlations to Study Scales and Variables (Case Conceptualization Diagrams)

Variable	*n*	*r*
Convergent Validity Variables		
Participant GPA	35	.31
Mental health skill and interest	62	.14
Need For Cognition total	64	.09
Empathy for Larry	62	.23
Written conceptualization scores	61	.59***
Interpersonal Reactivity Index: scale total		
Fantasy subscale	62	.31*
Perspective taking subscale	61	.19
Empathic concern subscale	61	.20
Personal distress subscale	58	.07
Openness	63	.10
Agreeableness	64	-.05
Mental health stigma items	63	-.08
Discriminant Validity Variables		
Number of Elements	63	.35**
Familiar with concept maps	63	.02
Other Big Five Traits		
Conscientiousness	60	-.05
Extraversion	62	-.09
Neuroticism	61	.07

note: *** p < .001, ** p < .01, * p < .05

As indicated in the table above, when BACQS total scores from participant written conceptualizations were correlated with the BACQS total scores from matching conceptualization diagrams the relationship was found to be significant. At the item level this remained true as well. All items demonstrated significant correlations between the scores from drawn and written conceptualizations. The strength of the correlations ranged from r = .34 to r = .56, and all were significant at the p < .01 level.

Just as in Study 1, a simultaneous multiple regression was conducted, though this time predicting BACQS total scores obtained from coding the conceptualization diagrams. The goal was again to determine whether certain variables were more robust predictors while accounting for shared variance. The "warm empathy" component, "affective empathy/distress" component, "detached openness and curiosity" component, and "introspection" components were included in the analysis. The number of words component was substituted for a standardized number of elements score, as this appeared to be the most similar variable. Data points with high influence, based on extreme DF beta values (as previously described), were removed prior to conducting the regression (a total of 13 participants).

The model was significant, R^2 = .27, *SE* = 3.41, $F(5,46) = 3.38$, p = .01, with the "warm empathy" component and standardized number of elements being the only two significant predictors. However, the magnitude of the relationship for the introspection variable was similar to those and may be worth interpreting as well, despite the statistical non-significance. Detailed results of the regression are below in Table 14.

Table 14

Regression Predicting BACQS Total Scores (Case Conceptualization Diagrams)

Variable	β	t
Warm Empathy	.28*	*2.05*
Affective Empathy/Distress	-.01	-0.05
Detached Openness	.08	-0.47
Introspection	.32	1.87

Number of Elements	.35**	2.74

Note: * = significant at the *p* < .05 level, ** = significant at the *p* < .01 level.

Between Group Analyses

Unlike in Study 1, the BACQS total scores did not differ significantly between students in psychology (*M* = 21.03, *SD* = 3.48) and non-psychology (*M* = 19.98, *SD* = 4.00) majors of study, *t* (64) = 1.10, *p* = .28. Scores also did not differ between English as first language (*M* = 20.82, *SD* = 3.34) or ESL (*M* = 19.14, *SD* = 3.86) participants, *t* (64) = 1.59, *p* = .12.

Discussion

Did the BACQS Demonstrate Adequate Inter-Rater Reliability for All Seven Items and the Scale Total Score?

Similar to in Study 1, the values from the single measures tests suggested that a single rater's scores do not reach adequate levels of reliability. In contrast to the written conceptualizations, the average measures reliabilities were not consistently above the "poor" range (above .50) across the items. In particular, item 6, the strength and resiliency item, had very low inter-rater reliability. It may be that this poor reliability is due in part to a restriction of range issue. This can be seen in the mean of this item, which was close to the "default" of 3 (where a conceptualization lacks notably positive or negative content), and the standard deviation which was much smaller than those of the other items. Indeed, across all the scores raters gave for this item (across the 71 conceptualization diagrams) 80.03% were scores of 3.

Items 3 and 5 also had lower reliability here, versus for the written conceptualizations, though these do not appear to be caused by restriction of range issues. This difference may be more attributable to the nature of the conceptualization diagrams themselves, which is returned to later. Of most importance, it does still appear that the scale total score reaches adequate reliability, even considering the 95% confidence interval. This lends some support to coders being able to use the BACQS reliably when coding conceptualization diagrams, but with more limits on at what level the scores can be interpreted (item versus scale total)

Did the BACQS Demonstrate High Recode Reliability?

Although only a subsample of the conceptualization diagrams were-coded a second time, the recode reliability obtained was still promising and similar to the results obtained in Study 1. It appears that even over a five-month period to forget the original scoring, the values obtained in the recoding remained relatively consistent.

Did the BACQS Demonstrate Good Internal Consistency and a Coherent Pattern of Inter-Item Correlations Suggestive of a Scale Assessing Overall Conceptualization Quality?

The internal consistency and item-total correlations obtained in this study were quite strong and indicate the measure's items contribute well to an overall score. The pattern of inter-item correlations appeared similar to those in Study 1, though the Strengths and Resiliency item had statistically significant associations to the other items; however, we did not test to see if the changes in strengths of correlations were significant.

Did the BACQS Demonstrate Evidence of Convergent and Divergent Validity Through Associations Between the Total Scale and Other Variables?

At first glance, the set of correlations between the BACQS scores and other variables in this study was less supportive of the scale's convergent and divergent validity when compared to Study 1. For example, in the correlational analyses the magnitudes of the relationships between

BACQS scores and convergent validity variables were generally weaker in this study, and the number of elements in a conceptualization was significantly correlated to the BACQS total scores (against the hypothesis for the scale demonstrating divergent validity from this variable) However, as demonstrated in Figure 2 in Appendix 8 the confidence intervals around the point estimates for the correlations overlap significantly between the two studies (given their modest sample size). As such, the differences in magnitudes (for the correlations) are unlikely to be statistically significant, which in turn, limits the ability to interpret these differences with confidence. With that consideration in mind, some variables (for example GPA and empathy for Larry) still demonstrated positive correlations that may have been approaching statistical significance but were likely underpowered due to the sample size. Additionally, the fantasy subscale of the IRI remained significantly related to the BACQS scores with a similar magnitude in the correlation.

The relationship between BACQS scores and the number of elements was also not particularly strong, suggesting it is unlikely coders were simply evaluating the conceptualization diagrams based on their visual complexity (number of elements and lines). Should a more robust test in the future research reveal that the number of elements does maintain a relationship to conceptualization quality ratings while number of words does not, that that could also be taken as an indication that the number of elements and the number of words in a drawn versus written conceptualization may not be as analogous as was initially assumed in this study. Longer writing to convey a conceptualization, when produced by a novice, may not reliably result in more useful ideas being added (hence a lower correlation), whereas increasing the number of key concepts in bubbles and the number of connections between them may be more closely tied to thoughtful conceptualizing, and in turn conceptualization quality. This also ties to some points below exploring possible differences between the visual versus language-based demands of this task.

Also against our hypotheses, we also did not see a difference between the psychology and non-psychology student's conceptualization quality as evaluated by the BACQS. This finding goes counter to previous research where psychology students have performed better on tasks emulating aspects of psychotherapy (Miles et al., 2020), and their demonstrated higher levels of psychological mindedness (Trudeau & Reich, 1995). Conversely, there was also no difference between English as first language and English not as first language speakers' scores, where there had been in Study 1. On this latter point, this may have been due to there being less of a reliance on fluency or writing ability in the conceptualization diagrams.

In the regression analysis only two of the variables entered were significantly predictive of the BACQS scores, the warm empathy component and the number of elements in a conceptualization diagram. However, the "introspection" component had nearly the same beta coefficient score as in Study 1 and the lack of statistical significance for this variable may have been the result of the pattern of inter-correlations between the predictors.

Several possible explanations for the poorer indications of validity and differing pattern of results in this study were considered. The first is it quite possible that given a smaller sample size and slightly different set of cases included the results may have shifted simply due to random error. It is also possible that the lower reliability for some items may have limited the ability to detect some smaller relationships versus what was found in Study 1. It may be that the items which had lower reliability in this study (integration, core mechanism, and strength and resilience focus) were those which contributed most to the relationship between BACQS scores and some of the other variables (such as a person's skill and familiarity with mental health, need for cognition, openness, and major of study).

Another possibility stems from the order of events in the study procedure. In particular, that participants created the case conceptualization diagrams before the written conceptualizations may have had some impact. Although participants were asked to use the same understanding to create both conceptualizations, it may be that a person's warm empathy has more impact on an immediately produced case conceptualization while the written conceptualization which was produced 15 minutes later may have been more impacted by their openness and need for cognition, as they continue to think about and refine their understanding. This might reflect participants sequentially using the “fast” and then “slow” routes of information processing (alternatively known as system 1 and system 2), where the rapid system draws more on emotional reactions, intuition, and habit, and the slower system - deliberate reasoning and deliberation (Geoffry et al, 2022; Kahneman, 2011). A point against this possibility is that scores on the conceptualization diagrams were actually higher on average than those obtained from the written conceptualizations.

Similar to the “fast” versus “slow” systems explanation above, the pattern of results may reflect well-studied differences in how humans process visual versus language-based information. A large body of research has consistently identified that the processing of visual versus language-based information draws on somewhat distinct cognitive systems (Carroll, 1993; Taub and McGrew, 2014). It may be that the change from conceptualization diagrams to written conceptualizations required participants to draw on different cognitive processes, which in turn, related to empathy and need for cognition differently.

In seeking to explain the differences between studies in the pattern of results, we can also consider the feedback from the coding team on their experiences coding the two forms of conceptualizations. The three coders all reported more difficulty coding the diagrams in comparison to coding written conceptualizations. The coders noted that the conceptualization diagrams and the ideas in them were more abstract and more ambiguous. They also tended to present fewer specific details and there was less clarity about what was considered important. As a result, coders felt they often made more inferences about what exactly the creator was intending to convey. Coders also reported that navigating through the more complex and visually busy concept maps was challenging.

Overall, these reflections suggested that there may simply be more error variance in these BACQS scores, that coders may have tended to give more benefit of the doubt when scoring (also possibly explaining the higher scores on the maps), and that factors such as the complexity or detail in a concept map might have biased scores more (which may explain why the differentiation item had more impact and why the number of elements appeared more strongly and directly linked to the BACQS scores here than in Study 1).

A synthesis of several of these possibilities could also be true, it may be that the broader level and more abstract diagrams, produced before the written conceptualizations, were influenced more by a participant's empathy rather than cognitive traits of openness and need for cognition, and also relied less on a participant's experience with psychological or mental health concepts. This may still have been more difficult for coders to score, causing them to rely more on heuristics in their scoring, for example the amount of information available, increasing how much the differentiation item correlated with the other items of the scale.

Limitations and Future Directions

This study faces the same limitations found in Study 1. There were issues relating to external validity to more realistic therapy situations and in the generalizability of results to a sample of actual psychologists and their performance. This study also shares the weaknesses found in several of the measures utilized and with participant self-selection into the study. Additionally, this study revealed potential impacts of the study procedure (the temporal ordering of which conceptualization was produced first remaining consistent across participants).

In the future several changes could benefit further applications of the BACQS to conceptualization diagrams. For example, depending on the study design, participants could be directed to produce conceptualizations that fall within a certain range of elements, possibly reducing the impact of this variable on the quality, allowing for other variables to demonstrate more of their impact. The instructions for creating the conceptualization diagrams could also be altered in some way so that participants have some method of better illustrating the importance of different elements included (for example colour coding or labeling elements by importance or the strength of relationships, or something similar). Instructions could also be provided that could increase how detailed/specific the conceptualizations were overall, so less coder inference might be required. Researchers could also ensure that the conceptualization diagrams are produced with less time pressure and/or with a first draft, so that participants have more time to make refinements, and see if this enhances the relationship of the conceptualization scores to other variables, such as openness. Alternatively, randomizing whether participants produce a written or diagrammed conceptualization first (should both be included) may allow for any impacts of this sequencing to be detected.

Another point relevant to the difficulties above is that the task of producing a conceptualization diagram without concurrently discussing it is unlike the typical process as it unfolds in either supervision meetings or psychotherapy. In either of these circumstances, it is typical for a case conceptualization diagram to be created in the context of a discussion which adds further details to the simple visual outline. As is returned to in the general discussion later, future research applying the BACQS to case conceptualization diagrams would likely benefit from adding in this important aspect of the process. Indeed, as noted in the introduction above an outline for how therapists and supervisors can structure conversations with clients and therapists (respectively) has been published and could serve as a useful guide in such future studies (Liese & Esterline, 2015)

Finally, coders would likely benefit from specific training and practice with case conceptualization diagrams. The coders in this study were trained on written conceptualizations and did not receive any additional specific training on how to evaluate conceptualization diagrams with the BACQS. Incorporating an opportunity for coders to collaboratively code several diagrams together and practice on a sample of diagram conceptualizations could potentially increase the reliability and validity of the scores obtained and would be in line with other research involving coding conceptualizations (Eells, 1998; Kuyken et al., 2016). This process could also lead to some revisions in the coding manual that would allow it to be more easily applied to conceptualization diagrams in the future.

Conclusions

In this study the BACQS demonstrated positive performance in some important areas. Several items, and most importantly the scale total score, achieved adequate inter-rater reliability. The scale demonstrated strong internal consistency, and a generally strong pattern of inter-item correlations. The scale's recode reliability appears similar to what was obtained in Study 1 and demonstrated consistency over relatively long periods of time (hopefully enough to allow the coder time to forget the original coding).

Another success was that the BACQS scores from this study correlated significantly with those from Study 1, which may be a sign that across both studies the BACQS is tapping into the quality of the ideas in a conceptualization, rather than simply the form of their expression. That this was true for the scale total score and each item is also a positive, as it may indicate that each item’s codable content can be expressed both in writing and a case conceptualization diagram (supporting the scale’s broad applicability). Finally, although the interpretation is somewhat more difficult than in Study 1, there were still some tentative indicators of the scale’s validity. For example, empathy remained a significant predictor of BACQS scores and several other variables had modest and perhaps simply underpowered relationships (GPA, the "introspection" component).

In general; however, it appears the performance of the BACQS on the case conceptualization diagrams was poorer than on the written conceptualizations from Study 1. Across several items the scale had lower inter-rater reliability and the scores appeared to demonstrate less validity through relationships to other variables and from between-group comparisons. Additionally, and perhaps related to these points, the coding team reported more difficulty with, and less confidence in, their scoring.

Some of the inter-rater reliability issues could be attributable to the content of this sample, for example that so few contained useful information for coding the strengths or resiliency item as anything but the default score. Should a sample with more strengths and resiliency relevant information be obtained in the future, higher reliability might have been obtained. Similarly, changes to the instructions for creating the concept maps (for example having a way for participants to express the relative importance or strengths of certain relationships) might have allowed coders to more reliably evaluate whether a core mechanism had been identified, and how well integrated the diagrams were.

Slight improvements in several of the items would have shifted them into the adequate range and so it is also possible that some of the inter-rater reliability issues could have been remedied by providing more training on coding conceptualization diagrams or by the addition of one or two more coders. While not ideal solutions for a tool designed to have ease of use and low requirements in terms of time or resources, this form of conceptualization may simply require slightly more effort to code accurately.

Study 3: Applying the BACQS to Conceptualizations Obtained from Case Study Articles

Although the first two studies provided data supporting the BACQS as a reliable and internally consistent measure with good recode reliability and (more so in Study 1) some indicators of construct validity, the limits of those studies made it difficult to generalize the results to more naturalistic contexts. An attempt was made to address this issue in this study (Study 3). Whereas the previous two studies used conceptualizations from an analogue and somewhat artificial experimental design, and were limited to describing a single client's case, this study utilized conceptualizations obtained from case study articles published to a variety of online academic journals. This allowed conceptualizations generated from actual courses of psychotherapy, which were used to guide real courses of treatment, and which were generated by practicing clinicians, to be coded. In addition to assessing a body of much more ecologically valid conceptualizations, a wide diversity of conceptualizations was also purposefully sought out (in terms of presenting problems, client characteristics, and psychotherapeutic approaches). As was begun in Study 2, where the applicability of the BACQS to another form of conceptualization was explored, the diversity in conceptualizations in this study allowed for another important examination of the scale’s broad applicability.

Case studies are important training and education tools in clinical psychology and other medical fields (Mackrill & Iwakabe, 2013). In them an author typically details a single client/patient and their care, highlighting aspects of diagnosis/assessment, treatment planning, and/or treatment that other clinicians may benefit from learning about vicariously (some case study articles illustrate a common theme across several similar cases, or focus on psychotherapy groups, or family therapy cases, but this study is focused on single case studies).

One method of presenting case studies for a wide audience, given their importance in medicine, is by publishing them in an academic journal. Some journals are exclusively dedicated to publishing case studies while others are primarily research focused journals that occasionally include relevant case studies. Some journals focus primarily on a specific type of psychotherapy or specific types of mental health disorders, while others are quite broad in focus. Many of these journals have conventions and guidelines for the reporting of case studies, (for example see Clinical Case Studies, 2021). Most psychotherapy case study articles include sections focused on client history and background, assessment and diagnoses, treatment approach and response, and most important for this study, a section on the case conceptualization. Also important for this context, most published articles describe the characteristics of their client (though often some details are slightly changed to help maintain confidentiality) and the psychotherapeutic approach (i.e. cognitive-behavioural therapy, psychodynamic therapy, integrative).

Aims and Hypotheses

In this study the main aim was to test the inter-rater reliability and internal consistency of the BACQS on a more naturalistic sample of conceptualizations. Another goal was to apply the BACQS to a sample of conceptualizations that represented a wide range of presenting mental health problems, psychotherapeutic approaches, and client characteristics (though limiting this to adult cases of individual psychotherapy).

Our hypotheses were:

1) The BACQS would demonstrate adequate inter-rater reliability for the 7 items and the scale total score.

2) The BACQS would demonstrate good recode reliability.

3) The BACQS would demonstrate good internal consistency and the items would relate to each other in a coherent pattern suggestive of a scale assessing overall conceptualization quality.

4) The BACQS scores would demonstrate signs of construct validity by:

4a) demonstrating little relationship to conceptualization length

4b) demonstrating little relationship to the grade level of the writing in conceptualizations

4c) being significantly higher than those of the written conceptualizations generated by undergraduates in Study 1

Methods

Case Studies

A search for case studies was conducted through the University of Waterloo's academic journal search engine and through Google Scholar. Searches utilized combinations of the following terms: case study, case example, psychotherapy, and clinical psychology paired with various psychotherapeutic orientations (cognitive-behavioural therapy, CBT, psychodynamic, emotion focused therapy, EFT, DBT, etc.). From the search results 60 articles were selected semi-randomly while trying to ensure variety in the case studies and based on indications from the title that they focused on single cases of psychotherapy. From these 34 were identified as meeting our inclusion criteria of focusing on single cases of individual adult psychotherapy and containing dedicated sections for describing the case conceptualization. In a few instances these sections were described using terms such as "patient psychodynamics". The majority of the articles were published after 2010, though several were published in the 2000s, and one in 1999. Our final sample of conceptualizations came from the following journals:

- *Clinical Case Studies*

-*Clinical Psychology Psychotherapy*

-*Journal of Clinical Psychology*

-*Journal of the American Psychoanalytic Association*

- *Psychodynamic Practice*

- *Psychotherapy: Theory, Research, Practice, Training*

- *Social and Behavioural Sciences*

-*International Journal of Cognitive Therapy*

- *Pragmatic Case Studies in Psychotherapy*

- *International Journal of Offender Therapy and Comparative Criminology*

- *The British Psychological Society*

- *Behaviour Modification*

-*Cognitive and Behavioural Practice*

- *Behavioural and Cognitive Psychotherapy*

-*Person-Centered and Experiential Psychotherapies*

- *Eating Weight Disorders*

- *International Journal of Integrative Psychotherapy*

- *American Journal of Psychotherapy*

After the characteristics of the clients and their treatments were recorded from the articles the conceptualizations (and only the conceptualizations) were copied into a separate document. The coders (whose training on applying the BACQS was described earlier) then coded all the conceptualizations independently and with the same instructions as in the previous two studies.

Data Analyses and Cleaning

Inter-rater reliability was again obtained through a series of intraclass-correlation coefficients, exactly as in studies 1 and 2. Internal consistency was evaluated with Cronbach's alpha, corrected item-total correlations, and inter-item correlations. We evaluated recode reliability through a Pearson's correlation between the original and re-test scores of one coder.

We evaluated construct validity through correlations between the BACQS total scores and the length of conceptualizations and BACQS total scores and the grade level of the writing. We also performed an independent samples t-test on the BACQS scores comparing the scores obtained from these case study conceptualizations to those from the written conceptualizations from Study 1.

Results

Sample Characteristics

In the table below, the characteristics of the clients and their courses of psychotherapy from the case study articles are presented.

Table 15

Characteristics of the Sample of Case Studies

Characteristic	*M (SD)*	*Range*	*% of Sample*
Client Age	38 (18.53)	18-93	
Client Sex			
Female			73.5
Male			26.5
Client Ethnicity			
Not reported			47
Caucasian			29
Non-Caucasian			23.5
Client Presenting Problem			
Personality disorder			15
Anxiety disorder			15
Depression			12
Eating disorder			12
Psychosis/bipolar disorders			12
Several comorbid T_x targets			12
Other*			22
Psychotherapeutic Orientation			
Psychodynamic			32.4
Cognitive-behavioural			27
Other†			41

Note: Approximate ages were sometimes given "a client in their 50s", resulting in some error in our estimation of the mean age of the sample.

* Obsessive-compulsive disorder, posttraumatic stress disorder, chronic pain, paraphilia, relational trauma

† Dialectical behavioural, interpersonal, acceptance and commitment, process experiential, narrative, compassion focused, integrative

Descriptive Statistics

The conceptualizations averaged a grade level for their writing of 14.48 (*SD* = 2.15) and a word count of 434.50 (*SD* = 243.20). In Table 16 below we report the descriptive statistics for the BACQS scores obtained from coding the case study conceptualizations.

Table 16

Descriptive Statistics for Case Study Conceptualization BACQS Scores

	n	Mean *(SD)*	Range	Skewness	Kurtosis
BACQS total coder 1	34	21.91 (4.12)	10-30	-.46	.95
BACQS total coder 2	34	22.62 (4.53)	10-31	-.21	-.19
BACQS total coder 3	34	22.79 (5.80)	10-32	-.42	-.70
All coders average					
Item 1: General Quality...	34	3.09 (.71)	1.33-4.67	-.11	.30
Item 2: Psych. Mindedness	34	3.42 (.68)	1.33-4.67	-.84	1.50
Item 3: Integration	34	3.25 (.73)	1.33-4.67	-.52	.20
Item 4: Differentiation	34	2.95 (.74)	1.00-4.67	-.14	.72
Item 5: Core Mechanism	34	3.19 (.70)	1.33-4.67	-.20	.34
Item 6: Strengths. and Resil...	34	3.03 (.38)	2.33-4.33	1.26	3.14
Item 7: Theoretical and...	34	3.52 (.54)	2.00-4.67	-.41	.87
BACQS total score	34	20.19 (3.57)	11.00-28.67	-.93	1.80

Inter-Rater Reliability

As in studies 1 and 2 we assessed the inter-rater reliability of the BACQS items and scale total score by calculating a series of intraclass correlation coefficients. The same models were utilized as in the prior studies and the method for obtaining a scale total score is the same here as well. The results of these ICC analyses are reported in Table 17 below.

Table 17

BACQS Inter-rater Reliability Estimates (Three Raters, Case Study Conceptualizations)

	Absolute Agreement ICC (95% CI)	Consistency ICC (95% CI)
Item 1: General Quality Impression		
Average measures	.61 (.32-.79)	.63 (.35-.80)
Single measures	.34 (.14-.55)	.36 (.15-.58)
Item 2: Psychological Mindedness		
Average measures	.61 (.32-.97)	.62 (.33-.80)
Single measures	.34 (.14-.56)	.36 (.14-.57)
Item 3: Integration		
Average measures	.52 (.15-.74)	.51 (.14-.74)
Single measures	.26 (.05-.49)	.26 (.05-.49)
Item 4: Differentiation/Breadth		
Average measures	.67 (.42-.82)	.69 (.46-.84))
Single measures	.40 (.19-.60)	.43 (.22-.61)
Item 5: Core Mechanism Identified		
Average measures	.21 (-.36-.57)	.22 (-.39-.58)
Single measures	.08(-.10-.31)	.08 (-.10-.32)
Item 6: Strengths and Resiliency Focus		
Average measures	.43 (-.02-.70)	.42 (-.01-.43)
Single measures	.20 (-.01-.43)	.20 (-.02-.69)
Item 7: Theoretical and Logical Grounding		
Average measures	.41(.01-.68)	.44 (.01-.70)
Single measures	.19 (.00-.41)	.21 (.00-.44)
Total Score		
Average measures	.57 (.25-.77)	.57 (.24-.77)
Single measures	.31 (.10-.53)	.31 (.10-.53)

Several possible causes for the markedly poorer performance of item 5, core mechanism identified, were explored. Using feedback from the coders two hypotheses were generated. The first was that the psychodynamic conceptualizations, which were often quite long, comprehensive, and included terminology unfamiliar to the coders may have been more difficult to evaluate for a main mechanism being emphasized. The undergraduate level coders also reflected that some conceptualizations did not address how to improve the disorders clients were facing. This situation appeared most likely to be the case when severe mental health issues (such as bipolar disorder or schizophrenia) were the focus, as these are often not directly expected to improve through psychotherapy, but instead clients are assisted in coping with, managing, or adapting to these difficulties.

With this slightly more nuanced treatment focus coders may have struggled to reach consensus on whether a core mechanism was being addressed in treatment. Emerging models of psychopathology may provide indirect support for this possibility (that bipolar disorders and schizophrenia may have distinct features making their conceptualization more difficult) as some emerging models attempting to identify commonalities and differences across diagnoses (through factor analyses) identified these disorders (along with cluster A personality disorders) as falling into a distinct “thought disorder” category (Kotov et al., 2017).

These two possibilities were tested by removing the psychodynamic conceptualizations or severe mental health conceptualizations from the sample and evaluating the impact on the ICCs obtained for the core mechanism item (item 5). In both cases the ICCs improved, however a significant ICC was only obtained when the psychodynamic conceptualizations were removed. With these conceptualizations removed the absolute agreement ICC for this item rose to .42.

Recode Reliability

Six months after initially coding the sample of conceptualizations the author recoded the sample a second time. The recode correlation for the scale total score was r = .75, p< .001.

Similar to Study 1, the changes in scores from original coding to recoding were examined; 52.9% of item scores remained the same, 41.9% differed by one point, and the remaining 5.2% differed by more than one point.

Internal Consistency and Item Inter-Correlations

The internal consistency of the scale was quite high, though slightly lower than in the previous two studies (Cronbach's alpha = .89). Most corrected item-total correlations were between *r* = .71 and *r* = .91, save for item 6 (Strengths and Resiliency) at *r* = .31 and item 7 (Theoretical and Logical Grounding), at *r* = .45.

Inter-item correlations are reported in Table 18 below. Similar to the results from Study 1, Item 6 (the Strengths and Resiliency item) appeared to have a lower set of correlations to the other items. The variation in scores was again low for this item, fully 75% of the scores from raters were 3s.

Table 18

BACQS Item correlations (Case Study Conceptualizations)

	1	2	3	4	5	6
Item 1: Overall qual...	-					
Item 2: Psych minded.	.79***	-				
Item 3: Integration	.85***	.72***	-			
Item 4: Different...	.76***	.71***	.70***	-		
Item 5: Core mech...	.76***	.74***	.78***	.57***	-	
Item 6: Strengths...	.36*	.23	.29	.10	.14	-
Item 7: Theoretical...	.48**	.35*	.39*	.31	.42*	.47**

note: *** *p* < .001, ** *p* < .01, * *p* < .05

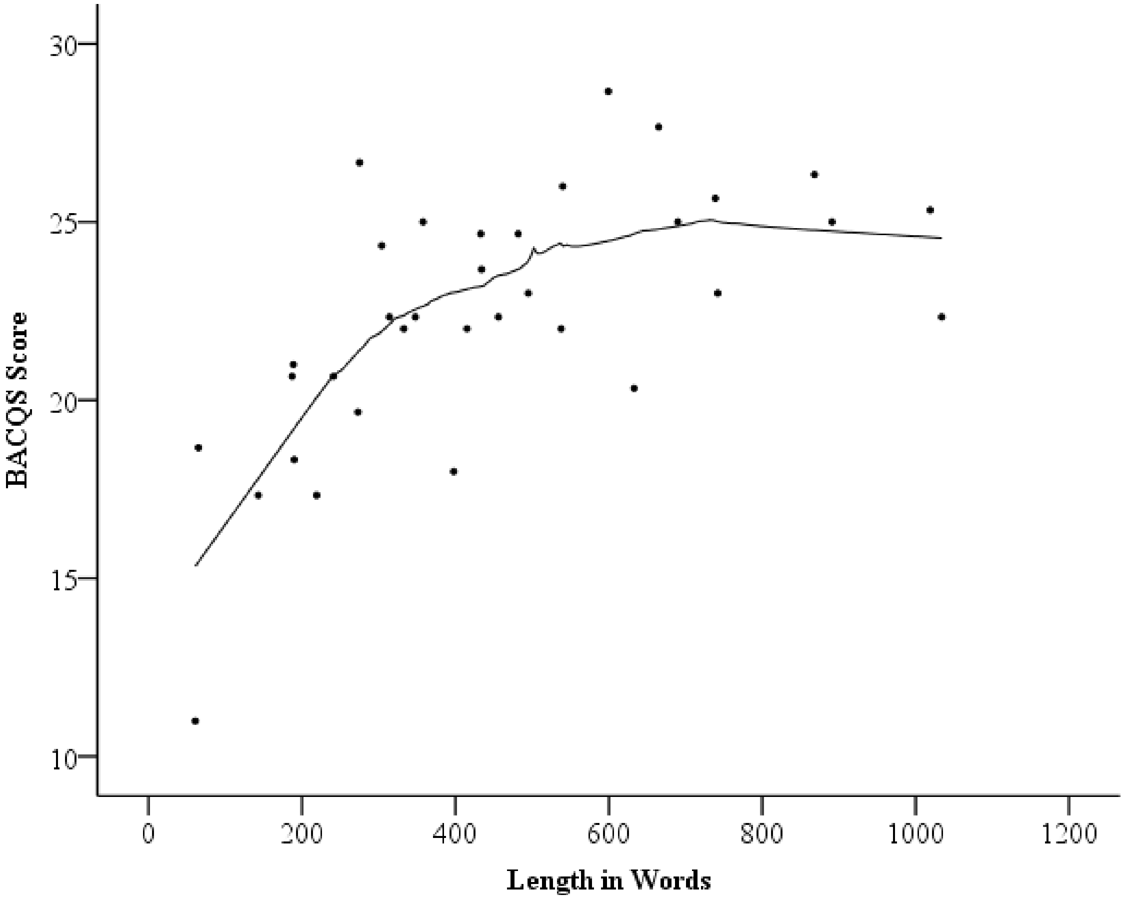
Construct Validity

As a preliminary examination of the performance of the BACQS on a sample of case conceptualizations obtained from real psychotherapy cases we did not have many avenues for assessing the scale's validity. However, similar to the previous two studies we were interested in testing how much the sophistication of the writing and the length of the conceptualizations could have been impacting the scoring of the coders. The grade level of the writing did not correlate significantly with the BAQCS scores ($r = -.17$, $p = .35$), while the length of conceptualizations did, $r = .61$, $p < .001$ (even after removing two cases with extreme influence [as was done in the previous studies using DF betas] $r = .64$, $p < .001$).

Upon inspecting a scatter plot the relationship between the length of the conceptualizations and BACQS scores appeared to be quadratic. A hierarchical multiple regression was performed to test this possibility. The first step included the linear term for number of words predicting the BACQS scores, $F(1,30) = 27.00$, $p<.001$, with an R^2 of .47. The addition of a quadratic term improved the model significantly $\Delta F(1,29) = 10.49$, $p = .003$, with an ΔR^2 of .14. The curve of this quadratic relationship suggests that BACQS scores only increased with the length of a conceptualization up a length of approximately 700 words, after which increasing length no longer contributed to better scores. This relationship is also demonstrated in Figure 1, a scatter plot of the BACQS scores by the length of the conceptualization (with a Loess fit line). This relationship was still observed after trimming out multivariate outliers with high influence, as determined by extreme DF beta values. The model of the first step of the hierarchical regression including the linear term, $F(1,28) = 16.68$, $p< 001$, with an R^2 of .35, was improved significantly by adding the quadratic term for the number of words $\Delta F(1,27) = 7.82$, $p = .009$, with an ΔR^2 of .14.

FIGURE 1

Scatter Plot of BACQS Scores by Length of Conceptualization (Case Study Conceptualizations)



Cross Sample Comparisons

As Study 1 utilized conceptualizations generated by undergraduate participants and this study focused on conceptualizations generated by professional clinicians, comparisons across the two samples were conducted as another examination of the scale's validity. Differences between scores for the individual items of the BACQS and the scale total across the two samples were examined through independent samples t-tests. Levene's tests showed equal variances across the items and the scale totals. Results of the t-tests are below in Table 19, they demonstrate significantly higher scores for the case study conceptualizations, except for item 4, where no significant difference was found.

TABLE 19

Independent Samples t-tests on BACQS Scores across Study 1 and Study 3

Variable		*M(SD)*	*t*	*p*
Item 1: General Quality Impression	Study 1 Study 3	2.37 (.70) 3.09 (.71)	4.94	<.001***
Item 2: Psychological Mindedness	Study 1 Study 3	2.57 (.72) 3.42 (.68)	5.79	<.001***
Item 3: Integration	Study 1 Study 3	2.79 (.79) 3.25 (.73)	2.83	.006**
Item 4: Differentiation/Breadth	Study 1 Study 3	2.72 (.60) 2.95 (.74)	1.70	.09
Item 5: Core Mechanism Identified	Study 1 Study 3	2.67 (.75) 3.19 (.70)	3.38	.001**
Item 6: Strength and Resiliency Focus	Study 1 Study 3	2.85 (.44) 3.03 (.38)	2.08	.04*
Item 7: Theoretical and Logical Grounding	Study 1 Study 3	2.69 (.67) 3.52 (.54)	6.38	<.001***
Scale Total	Study 1 Study 3	18.65 (3.85) 22.44 (3.57)	4.83	<.001***

Note: degrees of freedom equalled 104 for all analyses.

Discussion

The purpose of this study was to test the performance of the BACQS on a sample of case conceptualizations generated and utilized in actual cases of psychotherapy. As in the previous two studies the scale's inter-rater and recode reliability was examined, as was the scale's internal consistency. Efforts were also made to assess the scale's construct validity, though fewer means of testing this were available in this study.

In terms of the inter-rater reliability analyses, the scale demonstrated a more varied performance, with some items obtaining similar ICC values to those in the other two studies, while others scored much lower. Due to the smaller sample size the confidence intervals around the ICC values are much broader which made interpreting this varied performance more difficult. However, it appears unlikely that the individual items at the level of a single coder would reach adequate levels of reliability, which echoes the results from the previous two studies. The results did not suggest any notable differences between the absolute agreement or consistency ICCs. This suggested that the lower reliabilities for some items was not simply due to systematic bias among some raters.

If the broad confidence intervals are set aside, and considering the average measures ICCs, the initial quality impression and the differentiation/breadth items performed nearly the same as in the prior two studies, and within the acceptable range for inter-rater reliability. The psychological mindedness and integration items also performed similarly to the prior studies, if slightly lower, but also in the acceptable range. Items five, six, and seven were lower than in the previous studies and below the adequate range. The scale total score fell above .50, but significantly lower than in the previous two studies. In trying to understand this poorer performance, we can consider several factors that may be at play.

First, it appears that for item five (core mechanism identified) raters struggled the most when coding psychodynamic conceptualizations, as demonstrated by how this item’s ICC value improved greatly when those conceptualizations were removed. While this result should be considered quite tentatively given the small sample and ad hoc nature of the test, it appears plausible that something about this form of conceptualization was more difficult to code for this item and this coding team. The coding team (which may have less exposure to psychodynamic theories than to cognitive-behavioural approaches to therapy, for example) reported the psychodynamic conceptualizations were more comprehensive, difficult to understand, and contained more orientation specific terms. From the writer’s perspective they also tended to take long term developmental perspectives which illustrated client problems as the result of many cumulative and interacting events. Together, these factors may have made it more difficult to determine whether a “core mechanism” had been identified.

Second, the lower inter-rater reliability for item 6 (strengths and resiliency focus) may once again have been impacted by a restriction in range. As in the other studies few conceptualizations strayed from a score of three, the default score for when a conceptualization does not clearly include positive or negative elements related to this item. With little variation in scores the impact of instances of disagreement were likely inflated. Again, this could be less of an issue in a sample of conceptualizations where more strengths tended to be incorporated.

Third, a similar restriction of range issue, though less severe, also appears to possibly be impacting the scores on item 7 (theoretical and logical grounding). This item had the highest mean and the second lowest standard deviation. Retrospectively, that a sample of published case conceptualizations produced by professional therapists would tend to be judged as logical and grounded in psychological theory appears reasonable, but this may still have negatively impacted the inter-rater reliability.

Finally, with the somewhat lower ICC of the scale total score it appears possible that with a few small changes this score could be improved significantly. First, if there was more variety in the strength and resiliency content of conceptualizations or the logical and theoretical grounding scores, the reliability of these items could increase, in turn increasing the reliability of the total score. As such, in these cases it may be more a matter of the content being coded, rather than the scale itself which is limiting the inter-rater reliability. Another method of improving the scale's performance would be to simply omit psychodynamic conceptualizations from any sample, or when they are included, to ensure coders are more familiar with psychodynamic theories and approaches (which presumably would be the case amongst graduate level or higher coders). Another solution might simply be to exclude item 5 from the scale total should psychodynamic conceptualizations be included in the sample being coded. However, considering that identifying a core mechanism to target through therapy is a major proposed use for conceptualizations (Johnstone et al., 2011), this solution appears less than ideal.

The other aspects of the BACQS's performance appear to have been relatively strong in this study. For example, that over a period of six months a strong correlation was still found suggests the scale produced consistent scores over repeated applications. Also, that the internal consistency of the scale remained high, and that this is consistent with the previous two studies, continued to lend credit to the use of a scale total score and to the idea that the BACQS may function well in producing a score of a conceptualization's overall quality. Item 6's (Strengths and Resiliency item) lower corrected item-total and inter-item correlations may also be the result of restriction of range, though until a sample of conceptualizations is obtained that shows more variation in content around this item this possibility is still unclear.

Even in terms of signs of validity, the results from this study were promising. In examining the relationship between the grade level of the writing and BACQS scores no significant relationship was obtained, suggesting the complexity of writing was not used as a heuristic for evaluating the conceptualization quality. Although a larger than hoped for correlation between the length of the conceptualizations and their scores was obtained, the quadratic nature of this relationship was somewhat reassuring. At a certain point it appears longer conceptualizations no longer added to a coder's perception of quality, suggesting it was not simply that coders resorted to a 'longer is better' heuristic. Instead, it may be that additional information in a conceptualization is useful, particularly when produced by a professional clinician, and contributes to quality, but that this is only true to a point. It also appears that the shortest conceptualizations could be contributing to this quadratic relationship. Under approximately 300 words, the quality of conceptualizations dropped significantly. It may be that such short bodies of text are generally unable to contain enough information to reach high levels of quality.

Our test across the samples of written conceptualizations was also promising, in that professionally produced case conceptualizations scored significantly higher than those produced by undergraduates (save for item 4). Some research into whether deliberate training can improve the quality of conceptualizations has suggested this is possible, with even brief (2-hr) training sessions improving quality (Kendjelic & Eells, 2007). Past research has also found differences in case conceptualizations across levels of therapist experience, with higher experience resulting in more well-organized, inferential, and useful conceptualizations (Eells & Lombart, 2003; Eells et al., 2011).

These findings are in line with our results, where the case-study conceptualizations were scored as more integrated, stronger in terms of core-mechanisms identified, more psychologically minded, and of overall higher quality than the undergraduate conceptualizations. Thus, although the results of the cross-group comparisons in this study may not be surprising, they are consistent with evidence that case conceptualization is a skill that can be learned, and that they should be produced at a higher level of quality by a group of professionals who use them regularly. However, these comparisons are limited by the two samples of conceptualizations being generated in different circumstances, different cases being conceptualized, different amounts of time to produce the conceptualization, etc. (and so this being an imperfect comparison). Nonetheless it would have been a significant threat to the scale's validity had the difference in quality of the conceptualizations been non-significant (or worse, if the undergraduates had 'outperformed' the professionals).

Limitations

This study is limited in several important ways. The first is a small sample size. Given more time or resources to code a larger sample, useful statistical comparisons between subgroups of the sample could have been made, confidence intervals around the ICCs would have likely been tighter and more easy to interpret, and in general more confidence could be placed in the results of this first application of the BACQS to a more realistic set of conceptualizations (as this small sample may not have been representative of the larger body of case study conceptualizations that exist). However, the results from coding this small sample were still informative and highlighted that the BACQS may have some areas of weakness to address (which are discussed more shortly) before committing more resources and time into a more rigorous and larger study.

Secondly, although these conceptualizations should be quite similar to the conceptualizations produced by psychotherapists in actual practice, they still may differ in some ways. First, this sample included conceptualizations from after courses of psychotherapy had concluded. Conceptualizations are typically seen as evolving over a course of therapy, for example becoming more explanatory and including more breadth of information as a therapist and client share more information and refine their understanding of a client's difficulties (Kuyken et al., 2008) As such, the conceptualizations obtained in this study may represent "final versions" of the conceptualizations, and not the exact conceptualizations that guided earlier portions of these therapies. These conceptualizations are also likely to be different as they were selected by their authors, presumably over other therapy cases. Something about these cases was considered exceptional or worth sharing, which seems to indicate they differ somehow from more typical cases. Related to this, the authors may have emphasized or deemphasized aspects of the cases to better fit with the goal of educating or sharing insights through case studies.

Finally, our tests of validity were not particularly robust. It is promising that the quality scores obtained did not correlate with the grade level of the writing, only correlated with conceptualization length up to a point, and were significantly higher than the sample from Study 1, but more direct methods of assessing the construct validity are still needed.

Future Directions

There are several fruitful future directions that would build off this particular study. One would be to simply repeat the study with a larger and more randomly obtained sample and see if the results of this study replicate. As mentioned, a larger sample would also allow for tighter confidence intervals around the ICC values and might allow for comparisons to be made between subgroups. For example, it might be useful to determine whether item reliabilities or average scores on various items of the BACQS vary across therapeutic orientations or certain presenting problems. This would possibly allow for more useful revisions to the BACQS to be made, to address areas of weakness that have been more firmly established.

Another option to pursue would be to extend the findings into other client age ranges (child, adolescent), other modes of psychotherapy (group, couple), or even conceptualizations obtained from assessment cases, all contexts where case conceptualizations are also seen to be important. Ideally the BACQS should be able to score these types of conceptualizations, but even if the scale did not perform strongly, this could inform changes that might need to be made to the scale or the coding manual, or potentially suggest that related versions of the BACQS that are specifically tailored for those other contexts might need to be developed. This may be a longer-term goal; however, as it may be more useful to focus on adult focused conceptualizations until the scale has been further refined.

In any case, it appears likely that coders would benefit from further training or practice than were offered in advance of this study. It may be that in future uses of the BACQS on case study conceptualizations the coding team’s performance would be improved by learning more about psychodynamic terms and therapy, learning about psychotherapeutic approaches to severe mental health issues, and having a chance to practice coding on a subsample of these more ecologically valid conceptualizations.

Conclusions

This study served as an informative application of the BACQS to a more real-world sample of conceptualizations. The ICC values were suggestive of adequate reliability across most items and the scale total score, and recode reliability was strong. In terms of internal consistency and inter-item correlations the scale also appeared to be functioning relatively well as an overall quality measure. In terms of examinations of the scale's validity, it appears that the length of conceptualizations only correlates with quality up to a certain length, suggesting raters were not simply following a 'more is better' rule. Additionally, the cross-sample comparisons appeared consistent with the idea that professional mental health workers should have stronger case conceptualizations than non-professionals.

The weaknesses that the scale demonstrated in some areas, such as low inter-rater reliability for several items, and a correlation with length of conceptualizations are tempered by points discussed earlier. In some cases, there may simply not have been enough variation in certain markers of quality (for example strengths and resiliency related information) in the sample of conceptualizations and targeted training may have been needed to increase coder familiarity with psychodynamic theory and terms (to boost inter-rater reliability for item 5, for example). Nonetheless, even for other items the inter-rater reliability generally appeared lower in this study and more work appears needed to boost this important aspect of the scale's performance.

The small sample and methodological limitations should limit the confidence placed in the findings (both positive and negative for the scale), but broadly the results appear promising and give further support to this measure being mostly able to fulfill one of its major purposes, applicability across a wide variety of theoretical orientations and presenting problems.

General Discussion

Summary of Results

Above, the development and initial testing of a novel measure of case conceptualization quality was presented. This measure, the BACQS, was developed to address the need for an accessible and broadly applicable measure of case conceptualization quality, as highlighted by Bucci and colleagues (2016). The development of the BACQS was guided by existing research and clinical literature on the proposed benefits and functions of case conceptualizations, research suggesting links between psychological mindedness and case conceptualizations, and existing measures of case conceptualization quality (considering both the strengths and limitations of those measures). It is hoped that this research provides a solid foundation for future studies using the BACQS, where it may find use in exploring the link between case conceptualization quality and therapy outcomes. This is seen as a crucial, yet outstanding question by researchers in this field (Bieling & Kuyken, 2003). It is also hoped that the promising findings obtained may enable it to be applied in other settings where evaluating case conceptualization quality may be useful, such as in clinical supervision and training.

Considering the results obtained across the studies presented, the BACQS demonstrated adequate to good performance on most tests of its psychometric properties. The scale consistently demonstrated good recode reliability across the three studies and over relatively long periods. Where item level scores differed, the majority of the time this was only by one point. As outlined in their review, Bucci and colleagues (2016) highlighted how few measures of case-conceptualization quality have been assessed for evidence of test-retest reliability and the importance of this type of psychometric evaluation. The scale also consistently demonstrated high internal consistency. Even for a short scale with only seven items the internal consistency was in the excellent or very good ranges. Initial concerns that the scale may have had too high an internal consistency, suggesting some items could be redundant, were quelled somewhat by the performance of the scale in Study 3, which was the most ecologically valid application. However, this issue could return should all items improve in their inter-rater reliability, as this could also strengthen the correlations between some items, and in turn boost the internal consistency.

The scale also performed adequately on most tests of inter-rater reliability, and particularly promising was the inter-rater reliability of the scale's total score regularly falling above the adequate range. However, this is mostly true for the inter-rater reliability of the pool of coders, typically a single coder's ratings were not sufficiently reliable to use on their own. For the several instances where the scale items failed to reach adequate inter-rater reliability, even for the pooled reliability, it appeared plausible that restriction in range issues were contributing to this, particularly for the strengths and resiliency item. We also found that for the poorest inter-rater reliability score we obtained, the core mechanism item in Study 3, the issues may be focused mostly on psychodynamic conceptualizations and possibly some more biologically driven disorders. While not ideal, this suggests that it may be possible to learn from the poor performance in those areas and make adjustments to remediate the issue. For example, by providing more training to coders on psychodynamic theory and terms or reworking aspects of the coding manual.

In terms of reworking the coding manual, one solution may be to consult psychodynamic psychologists and seek their input on how to adjust the measure to include more relevant and better described examples for psychodynamic conceptualizations. Item 5 (core mechanism identified) appears particularly important to address. It may be, for example, that a psychodynamic mechanism is evidenced when a lifespan-focused conceptualization includes recurring patterns which, without explicitly being stated, imply an unconscious process to target in therapy. The manual may have to emphasize that the reader of a psychodynamic conceptualization should infer this type of mechanism more actively, versus in a CBT conceptualization where the mechanism is more likely to be explicitly stated.

Another perspective on this issue would be that if a coder had graduate level training focused on psychodynamic psychotherapy they may be able to apply the BACQS to psychodynamic conceptualizations as the scale is, due to their increased familiarity with the terms and concepts of that orientation. If true, this might suggest that the BACQS has a certain form of broad applicability (it could be used across therapeutic orientations), but only when coders apply the scheme to conceptualizations from theoretical orientations with which they are quite familiar. While not ideal, this could still enable this measure to be used in contexts such as supervision, psychotherapy, or research, with this caveat. It is also worth noting that in considering the issues with some areas of inter-rater reliability the coders did not have opportunities to practice on examples from the samples in studies 1-3, and this could suggest another way of improving the scale’s performance in the future that should not be particularly burdensome.

Our initial examinations of the scale’s validity focused mostly on construct validity and provided some modest support to the BACQS, but with limitations. The most robust evidence for construct validity was obtained in Study 1, in that the quality of the written conceptualizations positively correlated to other variables that we hypothesized they could relate to, such as mental health interest and skill, openness, need for cognition, some forms of empathy, and the academic performance of participants. Regression analysis also supported some of these variables' relationships to the BACQS, though as the number of variables had to be reduced through a PCA (due to our sample size) and given that GPA was excluded due to missing data, these results are still quite tentative. The one test of criterion validity was also in Study 1 but was promising. As conceptualizations increased in the number of "key features" from the master conceptualization of the fictional case they included their quality score according to the BACQS increased significantly.

The results from Study 2 also lend some support to the scale's construct validity, as empathy, a participant's GPA, and possibly the "introspection" component, appear to be predictive of BACQS scores. Additionally, that the scores from the conceptualization diagrams correlated moderately with the scores from the written conceptualizations was another positive indicator of the scale's validity – as this may suggest the BACQS was coding elements of the quality of ideas that were demonstrated across both forms, versus merely coding the quality of their expression through the writing or diagrams themselves. However, when compared to the results from Study 1, the conceptualization diagrams were less consistently or strongly linked to the variables they were predicted to be, perhaps due to coding issues, the nature of the diagrams themselves, or both. However, as noted in the discussion section of that study the changes in magnitude of the correlations are unlikely to be significant and a larger sample would be needed in a replication of the study to draw firmer conclusions about these findings.

Looking across Studies 1 and 2 it appears that the qualities that predicted better performance on the BACQS were largely consistent with traits that are often used to evaluate applicants to clinical psychology programs and are in line with some other research linking these traits (such as need for cognition) to a person's ability to conceptualize or reason about human behaviour (Cacioppo & Petty, 1982; Cacioppo et al., 1996). While empathy did not consistently predict case conceptualization quality, this possibility should be further explored before any strong conclusions are drawn, as some aspects of empathy (for example the fantasy subscale or empathy for Larry in particular) did show some relationships to conceptualization quality in some analyses.

That psychology undergraduates produced conceptualizations of higher quality than non-psychology students also appears to support the validity of the BACQS, and suggests that either the traits that bring students to study psychology, and/or the knowledge they gain from their studies enable them to understand mental health issues more effectively (or at least those the simulated client was facing). Similarly, that the professionally created and published case conceptualizations scored higher than those produced by undergraduates also adds slightly to the picture of the BACQS as valid, given past research linking conceptualization skill to training and experience with psychotherapy (Eells & Lombart, 2003; Eells et al., 2011; Kendjelic & Eells, 2007), and given how problematic the opposite finding would have been. These findings are also consistent with the idea that psychological mindedness has been found to be higher in individuals studying psychology and is relevant to a career in clinical psychology (Farber et al., 2005; Trudeau & Reich, 1995; Westen et al., 1991), and it may be that differences in this skill/competency/trait-like construct across the two subgroups may be a driver of the differences in conceptualization quality.

We also did not find evidence that the BACQS scores for the written conceptualizations were overly influenced by possible confounding variables such as the grade level of writing and length of conceptualizations. Though, for the length of the conceptualizations there were some relationships in Study 1 and 3 that appeared to be either small, or to dissipate at after a certain point. This may indicate that although more information in a conceptualization can improve their quality up to a point, coders were not simply relying on a basic ‘more is better’ heuristic while coding.

In Study 2, the number of elements did appear more strongly related to BACQS scores which could be a discriminant validity threat, and/or another sign that coding the conceptualization diagrams was difficult, and more influenced by the level of detail included. However, as discussed before, it may also be that the number of elements is not very analogous to length in words of a conceptualization. Writers can extend their text without contributing meaningfully to a better explanation whereas adding in more concepts and links between them in a conceptualization diagram may be more suggestive of conceptualizing thoughtfully.

There were also some significant differences in Study 1 between participants with or without English as their first language in BACQS scores which may indicate that confounding variables could be influencing BACQS scores. From most to least problematic for the BACQS’ validity, this could suggest that participant writing fluency was impacting scores, that those who are English as second language speakers had more difficulty understanding some of Larry’s session and this impacted their conceptualizations, or that cross-cultural differences in mental health literacy or understanding (Altweck et al., 2015) were impacting conceptualizations. Including a measure of mental health literacy (Jorm et al., 1997) in future research utilizing the BACQS may help to clarify these possibilities, given this measure has been sensitive to cross-cultural differences in previous research (Wong et al., 2017).

Clinical Implications

One implication stems from the generally adequate reliability of the scale total score or the reliability pooled across raters. This could be seen as evidence that the BACQS could help a supervisor and supervisee team to determine whether a case conceptualization is strong enough to proceed into treatment or to otherwise guide discussions around case conceptualization development. Indeed, the overall feedback from our coding team was that most conceptualizations were relatively quick to code (approximately 10-15 minutes each) and so incorporating the BACQS into supervisor-supervisee discussions may not demand a great deal of extra time.

Another related clinical implication stemming from these studies is that the BACQS may be an appropriate tool to aid in cross-profession consultation, which may increasingly occur as mental health needs continue to grow across Canada and the world. Consultation (which also can occur intra-professionally, i.e. within clinical psychology) refers to instances where a clinician is encountering difficulties in their treatment with a client and they seek out to another clinician (often with relevant expertise and often without them meeting directly with the client) for guidance, insight, and problem solving. Some reasons for a clinician seeking consultation include when they are faced with a particularly complex clinical presentation, presenting problems which lie outside their area of expertise, or instances where progress in treatment is unexpectedly not being reached.

Given that the scale demonstrated reasonably good evidence of its broad applicability (with some caveats), and given its brevity, ease of use, and low material and resource cost, it may be a useful tool to frame cross-profession consultation around, at least in part. Other mental health professionals may have different theoretical backgrounds, different skills and approaches, and different areas of expertise that may contribute to difficulties communicating about aspects of treatment across professions. The BACQS may be able to serve as an adaptable lens through which psychologists could consult on the conceptualizations and psychotherapeutic treatment plans of other professionals.

Another potential clinical implication of this research is more incidental but replicates findings from other researchers well. There does appear to be a paucity of strengths or resiliency focused information included in most case conceptualizations, across the three studies only 11% of conceptualizations scored above 3 on the strengths item (scores above 3 were given when any meaningful strengths or resiliency relevant information was included). This suggests that beyond the general lack of confidence many clinicians feel towards their case conceptualization skills (Glidewell & Livert, 1992), they may specifically benefit from improving their ability to incorporate client strengths and examples of resiliency into their case conceptualizations.

Although the above clinical implications (and applications) may be intriguing and promising, they should be considered with some caution as more research is needed to validate the performance of the BACQS in clinical contexts. Importantly, it appears unlikely that a single supervisor’s ratings for a supervisee’s conceptualization would be reliable enough to utilize independently, particularly at the item level. As such, the BACQS should not yet be used in any formal supervision evaluations. It is possible that clinical psychologists could more reliably and effectively apply the BACQS in comparison to our undergraduate level coders, perhaps even reaching high enough reliability to use their scores without pooling across raters, but until further research is conducted it would be premature to assume this.

Research Implications

The performance of the BACQS in these three studies may help guide future researchers exploring the role of case conceptualizations in psychotherapy in several ways. First, it may be useful to consider extending and refining the training of coders to improve inter-rater reliability, particularly if undergraduate research assistants are employed as coders. Specifically, providing more background knowledge of psychodynamic theories (or as mentioned previously, refining the coding manual to better address the coding of psychodynamic conceptualizations) or around more biologically driven disorders could be useful. However, it may also be fruitful to shift from undergraduates to graduate level or higher coders to see if this improves the inter-rater reliability of the scale significantly enough to warrant sacrificing some of the convenience of using undergraduate research assistants. Related to these changes, future researchers may wish to retain a slightly larger coding team as three coders appeared slightly too low for some items to reach adequate pooled reliability.

Other research implications from this project return to points raised earlier. It appears restriction of range impacted areas of the coding, and so it may be useful for researchers to consider if changing instructions about the conceptualization task or including more relevant information in case vignettes (on strengths for example) could help. Training clinical psychology graduate students on case conceptualization (perhaps using the items of the BACQS as a guide) might also increase the tendency of clinicians to include strengths in their conceptualizations. Previous research has indicated that training on case conceptualizations does improve their quality (Kendjelic & Eells, 2007), lending some support to this possibility. Finally, conducting a semi-structured interview which focuses on a clinician’s case conceptualization, and which prompts for, but does not unduly pull for, the inclusion of strengths related information could be another solution.

Ultimately, the restriction of range issues for the strengths and resiliency item did appear quite persistent across studies and the possibility of dropping this item from the scale merits some discussion. An advantage of dropping the item would be that the scale's overall reliability and internal consistency would likely improve, particularly should other samples of conceptualizations also include little in the way of codable strengths and resiliency relevant information. Dropping this item would also shorten the scale, potentially increasing the ease of its application even further (though perhaps only marginally).

In considering reasons to retain this item, it did have small to moderate relationships to some other BACQS items in Study 2 and Study 3, despite the restriction of range issues, and it also performed adequately in terms of inter-rater reliability in the pilot study and Study 1. These results may be indications that the focus of this item (strengths and resiliency relevant information in conceptualizations) can be reliably coded and can have a meaningful relationship to the overall quality of a conceptualization, in the appropriate context. Given some evidence that incorporating strengths and examples of resiliency is a common occurrence for many clinicians in their therapeutic work (Scheel et al., 2012), it should be possible to elicit more strengths-based information in case conceptualizations in future studies. For these reasons, dropping item 6 may prematurely foreclose on a potentially useful indicator of quality.

Another less concrete justification for retaining the item may be found by returning to the importance of content validity during measure creation, which emphasizes that measures should assesses all important aspects of construct. Many authors writing on case conceptualization, some with empirical evidence (Capobianco, 2013; Welfare et al., 2013), support the idea that including strengths in a conceptualization improves their utility and quality (Johnstone et al., 2011; Kuyken et al., 2008; 2011). Similarly, indications that some individuals have had negative reactions to their case conceptualizations being shared (Chadwick et al., 2003; Evans & Parry, 1996) also speaks to the importance of including this item, given that hope and engagement are predictors of treatment outcomes (Coppock et al., 2010; Irving et al., 2004; Milovanov, 2017), which case conceptualizations are meant to improve.

The research in this project also highlights the importance of selecting robust measures (as discussed previously) and obtaining a large enough and representative sample of conceptualizations. Relating to the first point, across studies 1 and 2 only a relatively small proportion of the variance in case conceptualization quality could be explained through the variables included in the regression analyses. This suggests future researchers could continue to identify other distinct but relevant traits or characteristics that contribute to the variation in conceptualization quality. One important characteristic would likely be level of expertise and experience. Other research into case conceptualization quality has focused on differences between experts and novices and has found higher quality has generally correlated with increasing expertise and experience (Eells et al., 2005). Applying a similar analysis to scores obtained from the BACQS could be another method of testing the scale's validity and explaining variance in the conceptualization quality. Relating to the issue of obtaining a representative sample, future research would likely benefit from gathering conceptualizations from ongoing therapy cases and from sources beyond case study articles. It may be useful to gather samples from psychology training clinics, private practice psychology clinics, hospital settings, and other institutions where clinical psychologists provide psychotherapy. Although a truly random sample of conceptualizations would likely not be obtainable, these sites would likely offer a more representative sample of conceptualizations by drawing from locations of real-life, everyday clinical psychology practice. Ensuring the sample of conceptualizations was large enough (and diverse enough) to provide better opportunities for cross-group comparisons and a factor analysis of the scale items, etc., would also be ideal.

Limitations

While this research project focuses on areas of scale reliability and validity that have not frequently been examined in other measures of case conceptualization quality, there are a few important aspects of scale validation this research project could not address. Most important among these, is that this research project did not address questions about predictive validity. In some ways this would be the ultimate goal of the BACQS, ideally higher scores obtained for a case conceptualization's quality would predict faster, greater, or more sustained gains in courses of psychotherapy, or secondary benefits such as higher working alliance between client and therapist, or a greater sense of understanding of the issues and more confidence in forming a treatment plan.

Demonstrating these relationships would help lend some credibility to the claims made by many researchers and clinicians (as noted earlier) that a case conceptualization's quality impacts the treatment of mental health difficulties in meaningful ways. Despite not being able to include examinations of predictive validity in this research the important first step of demonstrating other forms of psychometric soundness (reliability, internal consistency, face validity, content validity, construct validity) could help open the door for future researchers to answer these important questions more directly.

An additional limitation of this project was that the BACQS was applied to conceptualizations gathered at single points in time and only from the perspective of the "therapist"/therapist. A significant critique of early case conceptualization research was that most researchers examined their impact without considering their evolving and co-created nature. The CCC-RS was specifically developed with this issue in mind (Kuyken et al., 2008; 2011), with items assessing the level of agreement and collaboration around the conceptualization, and items meant to assess whether the conceptualization is being refined, tested, and deepened over time (though as such, the measure may be seen as focusing more on processes involved in conceptualization, versus conceptualizations themselves). As a preliminary effort to assess conceptualizations through the application of the BACQS, beginning with the therapist's perspective and assessing the conceptualizations at a single time was seen as an appropriate starting point. Below, some ideas for future directions are discussed, including some for addressing this issue.

Future Directions

Aside from some of the more specific follow up studies, changes to the BACQS, or changes to study procedures outlined previously, several more general future directions appear useful to outline and consider here. First, it could be useful to examine the BACQS coding manual itself and re-evaluate whether any changes could help improve the agreement in ratings between coders. As one example, while most items varied little across the absolute agreement or consistency ICCs, item 1 (overall quality) did appear to show a difference across these two in some cases suggesting that this item may be more subject to coder biases towards more liberal or conservative scoring. As such, it may be worth expanding the descriptions for the quality levels of this item to help guide coders more consistently to the same scores. It may also be useful to consider whether changes in the wording of the BACQS coding scheme could be needed in advance of applying it to other psychodynamic conceptualizations, particularly the item on a core mechanism being articulated.

In terms of specific future research directions, the BACQS has several frontiers where it has not yet been tested at all, but which appear relevant. Some of these have been mentioned previously, but one which appears particularly important would be to see if this measure can be applied in supervisory settings. One possible way of doing this would be to use conceptualization diagrams again, in this case obtained from supervisor-supervisee dyads, while additionally having coders listen to the supervisor-supervisee discussion as the diagram is being created. The addition of the discussion may be useful given the feedback from our coding team that coding the conceptualization diagrams was difficult due to the inferences that had to be made and the smaller amount of detail present. As discussed previously, there are guidelines which have been proposed for how to structure conversations held around a conceptualization diagram, which could be useful in designing these studies (Liese & Esterline, 2015). While applying the BACQS to conceptualizations in these supervisory settings it may also be useful to examine how acceptable, useful, and accessible the measure was to both the supervisee and supervisor, and whether including the BACQS contributed to trainee learning and growth.

Another very important test of the BACQS would be to test the validity of the scores more rigorously. One way of doing this would be to obtain case conceptualizations from a range of ongoing therapy cases and assess whether the quality of conceptualizations predict later treatment outcomes or other important psychotherapy processes such as the strength of the working alliance. This would expand on the limited research examining case conceptualization impacts on outcomes and the alliance (Bieling & Kuyken, 2003; Nattrass et al., 2015), importantly providing the opportunity to assess this relationship on a broader sample of presenting problems. In such studies it may be important to measure and/or statistically control for other factors that are known to have sizeable impacts on psychotherapy outcomes, such as client characteristics and the so called "common factors" of psychotherapy (Bohart, 2000; Wampold, 2015). This could allow researchers to find the unique contribution of case conceptualization on outcomes and to also explore how case conceptualizations interact with those other factors that are known to contribute to outcomes.

This type of study may also present an opportunity to address the issue of examining the impacts of case conceptualizations only at a single timepoint (and potentially the issue of excluding the client’s perspective as well). It may be beneficial to gather case conceptualizations several times during a course of therapy, and to track and compare the quality as assessed by the BACQS over these repeated samplings. The prediction, guided by the view that conceptualizations evolve and are refined (Johnstone et al., 2011; Kuyken et al., 2008; 2011), would be that BACQS scores should increase over time or with the number of sessions completed. Such a design could potentially reveal that the relationship between conceptualization quality and outcomes varies across stage of therapy or that therapy outcomes are guided more by the slope of the BACQS-rated quality over time.

On the point of including the client’s perspective, it may be valuable to gather the client’s own conceptualization of their difficulties at various points (and apply the BACQS to their conceptualization), to assess the client’s level of agreement with their therapist’s version of the conceptualization, and assess how useful they find their therapist’s conceptualization. These may offer other important methods of assessing the validity (and utility) of the BACQS. Presumably, a client’s own understanding of their difficulties should improve over sessions (reflected in higher BACQS scores over time), their agreement with their therapist’s conceptualization should also correlate positively with BACQS scores (speaking to the proposed benefit that conceptualizations should increase treatment engagement; Johnstone et al., 2011), and higher scoring conceptualizations would also be expected to be as more useful by clients.

During that study, or as a separate study (as perhaps all these questions are too many to address all at once), it may also be useful to measure the psychological mindedness of the therapists and assess whether this construct demonstrates a relationship to the quality of conceptualizations they produce, and perhaps to treatment outcomes in turn. Should such a relationship be found, further research could also examine whether specific training to enhance psychological mindedness has any impact on case conceptualization quality, as this may further highlight the value of this construct in training programs and of viewing it as a competency in clinical psychology.

Another method of assessing validity would be to compare the scores from the BACQS to judgments of conceptualization quality from a small set of registered clinical psychologists representing a sample of "psychotherapy experts", looking for a high rate of agreement across the two evaluations of conceptualization quality. Additionally, scores from the BACQS could be compared to other conceptualization quality measures, this would be another way of demonstrating convergent validity.

Finally, greater effort could be put into testing the divergent validity of the BACQS, to ensure that coders are not simply evaluating persuasiveness, sophistication of the writing, or the length of writing, but instead coding the quality of the conceptualizations.

Conclusions

Through this program of research a new tool for evaluating the quality of case conceptualizations was developed and preliminary evidence for its psychometric performance and validity was obtained. The BACQS was designed with previous research and clinical literature on case conceptualizations informing it, and with the goal of being broadly applicable, accessible, and brief. It was tested across three studies with generally promising results.

The scale obtained adequate levels of inter-rater reliability for most items, and consistently for the overall scale total. The measure demonstrated strong internal consistency across all three studies, and the pattern of inter-item correlations typically were suggestive that the items related strongly to each other, which would be consistent with a measure of overall case conceptualization quality. The BACQS also demonstrated good recode reliability even over periods up to six months. This demonstration of consistency in scores when a measure is applied repeatedly has been lacking in some previous case conceptualization quality measures. Some signs of validity were also obtained from the three studies. For example, that interest and familiarity with mental health, higher academic achievement, being in psychology programs versus other programs of study, and being a professional therapist versus an undergraduate were all predictive of higher quality conceptualizations each lend some credence to the scale measuring skill in conceptualizing mental health difficulties.

Some results point toward the need for further research, for example to further explore the link between empathy and case conceptualization quality, or to better determine if the BACQS is uniquely struggling to achieve adequate inter-rater reliability for certain forms of psychotherapy. Additionally, further research is required to establish more evidence of the scale's validity, and to further extend the areas where the scale has been applied.

References

Altweck, L., Marshall, T. C., Ferenczi, N., & Lefringhausen, K. (2015). Mental health literacy: A cross-cultural approach to knowledge and beliefs about depression, schizophrenia and generalized anxiety disorder. *Frontiers in Psychology, 6*, 1-17. https://doi.org/10.3389/fpsyg.2015.01272

Atzil-Slonim, D., Bar-Kalifa, E., Fisher, H., Lazarus, G., Hasson-Ohayon, I., Lutz, W., Rubel, J., & Rafaeli, E. (2019). Therapists' empathic accuracy toward their clients' emotions. *Journal of Consulting and Clinical Psychology*, *87*(1), 33–45. https://doi.org/10.1037/ccp0000354

Bachrach, H. M., & Leaff, L. A. (1978). "Analyzability": A systematic review of the clinical and quantitative literature. *Journal of the American Psychoanalytic Association*, *26*(4), 881-920. https://doi.org/10.1177/000306517802600409

Barber, J. P., & Crits-Christoph, P. (1993). Advances in measures of psychodynamic formulations. *Journal of Consulting and Clinical Psychology*, *61*(4), 574-585. https://doi.org/10.1037//0022-006X.61.4.574

Bamford, J. M. S., & Davidson, J. W. (2019). Trait Empathy associated with Agreeableness and rhythmic entrainment in a spontaneous movement to music task: Preliminary exploratory investigations. *Musicae Scientiae*, *23*(1), 5-24. https://doi.org/10.1177/1029864917701536

Batson, C. D., Sympson, S. C., Hindman, J. L., Decruz, P., Todd R. M., Weeks J. L., Jennings, G., Burns, C. T. (1996). "I've been there, too": Effect on empathy of prior experience with a need. *Personality and Social Psychology Bulletin*, *22*(5), 474–482. https://doi.org/10.1177/0146167296225005

Beck, J. S. (1995). *Cognitive behavior therapy: Basics and beyond*. Guilford Press.

Beitel, M., Ferrer, E., & Cecero, J. J. (2005). Psychological mindedness and awareness of self and others. *Journal of Clinical Psychology*, *61*(6), 739-750. https://doi.org/10.1002/jclp.20095

Benjamin, L. S. (2003). *Interpersonal reconstructive therapy: Promoting change in non-responders.* New York, NY: Guilford Press.

Bieling, P. J., & Kuyken, W. (2003). Is cognitive case formulation science or science fiction? *Clinical Psychology: Science and Practice*, *10*(1), 52-69. https://doi.org/10.1093/clipsy/10.1.52

Bohart, A. C. (2000). The client is the most important common factor: Clients' self-healing capacities and psychotherapy. *Journal of Psychotherapy Integration*, *10*(2), 127-149. https://doi.org/10.1023/A:1009444132104

Bordage, G., & Lemieux, M. (1991). Semantic structures and diagnostic thinking of experts and novices. *Academic Medicine*, *66*(9 Suppl), S70-S72. https://doi.org/10.1097/00001888-199109000-00045

Bucci, S., French, L., & Berry, K. (2016). Measures assessing the quality of case conceptualization: A systematic review. *Journal of Clinical Psychology*, *72*(6), 517-533. https://doi.org/10.1002/jclp.22280

Butler, G. (1998). Clinical formulation. In A.S. Bellack & M. Hersen (Eds.), *Comprehensive Clinical Psychology*. Oxford: Pergamon.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116-131. https://doi.org/10.1037/0022-3514.42.1.116

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197-253. https://doi.org/10.1037/0033-2909.119.2.197

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306-307. https://doi.org/10.1207/s15327752jpa4803_13

Canadian Psychological Association (2017). *Canadian Code of Ethics for Psychologists 4th edition.* Retrieved from: https://cpa.ca/docs/File/Ethics/CPA_Code_2017_4thEd.pdf

Cape, J., Morris, E., Burd, M., & Buszewicz, M. (2008). Complexity of GPs' explanations about mental health problems: development, reliability, and validity of a measure. *British Journal of General Practice*, *58*(551), 403-410. https://doi.org/10.3399/bjgp08X299281

Capobianco, K. P. (2015). *Evaluating Case Conceptualizations in Psychotherapy Reports: Links to Therapy Outcome and the Alliance.* (Master's Thesis). Retrieved from UW Space. http://hdl.handle.net/10012/9613

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies (1st ed.).*

Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Chadwick, P., Williams, C., & Mackenzie, J. (2003). Impact of case formulation in cognitive behaviour therapy for psychosis. *Behaviour Research and Therapy, 41*(6), 671-680. https://doi.org/10.1016/S0005-7967(02)00033-5

Chapman, B. P., Talbot, N., Tatman, A. W., & Britton, P. C. (2009). Personality traits and the working alliance in psychotherapy trainees: An organizing role for the Five Factor Model? *Journal of Social and Clinical Psychology*, *28*(5), 577-596. https://doi.org/10.1521/jscp.2009.28.5.577

Chung, R.C., & Bemak, F. (2002). The relationship of culture and empathy in cross-cultural counseling. *Journal of Counseling & Development*, *80*(2), 154–159. https://doi.org/10.1002/j.1556-6678.2002.tb00178.x

Clinical Case Studies. (2021.). *Journal description: Clinical case studies*. SAGE Journals. Retrieved from: https://journals.sagepub.com/description/ccs.

Cogburn, M., Raihan, N., Scott, H., Cogburn, H. (2022). A missing step: The value of psychological mindedness training during psychiatry residency. *Open Journal of Psychiatry, 12*¸ 73-77. https://doi.org/10.4236/ojpsych.2022.121007

College of Psychologists of Ontario (2019). *Requirements and Registration Process*. Retrieved from: http://www.cpo.on.ca/Resources.aspx

Collie, R., Ward, T., Vess, J. (2008). Assessment and Case Conceptualization in Sex Offender Treatment. *Journal of Behavior Analysis of Offender and Victim Treatment and Prevention, 1*(1), 65-80. https://doi.org/10.1037/h0100435

Connelly, B. S., Ones, D. S., & Chernyshenko, O. S. (2014). Introducing the special section on openness to experience: Review of openness taxonomies, measurement, and nomological net. *Journal of Personality Assessment*, *96*(1), 1-16. https://doi.org/10.1080/00223891.2013.830620

Constantino, M. J., Ametrano, R. M., & Greenberg, R. P. (2012). Clinician interventions and participant characteristics that foster adaptive patient expectations for psychotherapy and psychotherapeutic change. *Psychotherapy*, *49*, 557-569. https://doi.org/10.1037/a0029440

Coppock, T. E., Owen, J. J., Zagarskas, E., & Schmidt, M. (2010). The relationship between therapist and client hope with therapy outcomes. *Psychotherapy Research*, *20*(6), 619-626. https://doi.org/10.1080/10503307.2010.497508

Crits-Christoph, P., Luborsky, L., Dahl, L., Popp, C., Mellon, J., & Mark, D. (1988). Clinicians can agree in assessing relationship patterns in psychotherapy: The Core Conflictual Relationship Theme method. *Archives of General Psychiatry*, *45*(11), 1001-1004. https://doi.org/10.1001/archpsyc.1988.01800350035005

Crowe, M., Carlyle, D., & Farmar, R. (2008). Clinical formulation for mental health nursing practice. *Journal of Psychiatric and Mental Health Nursing*, *15*(10), 800-807. https://doi.org/10.1111/j.1365-2850.2008.01307.x

Cummings, A. L., Hallberg, E. T., Martin, J., Slemon, A., & Hiebert, B. (1990). Implications of counselor conceptualizations for counselor education. *Counselor Education and Supervision, 30*(2), 120–134. https://doi.org/10.1002/j.1556-6978.1990.tb01189.x

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, *10*, 85-104.

Daw, B., & Joseph, S. (2010). Psychological mindedness and therapist attributes. *Counselling and Psychotherapy Research*, *10*(3), 233-236. https://doi.org/10.1080/14733140903226982

Del Re, A. C., Flückiger, C., Horvath, A. O., Symonds, D., & Wampold, B. E. (2012). Therapist effects in the therapeutic alliance–outcome relationship: A restricted-maximum likelihood meta-analysis. *Clinical Psychology Review*, *32*(7), 642-649. https://doi.org/10.1016/j.cpr.2012.07.002

Del Re, A. C., Flückiger, C., Horvath, A. O., & Wampold, B. E. (2021). Examining therapist effects in the alliance–outcome relationship: A multilevel meta-analysis. *Journal of Consulting and Clinical Psychology*, *89*(5), 371–378. https://doi.org/10.1037/ccp0000637

De Simone, C. (2007). Applications of concept mapping. *College Teaching*, *55*(1), 33-36. https://doi.org/10.3200/CTCH.55.1.33-36

Dudley, R., Park, I., James, I., & Dodgson, G. (2010). Rate of agreement between clinicians on the content of a cognitive formulation of delusional beliefs: The effect of qualifications and experience. *Behavioural and Cognitive Psychotherapy*, *38*(2), 185-200. https://doi.org/10.1017/S1352465809990658

Easden, M. H., & Kazantzis, N. (2018). Case conceptualization research in cognitive behavior therapy: A state of the science review. *Journal of Clinical Psychology*, *74*(3), 356-384. https://doi.org/10.1002/jclp.22516

Eells, T. D. (Ed.). (2011). *Handbook of psychotherapy case formulation*. Guilford Press.

Eells, T. D., Kendjelic, E. M., & Lucas, C. P. (1998). What's in a case formulation?: development and use of a content coding manual. *The Journal of Psychotherapy Practice and Research*, *7*(2), 144-153.

Eells, T. D., & Lombart, K. G. (2003). Case formulation and treatment concepts among novice, experienced, and expert cognitive-behavioral and psychodynamic therapists. *Psychotherapy Research*, *13*(2), 187-204.

Eells, T. D., Lombart, K. G., Kendjelic, E. M., Turner, L. C., & Lucas, C. P. (2005). The quality of psychotherapy case formulations: a comparison of expert, experienced, and novice cognitive-behavioral and psychodynamic therapists. *Journal of Consulting and Clinical Psychology*, *73*(4), 579-589. https://doi.org/10.1037/0022-006X.73.4.579

Eells, T. D., Lombart, K. G., Salsman, N., Kendjelic, E. M., Schneiderman, C. T., & Lucas, C. P. (2011). Expert reasoning in psychotherapy case formulation. *Psychotherapy Research*, *21*(4), 385-399.

Elliot, R., Bohart, A. C., Watson, J. C., Greenberg, L. S., & Norcross, J. C. (2011). *Psychotherapy relationships that work*: *Therapist contributions and responsiveness to patient needs*. New York: Oxford University Press

Elliott, R., Bohart, A. C., Watson, J. C., & Murphy, D. (2018). Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, *55*(4), 399-410. https://doi.org/10.1037/pst0000175

Emmelkamp, P. M., Bouman, T. K., & Blaauw, E. (1994). Individualized versus standardized therapy: A comparative evaluation with obsessive-compulsive patients. *Clinical Psychology & Psychotherapy*, *1*(2), 95-100. https://doi.org/10.1002/cpp.5640010206

Evans, G. & Parry, J. (1996). The impact of reformulation in cognitive-analytic therapy with difficult-to-help clients. *Clinical Psychology and Psychotherapy*, *3*(2), 109–117. https://doi.org/10.1002/(SICI)1099-0879(199606)3:2<109::AID-CPP65>3.0.CO;2-U

Falender, C. A., & Shafranske, E. P. (2014). Clinical supervision: The state of the art. *Journal of Clinical Psychology*, *70*(11), 1030-1041. https://doi.org/10.1002/jclp.22124

Falvey, J. E., Bray, T. E., & Hebert, D. J. (2005). Case conceptualization and treatment planning: Investigation of problem-solving and clinical judgment. *Journal of Mental Health Counseling*, *27*(4), 348-372.

Farber, B. A. (1985). The genesis, development, and implications of psychological-mindedness in psychotherapists. *Psychotherapy: Theory, Research, Practice, Training*, *22*(2), 170-177. https://doi.org/10.1037/h0085490

Farber, B. A., Manevich, I., Metzger, J., & Saypol, E. (2005). Choosing psychotherapy as a career: Why did we cross that road? *Journal of Clinical Psychology, 61*(8), 1009-1031. https://doi.org/10.1002/jclp.20174

Featherston, R., Downie, L.E., Vogel, A.P., Galvin, K.L. (2020). Decision making biases in the allied health professions: A systematic scoping review. *Plos One*, *15*(10). doi: 10.1371/journal.pone.0240716

Feldman, D. B., & Crandall, C. S. (2007). Dimensions of mental illness stigma: What about mental illness causes social rejection? *Journal of Social and Clinical Psychology*, *26*(2), 137-154. https://doi.org/10.1521/jscp.2007.26.2.137

Fernando, I., Cohen, M., & Henskens, F. (2013). A systematic approach to clinical reasoning in psychiatry. *Australasian Psychiatry*, *21*(3), 224-230. https://doi.org/10.1177/1039856213486209

Fleischhauer, M., Enge, S., Brocke, B., Ullrich, J., Strobel, A., & Strobel, A. (2010). Same or different? Clarifying the relationship of need for cognition to personality and intelligence. *Personality and Social Psychology Bulletin*, *36*(1), 82-96. https://doi.org/10.1177/0146167209351886

Flitcroft, A., James, I. A., Freeston, M., & Wood-Mitchell, A. (2007). Determining what is important in a good formulation. *Behavioural and Cognitive Psychotherapy*, *35*(3), 325-333. https://doi.org/10.1017/S135246580600350X

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, *55*(4), 316–340. https://doi-org.proxy.lib.uwaterloo.ca/10.1037/pst0000172

Fothergill, C. D., & Kuyken, W. (2002). The quality of cognitive case formulation rating scale. *Unpublished manuscript*.

Fox, A. B., Earnshaw, V. A., Taverna, E. C., & Vogt, D. (2018). Conceptualizing and measuring mental illness stigma: The mental illness stigma framework and critical review of measures. *Stigma and Health*, *3*(4), 348-376. https://doi.org/10.1037/sah0000104

Frank, G., (1984). The Boulder Model: History, rationale, and critique. *Professional Psychology: Research and Practice*, *15*(3), 417–435. https://doi.org/10.1037/0735-7028.15.3.417

Ghaderi, A. (2006). Does individualization matter? A randomized trial of standardized (focused) versus individualized (broad) cognitive behavior therapy for bulimia nervosa. *Behaviour Research and Therapy, 44*(2), 273-288. https://doi.org/10.1016/j.brat.2005.02.004

Glidewell, J. C., & Livert, D. E. (1992). Confidence in the practice of clinical psychology. *Professional Psychology: Research and Practice, 23*(5), 362-368. https://doi.org/10.1037/0735-7028.23.5.362

Goldfried, M.R. (1995). *From cognitive-behaviour therapy to psychotherapy integration*. New York: Springer-Verlag.

Gong, Y., Ericsson, K. A., & Moxley, J. H. (2015). Recall of briefly presented chess positions and its relation to chess skill. *PLoS One*, *10*(3). https://doi.org/10.1371/journal.pone.0118756

Goodyear, R. K. (1997). Psychological expertise and the role of individual differences: An exploration of issues. *Educational Psychology Review, 9*(3), 251-265.

Gower, P. (2011). *Therapist competence, case conceptualization and therapy outcome in cognitive behavioural therapy* (Doctoral dissertation, University of Exeter, Exeter, United Kingdom). Retrieved from https://ore.exeter.ac.uk/repository/handle/10036/3275

Haarhoff, B. A., Flett, R. A., & Gibson, K. L. (2011). Evaluating the content and quality of cognitive-behavioural therapy case conceptualisations. *New Zealand Journal of Psychology*, *40*(3), 104-114.

Habashi, M. M., Graziano, W. G., & Hoover, A. E. (2016). Searching for the prosocial personality: A Big Five approach to linking personality and prosocial behavior. *Personality and Social Psychology Bulletin*, *42*(9), 1177-1192. https://doi.org/10.1177/0146167216652859

Hartley, S., Jovanoska, J., Roberts, S., Burden, N., & Berry, K. (2016). Case formulation in clinical practice: Associations with psychological mindedness, attachment and burnout in staff working with people experiencing psychosis. *Psychology and Psychotherapy: Theory, Research and Practice*, *89*(2), 133-147. https://doi.org/10.1111/papt.12074

Harton, H. C., & Lyons, P. C. (2003). Gender, empathy, and the choice of the psychology major. *Teaching of Psychology*, *30*(1), 19-24. https://doi.org/10.1207/S15328023TOP3001_03

Henggeler, S.W., Schoenwald, S.K., Rowland, M.D., & Cunningham, P.B. (2002). *Serious emotional disturbance in children and adolescents: Multisystemic therapy.* New York: Guildford Press.

Hessen, E., Hokkanen, L., Ponsford, J., van Zandvoort, M., Watts, A., Evans, J., & Haaland, K. Y. (2018). Core competencies in clinical neuropsychology training across the world. *The Clinical Neuropsychologist*, *32*(4), 642-656. https://doi.org/10.1080/13854046.2017.1413210

Hill, C. E., & O'Brien, K. M. (2004). *Helping skills: Facilitating exploration, insight, and action* (pp. 25-37). Washington, DC: American Psychological Association.

Hing, N., Russell, A. M., & Gainsbury, S. M. (2016). Unpacking the public stigma of problem gambling: The process of stigma creation and predictors of social distancing. *Journal of Behavioral Addictions, 5*(3), 448-456. https://doi.org/10.1556/2006.5.2016.057

Irving, L. M., Snyder, C. R., Cheavens, J., Gravel, L., Hanke, J., Hilberg, P., & Nelson, N. (2004). The relationships between hope and outcomes at the pretreatment, beginning, and later phases of psychotherapy. *Journal of Psychotherapy Integration*, *14*, 419-443. https://doi.org/10.1037/1053-0479.14.4.419

Jaafarpour, M., Aazami, S., & Mozafari, M. (2016). Does concept mapping enhance learning outcome of nursing students? *Nurse Education Today, 36*, 129-132. https://doi.org/10.1016/j.nedt.2015.08.029

Jacobson, N. S., Schmaling, K. B., Holtzworth-Munroe, A., Katt, J. L., Wood, L. F., & Follette, V. M. (1989). Research-structured vs clinically flexible versions of social learning-based marital therapy. *Behaviour Research and Therapy*, 27(2), 173-180. https://doi.org/10.1016/0005-7967(89)90076-4

Jennings, L., & Skovholt, T. M. (1999). The cognitive, emotional, and relational characteristics of master therapists. *Journal of Counseling Psychology*, *46*(1), 3-11. https://doi.org/10.1037/0022-0167.46.1.3

Johnstone, L., Whomsley, S., Cole, S., & Oliver, N. (2011). *Good practice guidelines on the use of psychological formulation.* Retrieved from: http://www.sisdca.it/public/pdf/DCP-Guidelines-for-Formulation-2011.pdf.

Jorm, A. F., Korten, A. E., Jacomb, P. A., Rodgers, B., Pollitt, P., Christensen, H., & Henderson, S. (1997). Helpfulness of interventions for mental disorders: Beliefs of health professionals compared with the general public. *The British Journal of Psychiatry*, *171*(3), 233-237.

Kahneman, D. (2013). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kanter, J. W., Weeks, C. E., Bonow, J. T., Landes, S. J., Callaghan, G. M., & Follette, W. C. (2009). Assessment and case conceptualization. *In A Guide to Functional Analytic Psychotherapy* (pp. 1-23). Springer, Boston, MA.

Kendjelic, E. M., & Eells, T. D. (2007). Generic psychotherapy case formulation training improves formulation quality. *Psychotherapy: Theory, Research, Practice, Training*, *44*(1), 66-77. https://doi.org/10.1037/0033-3204.44.1.66

Kim, H., & Han, S. (2018). Does personal distress enhance empathic interaction or block it? *Personality and Individual Differences*, *124*, 77-83. https://doi.org/10.1016/j.paid.2017.12.005

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Naval Technical Training Command Millington TN Research Branch.

Koo, T.K., & Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kotov, Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., … Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A Dimensional Alternative to Traditional Nosologies. *Journal of Abnormal Psychology*, 126(*4*), 454–477. https://doi.org/10.1037/abn0000258

Kuyken, W., Beshai, S., Dudley, R., Abel, A., Görg, N., Gower, P., McManus, F., & Padesky, C. A. (2016). Assessing competence in collaborative case conceptualization: Development and preliminary psychometric properties of the Collaborative Case Conceptualization Rating Scale (CCC-RS). *Behavioural and Cognitive Psychotherapy, 44*(2), 179-192. https://doi.org/10.1017/S1352465814000691

Kuyken, W., Fothergill, C. D., Musa, M., & Chadwick, P. (2005). The reliability and quality of cognitive case formulation. *Behaviour Research and Therapy*, *43*(9), 1187-1201. https://doi.org/10.1016/j.brat.2004.08.007

Kuyken, W., Padesky, C. A., & Dudley, R. (2008). The science and practice of case conceptualization. *Behavioural and Cognitive Psychotherapy, 36*(6), 757-768. https://doi.org/10.1017/S1352465808004815

Kuyken, W., Padesky, C. A., & Dudley, R. (2011). *Collaborative case conceptualization: Working effectively with clients in cognitive-behavioral therapy*. Guilford Press.

Liese, B. S., & Esterline, K. M. (2015). Concept mapping: A supervision strategy for introducing case conceptualization skills to novice therapists. *Psychotherapy, 52*(2), 190-194. https://doi.org/10.1037/a0038618

Llinás, J. G., Macías, F. S., & Márquez, L. M. T. (2020). The use of concept maps as an assessment tool in physics classes: Can one use concept maps for quantitative evaluations? *Research in Science Education*, *50*(5), 1789-1804. https://doi.org/10.1007/s11165-018-9753-4

Luborsky, L. (1977). Measuring a pervasive psychic structure in psychotherapy: The core conflictual relationship theme. In *Communicative structures and psychic structures* (pp. 367-395). Springer, Boston, MA.

Mackrill, T., & Iwakabe, S. (2013). Making a case for case studies in psychotherapy training: A small step towards establishing an empirical basis for psychotherapy training. *Counselling Psychology Quarterly*, *26*(3-4), 250-266. https://doi.org/10.1080/09515070.2013.832148

McCallum, M., & Piper, W. E. (1990). The Psychological Mindedness Assessment Procedure. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*(4), 412-418. https://doi.org/10.1037/1040-3590.2.4.412

McCallum, M., & Piper, W. E. (1996). Psychological mindedness. *Psychiatry*, *59*(1), 48-64. https://doi.org/10.1521/00332747.1996.11024750

McMurran, M., Logan, C., & Hart, S. (2012). Case formulation quality checklist. *Unpublished checklist*.

McNeil, D. W., & Hayes, S. E. (2015). Psychotherapy Analog Studies. *The Encyclopedia of Clinical Psychology*, 1-3. https://doi.org/10.1002/9781118625392.wbecp406

Meichenbaum, D. (2014). *The Role of a Case Conceptualization Model and Core Tasks of Intervention.* Retrieved from: https://www.melissainstitute.org/documents/Conf18-Meichenbaum-CaseConcept.pdf

Meier, S. T. (2003). *Bridging case conceptualization, assessment, and intervention*. Thousand Oaks, CA: Sage.

Miles, R., Rabin, L., Krishnan, A., Grandoit, E., & Kloskowski, K. (2020). Mental health literacy in a diverse sample of undergraduate students: demographic, psychological, and academic correlates. *BMC Public Health*, *20*(1), 1699. https://doi.org/10.1186/s12889-020-09696-0

Milovanov, A. (2017). *Common Therapeutic Factors in Psychotherapy and Complementary and Alternative Medicine Treatments*. (Master's Thesis). Retrieved from UW Space. http://hdl.handle.net/10012/12315

Murre, J.M., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *Plos One 10*(7). https://doi.org/10.1371/journal.pone.0120644

Nattrass, A., Kellett, S., Hardy, G. E., & Ricketts, T. (2015). The content, quality and impact of cognitive behavioural case formulation during treatment of obsessive compulsive disorder. *Behavioural and Cognitive Psychotherapy*, *43*(5), 590-601. https://doi.org/10.1017/S135246581400006X

Needleman, L. D. (1999). *Cognitive case conceptualization: A guidebook for practitioners*. Routledge.

Negd, M., Mallan, K. M., & Lipp, O. V. (2011). The role of anxiety and perspective-taking strategy on affective empathic responses. *Behaviour Research and Therapy*, *49*(12), 852-857. https://doi.org/10.1016/j.brat.2011.09.008

Nezu, C. M., Nezu, A. M., & Colosimo, M. M. (2015). Case formulation and the therapeutic alliance in contemporary Problem-Solving Therapy (PST). *Journal of Clinical Psychology*, *71*(5), 428-438. https://doi.org/10.1002/jclp.22179

Nicoll, G. (2001). A three-tier system for assessing concept map links: a methodological study. *International Journal of Science Education*, *23*(8), 863-875. https://doi.org/10.1080/09500690010025003

Norcross, J. C., & Karpiak, C. P. (2012). Clinical psychologists in the 2010s: 50 years of the APA division of clinical psychology. *Clinical Psychology: Science and Practice*, *19*(1), 1-12. https://doi.org/10.1111/j.1468-2850.2012.01269.x

Novak, J. D. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching*, *27*(10), 937-949. https://doi.org/10.1002/tea.3660271003

O'Brien, T. J. (2001). The development and impact of psychological mindedness in clinical psychology doctoral students (Doctoral dissertation, The Chicago School of Professional Psychology).

Ortega-Tudela, J. M., Lechuga, M. T., & Gómez-Ariza, C. J. (2019). A specific benefit of retrieval-based concept mapping to enhance learning from texts. *Instructional Science*, *47*(2), 239-255. https://doi.org/10.1007/s11251-018-9476-y

Padesky, C. A, Kuyken, W., & Dudley, R. (2011). *Collaborative Case Conceptualisation Rating Scale and Coding Manual*. Unpublished manual.

Page, A. C., Stritzke, W. G., & Mclean, N. J. (2008). Toward science-informed supervision of clinical case formulation: A training model and supervision method. *Australian Psychologist, 43*(2), 88-95. https://doi.org/10.1080/00050060801994156

Pascual-Leone, A., Andreescu, C. A., & Yeryomenko, N. (2014). Training novice psychotherapists: Comparing undergraduate and graduate students' outcomes. *Counselling and Psychotherapy Research*, *15*(2), 137-146. https://doi.org/10.1002/capr.12007

Persons, J. B. (1989). *Cognitive therapy in practice: A case formulation approach* (pp. 109-118). New York: Norton

Persons, J. B. (2006). Case formulation–driven psychotherapy. *Clinical Psychology: Science and Practice*, *13*(2), 167-170. https://doi.org/10.1111/j.1468-2850.2006.00019.x

Persons, J. B., & Bertagnolli, A. (1999). Inter-rater reliability of cognitive-behavioral case formulations of depression: A replication. *Cognitive Therapy and Research*, *23*(3), 271-283. https://doi.org/10.1023/A:1018791531158

Persons, J. B., Roberts, N. A., Zalecki, C. A., & Brechwald, W. A. (2006). Naturalistic outcome of case formulation-driven cognitive-behavior therapy for anxious depressed outpatients. *Behaviour Research and Therapy*, *44*(7), 1041-1051. https://doi.org/10.1016/j.brat.2005.08.005

Persons, J. B., Mooney, K. A., & Padesky, C. A. (1995). Interrater reliability of cognitive-behavioral case formulations. *Cognitive Therapy and Research*, *19*(1), 21-34. https://doi.org/10.1007/BF02229674

Quilty, L.C., Taylor, G.J., McBride, C., Bagby, R.M. (2017). Relationships among alexithymia, therapeutic alliance, and psychotherapy outcome in major depressive disorder. *Psychiatry Research*, *254*, 75-79. https://doi.org/10.1016/j.psychres.2017.04.047.

Restifo, S. (2011). Beyond diagnosis-and therapy-centred therapy: The case for incorporating contextual factors in psychotherapy treatment planning. *Australian and New Zealand Journal of Psychiatry*, *19*(4), 309-312. https://doi.org/10.3109/10398562.2011.579125

Ridley, C. R., Jeffrey, C. E., & Roberson III, R. B. (2017). Case mis-conceptualization in psychological treatment: An enduring clinical problem. *Journal of Clinical Psychology*, *73*(4), 359-375. https://doi.org/10.1002/jclp.22354

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118-137. https://doi.org/10.1016/j.intell.2015.09.002

Saito, T., Takeda, S., Yamagishi, Y., Kubo, R., Kitamura, T. (2017). Psychotherapy training on psychological mindedness in a Japanese nurse population: Effects and personality correlates. *Healthcare*, 5(*3),* 43-52. https://doi.org/10.3390/healthcare5030043

Samstag, L. W., Muran, J. C., & Safran, J. D. (2004). Defining and identifying alliance ruptures in D.P. Charman (Ed.), *Core processes in brief psychodynamic psychotherapy: Advancing effective practice*, 187-214. Routeledge.

Scheel, M. J., Davis, C. K., & Henderson, J. D. (2013). Therapist use of client strengths: A qualitative study of positive processes. *The Counseling Psychologist*, *41*(3), 392-427. https://doi.org/10.1177/1469787420950589

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763-1768.

Schulte, D., Künzel, R., Pepping, G., & Schulte-Bahrenberg, T. (1992). Tailor-made versus standardized therapy of phobic patients. *Advances in Behaviour Research and Therapy, 14*(2), 67-92. https://doi.org/10.1016/0146-6402(92)90001-5

Shestowsky, D., Wegener, D. T., & Fabrigar, L. R. (1998). Need for cognition and interpersonal influence: Individual differences in impact on dyadic decisions. *Journal of Personality and Social Psychology*, *74*(5), 1317-1328. https://doi.org/10.1037/0022-3514.74.5.1317

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420

Shrout, P. E., & Lane, S. P. (2012). Psychometrics. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 302–320). The Guilford Press.

Sim, K., Gwee, K. P., & Bateman, A. (2005). Case formulation in psychotherapy: Revitalizing its usefulness as a clinical tool. *Academic Psychiatry*, *29*(3), 289-292. https://doi.org/10.1176/appi.ap.29.3.289

Spreng, R. N., McKinnon, M. C., Mar, R. A., & Levine, B. (2009). The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment*, *91*(1), 62-71. https://doi.org/10.1080/00223890802484381

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality, 68*, 69-81. https://doi.org/10.1016/j.jrp.2017.02.004

Streiner, D.L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99-103.

Taub, G. E., & McGrew, K. S. (2014). The Woodcock–Johnson Tests of Cognitive Abilities III's cognitive performance model: Empirical support for intermediate factors within CHC theory. *Journal of Psychoeducational Assessment*, *32*(3), 187-201. https://doi.org/10.1177/0734282913504808

Toews, J. A. (1993). Case formulation in psychiatry: Revitalizing an ailing art. *Canadian Journal of Psychiatry*, *38*(5), 344-344. https://doi.org/10.1177/070674379303800511

Truax, C. B., Wargo, D. G., Frank, J. D., Imber, S. D., Battle, C. C., Hoehn-Saric, R., Nash, E.H., & Stone, A. R. (1966). Therapist empathy, genuineness, and warmth and patient therapeutic outcome. *Journal of Consulting Psychology*, *30*(5), 395-401. https://doi.org/10.1037/h0023827

Trudeau, K. J., & Reich, R. (1995). Correlates of psychological mindedness. *Personality and Individual Differences*, *19*(5), 699-704. https://doi.org/10.1016/0191-8869(95)00110-R

Van Loo, H. M., Romeijn, J. W., de Jonge, P., & Schoevers, R. A. (2013). Psychiatric comorbidity and causal disease models. *Preventive Medicine*, *57*(6), 748-752. https://doi.org/10.1016/j.ypmed.2012.10.018

Vedel, A. (2016). Big Five personality group differences across academic majors: A systematic review. *Personality and Individual Differences*, *92*, 1-10. https://doi.org/10.1016/j.paid.2015.12.011

Verplanken, B., Hazenberg, P. T., & Palenewen, G. R. (1992). Need for cognition and external information search effort. *Journal of Research in Personality*, *26*(2), 128-136. https://doi.org/10.1016/0092-6566(92)90049-A

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry*, *14*(3), 270-277. https://doi.org/10.1002/wps.20238

Welfare, L. E., Farmer, L. B., & Lile, J. J. (2013). Empirical evidence for the importance of conceptualizing client strengths. *The Journal of Humanistic Counseling*, *52*(2), 146-163. https://doi.org/10.1002/j.2161-1939.2013.00039.x

Westen, D., Lifton, N., Boekamp, J., Huebner, D., & Silverman, M. (1991). Assessing complexity of representations of people and understanding of social causality: A comparison of natural science and clinical psychology graduate students. *Journal of Social and Clinical Psychology*, *10*(4), 448-458. https://doi.org/10.1521/jscp.1991.10.4.448

Whelehan, D. F., Conlon, K. C., & Ridgway, P. F. (2020). Medicine and heuristics: cognitive biases and medical decision-making. *Irish Journal of Medical Science (1971-)*, *189*(4), 1477-1484. https://doi.org/10.1007/s11845-020-02235-1

Withall, A.A., & Sagi, E. (2021). The impact of readability on trust in information. *The Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 2370-2376.

Wilson, G.T. (1996). Manual-based treatments: The clinical application of research findings. *Behavior Research and Therapy*, *34,* 295–314. https://doi.org/10.1016/0005-7967(95)00084-4

Wong, F.K., Cheng, C., Zhuang, X.Y., Ng, T.K., Pan, S., He, X., Poon, A. (2017). Comparing the mental health literacy of Chinese people in Australia, China, Hong Kong and Taiwan: Implications for mental health promotion. *Psychiatry Research, 256*, 258-266.

Zimmerman, M., McGlinchey, J.B., Chelminski, I., & Young, D. (2008). Diagnostic co-morbidity in 2300 psychiatric outpatients presenting for treatment evaluated with a semi-structured diagnostic interview. *Psychological Medicine*, *38*(2), 199-210. https://doi.org/10.1017/S0033291707001717

Zivor, M., Salkovskis, P. M., Oldfield, V. B., & Kushnir, J. (2013). Formulation in cognitive behavior therapy for obsessive–compulsive disorder: Aligning therapists, perceptions and practice. *Clinical Psychology: Science and Practice*, *20*(2), 143-151. https://doi.org/10.1111/cpsp.12030

Appendices

Appendix 1 | Broadly Applicable Conceptualization Quality Scale

Please read through the following coding instructions fully before you begin scoring the conceptualizations provided.

Working one conceptualization at a time, examine a conceptualization once fully and then consult the following coding criteria to score the conceptualizations on the dimensions below. You may re-examine the conceptualization after the first time, and are also free to adjust the scores you give. However, there is an exception to this for the overall score. Provide this score after examining the conceptualization the first time, and do not change your initial score.

1) General Quality Impression

Based on your initial impression, score the conceptualization for your impression of its quality, considering they are meant to be used to understand a client better and guide a course of psychotherapy.

(1) - Very Poor - There are fundamental and serious problems with this case conceptualization. It is weak in several ways and should not be used to guide a course of psychotherapy.

(2) - Poor - This case conceptualization is approaching adequate, but is still not appropriate to guide a course of psychotherapy at this point. You can take something useful from this conceptualization, but it isn't enough.

(3) - Adequate - With this case conceptualization you believe a psychotherapist should be able to help a client in psychotherapy. It is developed enough to serve its useful functions without being above average in quality.

(4) - Strong - This case conceptualization is quite obviously going to be useful in psychotherapy, and goes above and beyond being merely adequate in some way(s). There are no serious flaw or weaknesses.

(5) - Outstanding - In almost every aspect this case conceptualization exceeds. It appears very skillfully developed and has very high potential in guiding psychotherapy and enhancing important aspects of the psychotherapeutic process.

2) Psychological Mindedness and Depth

Psychological mindedness is the propensity to consider psychological constructs when trying to understand the self and others. It is shown when emotional nuance, the impact of the past on the present, and the relationships between thoughts, feelings, and actions are explored. It is also reflected in an awareness of the messy, unconscious, conflicted, and sometimes self-defeating human nature. A conceptualization high in psychological mindedness will not be limited to concrete, physical, situational, or reductionist psychological explanations, instead working at a more abstract/complex/deeper level.

(1) - Very low - this conceptualization lacks any appreciation for the internal, psychological world of the client. The conceptualization references only situational factors, biology, and concrete events without linking these at all to any emotions, thoughts, or other intra-psychic phenomena. At its worst, this level of conceptualization has reduced a complex and nuanced problem to an overly simplified explanation and has failed to explore why, if the problem is so simple, it has not already been resolved.

(2) - Low - this conceptualization considers some basic internal psychological processes of the client, for example noting an emotion or a particular thought, but has remained overall focused on concrete and external factors. When internal processes are noted they are not considered fully, and are often seen as effects of, rather than causes of, other phenomena. There may still be oversimplifying of the client's difficulties, and the solution to these is seen to be a matter of concrete and easy changes.

(3) - Moderate - this conceptualization uses internal psychological processes and phenomena to help explain other events and outline the client's difficulties. The internal psychological processes are given some centrality in the conceptualization, and often connect to other concrete and external phenomena. However, the level of complexity of the psychological processes remains generally low, and the conceptualization does not consider inner-conflict or complex cycles of thoughts-behaviours-feelings, and explores emotions without much recognition of varying intensity or mixed feelings.

(4) - High - this conceptualization considers internal psychological processes and phenomena of the client with depth and openness. The client's perceptions, personality, thoughts, and emotions are used to explain their behaviours and struggles. When more concrete concepts are incorporated into the conceptualization they are explored in terms of their interaction with psychological processes. Some nuance is present in the consideration of emotions, cycles of thoughts-behaviours-feelings, conflicting desires, and unconscious processes. The conceptualization considers past events and experiences of the client, including childhood experiences, and how these have shaped their psychology today.

(5) - Very high - This conceptualization explores the client's psychology in a remarkably rich, open, and thoughtful manner. Unique client perceptions, experiences, personality traits, behaviours, and history are considered in relation to unconscious drives, held beliefs, fears, and inner-conflicts. The intensity and variation of the client's emotional experience is considered. Complex patterns of thoughts, feelings and actions are explored. Links are made to past events with care and consideration, and add depth to the understanding of the client. The fundamental worldviews, core beliefs, attachment styles, interpersonal patterns, and self-concepts of the client are used to enhance the understanding of their larger life story, and also the nature of the more immediate difficulties being addressed.

3) Integration

A conceptualization high in integration links the ideas in the conceptualization (such as the history, characteristics, thoughts, feelings, and situations of the client) together in meaningful ways. In this way a well-integrated conceptualization should not read as a list of ideas, but has a flow, narrative, structure, and cohesion. A well-integrated conceptualization suggests causal links or relationships. The conceptualization will have a logic to it, such that things are explained, not merely described.

(1) - Unintegrated - The conceptualization reads as a list of separate, unexplained, and unexplored facts. There are nearly no examples where the creator has tried to link two separate ideas in any meaningful way. Although the creator may have used the client's own explanations in their conceptualization ("they lost their job so they moved to a new city"), this link does not appear to represent any new connection and is not a meaningful integration.

(2) - Weakly Integrated - The conceptualization may have a sparse or limited linking of some facts together. While a majority of the information still does not appear to be linked to other parts meaningfully, at least a few instances occur where the creator has demonstrated a deeper understanding of how two ideas are linked.

(3) - Moderately integrated - There is a moderate amount of tying of ideas and facts together. The creator has begun to form connections between several ideas that were not likely linked by the client themselves. However, there is not yet a clear overall structure or larger picture in the conceptualization.

(4) - Highly Integrated - The conceptualization only rarely contains a piece of information without at some point making use of it to explain another idea, or without linking it with other concepts. Clusters of ideas are formed and a larger picture for the client's case may be apparent. A large portion of the ideas being linked appear to be hypotheses of the creator, and not merely representations of information the client provided.

(5) - Extremely Integrated - There is clearly a larger picture in this conceptualization and nearly all ideas are linked in a logical, useful, and understandable way. There has been a sorting and clustering of ideas which helps to form patterns and enhances comprehension. The creator of the conceptualization has made many important links beyond those the client themselves provided, and has made some non-obvious and useful hypotheses relating disparate aspects of the client's case.

4) Differentiation/Breadth

A conceptualization with high differentiation will likely contain a higher number of "types" or "categories" of information, and also more examples in categories of information present. A differentiated conceptualization does not seem to focus on just one main idea, or a few main ideas, but instead appears to be an attempt at a thorough account of the client. A differentiated conceptualization may look at the problem from several perspectives or across places, times, and situations. It could also reference ideas from different theories for human behaviour.

(1) - Extremely Sparse - There is a single main perspective, group of ideas, and a general sparseness of information. The conceptualization is overly minimalist, and it appears the creator believed only a few ideas to be worth considering.

(2) - Sparse - There are several categories or types of information used, but the conceptualization still appears pretty sparse and incomprehensive , you feel there must still be a lot more to know about the case

(3) - Moderate differentiation - There's a moderate amount of information from several categories or perspectives. The creator may provide a couple of examples for some of the ideas they consider. You feel with the amount of information present you likely have a reasonable starting description of the client. However, you still feel as though there are some areas where more information should be included.

(4) - High differentiation - There is a rich amount of information taken from different times, situations, and categories, the creator is clearly considering a great deal of the client information available. Several ideas are presented with multiple examples. You feel you have a solid and pretty thorough description of the client available. You feel there is at least one piece of information available for nearly all important categories that could be considered.

(5) - Very high differentiation - The conceptualization has considered the client and their difficulties in high detail: looking at many times and situations, including information from many categories, and thinking from several perspectives. The description, facts, and ideas available appear very thorough and you are really able to understand and perceive many aspects of this client's life, situation, and characteristics. You find yourself very satisfied with the amount and types of information provided, and would have few suggestions for important things that should also be added into the conceptualization.

5) Core mechanism identified

A conceptualization with a core mechanism identified will have theorized that at least one major psychological mechanism explains a large portion of the client's difficulties. Adequate detail will explain the psychological mechanism and the processes that are sustaining it. The mechanism should also not simply be a statement of a symptom disguised as a cause. It should be clear that if this core problem was addressed (overcome, eliminated, adapted to, remediated, etc.) the client would experience meaningful improvement in their mental health difficulties. The types of mechanisms identified should appear credible/justifiable and should be addressable in psychotherapeutic work. A conceptualization that lacks a core problem will not appear particularly useful for guiding psychotherapy.

(1) - No core mechanism identified - The conceptualization has not captured any meaningful reason for why the client is facing their challenges. If the creator of the conceptualization has included something important, they do not appear to have realized it, and it is not used to explain any significant part of the client's difficulties.

(2) - A psychological mechanism touched on - There are a few places where a useful psychological mechanism is mentioned, or briefly described, but it is not well used in explaining the client's difficulties. The conceptualization does not highlight why this mechanism is an important driver of the client's difficulties above and beyond the other information in the conceptualization. Nothing stands out as central, major, or critical in explaining why the client is facing the challenges they are facing.

(3) - Not quite strong enough mechanism - The conceptualization does highlight a psychological mechanism (or a few main mechanisms) which are roughly linked to the client's difficulties. However, it seems a more central mechanism, a more clearly articulated mechanism, or including other mechanisms would be useful as a large portion of the client's difficulties remain unexplained. If this mechanism was remediated in psychotherapy there would be some benefit, but you do not feel the client's main difficulties would be adequately addressed.

(4) The conceptualization contains a main psychological mechanism (or a few main mechanisms) which explains the majority of the client's difficulties in an accessible and clear manner. It is easy to follow the reasoning for how these mechanisms are creating/sustaining the problems. You feel confident that the nature of these mechanisms can be addressed in psychotherapy and if this was done the client would experience a significant and sustainable improvement in their difficulties.

(5) This conceptualization has identified and highlighted a main psychological mechanism (or a few) to explain the client's difficulties and why they are being maintained. This explanation is clear and accessible, and you feel confident addressing these mechanisms would result in meaningful change for the client. In addition, this conceptualization has at least some of the following features: the origin of the psychological mechanism is considered, the reasons for any recent exacerbation or current worsened impact are considered, the psychological mechanism is placed in a cultural, sociological, or biological context as well, or the reasons for why this mechanism has not yet been overcome by the client are considered. In this way the mechanism is not only clearly identified, but information relevant to how it may eventually be tackled has also been included.

6) Strengths focus, resiliency, positive regard.

Case conceptualizations may vary in terms of how solely negative, problem focused, and disheartening they are, versus being hopeful, strengths focused, positive, normalizing, and resiliency focused. A focus on strength and resiliency will highlight ways that the client's own positive characteristics will be useful for treatment and could be used to help engender hope and marshal client effort. Descriptions and conceptualizations express positive regard and normalize problems may help the client and clinician feel more open to approaching difficult topics, and help the client feel less shame and fear. At their worst, a case conceptualization may contain ideas or wording that is harmful, stigmatizing, or client blaming, and contain very little in the way of empathy, hope, and non-judgmental curiosity.

(1) The conceptualization has elements that you feel may seriously insult the client, seriously fail in terms of empathy and positive regard, or place an unreasonable amount of blame and harsh responsibility on the client. This conceptualization may have been written by someone who views other people's problems as "their fault" and has a critical tone. The conceptualization may also be fatalistic, implying nothing can change and that the client is doomed to remain unwell or locked in their challenges.

(2) The conceptualization has a few somewhat negative elements such as those listed for level one: perhaps being a bit harsh in one spot, or using language which is stigmatizing. However, the negative aspects of the conceptualization are not overwhelming or striking, and are balanced by other more neutral or even kind elements.

(3) This conceptualization contains no obvious insulting, fatalistic, or overly harsh elements. If a client's personal role in their difficulties is discussed, the tone and wording is overall quite neutral. There are no, or very few references to client strengths, examples of client resiliency, and the conceptualization may have a somewhat detached and impersonal feeling.

(4) The conceptualization conveys at least a small amount of positivity, hopefulness, and respect for the client. Client strengths, examples of resiliency, and client characteristics that may be useful for psychotherapy success are included. Where a client's personal role in their difficulties is considered, or when self-defeating patterns are identified, the conceptualization of these are done in a sensitive manner. The tone and nature of the conceptualization is less impersonal and has started to show the client as a person the clinician can connect with.

(5) The conceptualization conveys a strong feeling of hopefulness, positivity, and respect for the client. Several strengths, examples or resiliency, and positive client characteristics are considered and highlighted. Effort has been put in to show how the client has strength and potential for growth, through highlighting resiliency, or traits that will enhance the likelihood of success. Where a client's role in their difficulties or self-defeating patterns are identified the conceptualization does so in a way that helps normalize them, or explains their cause in a way that would minimize blame and shame. The tone of the conceptualization is warm and caring, demonstrating a liking and positive regard for the client.

7) Theoretical and Logical Grounding

Case Conceptualizations should be grounded in existing theories on the nature of, development of, and maintenance of psychological difficulties. Where this is not possible they should at least be logical, plausible, and not pseudoscientific or demonstrably incorrect. For this item examine the claims explicitly or implicitly present in the conceptualization and evaluate them for how logically sound they are, and more how closely they correspond with theory and practice in psychotherapy. The lowest scores represent claims that are non-testable or not grounded in viable psychotherapeutic theory (the client is depressed due to past-life experiences), and/or claims with logical contradictions or non-sequiturs. Moderate scores represent conceptualizations that contain few claims or ones that are logical and plausible but lie outside typical approaches in psychotherapy. High scores will represent logical conceptualizations which draw from established approaches in clinical psychology and which appear supported by the information in the conceptualization.

(1) This conceptualization contains claims or perspectives which have glaring logical non-sequiturs or contradictions, or references pseudoscientific theories which are clearly not in the scope of any established psychotherapeutic tradition. The creator of this conceptualization appears very unfamiliar with clinical psychological frameworks for understanding human well-being or psychopathology, and has created explanations that are difficult to follow, or which are factually incorrect.

(2) This conceptualization may include relatively few (or even no) claims, and/or those which are included could be considered plausible. As such, the conceptualization does not contain any clearly false, illogical, or grossly pseudoscientific claims. The conceptualization nonetheless has few claims in line with established psychotherapeutic tradition and practice.

(3) Many of the claims in this conceptualization appear grounded in traditional or current approaches to psychotherapy. There are signs that the creator has at least some familiarity with psychological, sociological, or biological theories with relevance to psychotherapy. Remaining claims are plausible, do not represent logical leaps or stretches of the imagination, but are not critically examined.

(4) Most claims in this conceptualization are consistent with established approaches to understanding mental health difficulties in clinical psychology and psychotherapy. When any claims are not obviously based off established theories in the field, the creator has still produced a plausible idea which fits logically in the conceptualization. Overall; however, these less established claims appear to be made somewhat offhandedly, with minimal efforts to justify them.

(5) The creator of this conceptualization has very successfully interpreted the client’s situation through well supported theories and psychotherapeutic approaches on human well-being and psychopathology. Claims are made in accordance with clinical psychological traditions and are expressed with appropriate firmness or tentativeness based on the level of inference or the science available. When integrating different theories, or when conceptualizing unique or under-researched client information, the creator still utilizes logic, solid reasoning, or broader theoretical frameworks, and appears particularly mindful of straying from established ideas.

Appendix 2 | Examples of Pilot Study Conceptualizations

Examples of High Scoring Conceptualizations

Example 1: Irene is a young adult woman pursuing a diploma in business at a local college. She is hoping psychotherapy will help her overcome social anxiety which she describes as having been life-long, but never as impairing as it is currently. Her business courses require many presentations and she also feels there is a pressure to network and engage with local businesses. When presenting or meeting new people Irene described symptoms of what appear to be panic attacks leading up to the social situation, which eventually morph into feeling numb and as though everything around her is a movie. She finds herself struggling to attend to anything except her own nervous thoughts during social situations, and will ruminate for some time afterwards on how her anxiety made her perform poorly, despite not remembering nearly anything about how the event actually went. Irene's anxiety appears to have worsened as a result of the increased demands for socializing in this program, but she appears determined not to drop out of college. At a more immediate level, her very high anticipatory anxiety appears to provoke panic attacks, which in turn leave her feeling numb and with a blank mind while actually in the social situation. The effect is she cannot reflect on her objective performance, which although likely impaired somewhat by her anxiety, is unlikely to be as terrible as she feels. Irene has a supportive group of new friends through this program who appear somewhat able to empathize with her struggles, and who may be able to help her reflect more accurately on her performance in presentations. Irene's family appears less positive, as she described having intentionally distanced herself from her parents and brother as they often were angry at her for missing family gatherings, likely due to her social anxiety.

Example 2: Meeting with Mark offered me a chance bring some clarity to a complicated and long history of mental illness. Mark has had a difficult life, their parents divorced when he was young, he moved cities several times before high school, and he was diagnosed with a reading disorder which has disrupted his ability to thrive in school. Despite this he had a small, but close group of friends who he remained in touch with, even after moving. This was fortunate as Mark began experiencing anxiety and depression in his early teen years, and claims that if it wasn't for his friends he "might not still be around". However, receiving support through online interactions was not sufficient to remit his depression, which has been with him nearly constantly since his teenage years. Mark's anxiety has shifted focus over the years, and currently he spends a significant amount of time worrying about his health. The larger theme that appear to be driving his difficulties are concerns about something "being wrong with him", Mark seems to blame himself (and his academic difficulties) for his parent's divorce, he feels his limited social life is due to him being a "weird guy", and he voices a hopelessness about his future because "no matter else what changes, I'll still be me". It seems his time and effort worrying about health issues could be a manifestation of his belief that there is something fundamentally wrong with him, and holding such a view has apparently left him depressed without relief for some time. It appears more likely that Mark is a victim of his life circumstances, his guilt and self-directed anger of his parents' divorce is misplaced, his frequent moves help explain his limited real life social circle, and there is nothing objectively "wrong" about him - he is pleasant and intelligent.

Example 3: The client's struggles can best be captured by a diagnosis of histrionic personality disorder. They report a history of impulsive decisions, mood swings, rocky relationships, and inappropriate attention seeking leading back to young adulthood. In the therapy context they've overestimated my degree of attachment to them and the depth of our connection (so soon into therapy) and also made several suggestive comments about how attractive I must find them. When, after a session, I had to speak to another professional and only gave the client a quick goodbye, they erupted into tears, which ended as soon as I checked back with them, suggesting a somewhat shallow and intentional emotional reaction was being displayed. They may have inherited a genetic predisposition towards this form of personality disorder, as both parents were reportedly very emotionally intense and "theatrical". As well, that chaotic early childhood environment may have left the client struggling to receive parental affection and attention, leading them to also adopt an dramatic persona in order to compete better for any attention. The exact reason for seeking treatment is unclear, but the client's current friendships appear to be weakening, and the client reported many disruptive and likely hurtful behaviours they engage in while socializing, such as teasing others who appear more likeable, complaining loudly about personal struggles to the exclusion of others input, "starting shit" between friends in order to "liven parties up", and making sexual passes at people who are already in romantic relationships as the client "enjoys flirting". As with many personality disordered clients, there is a lack of true awareness of the impact of their behaviors and a lack of awareness of alternative means of getting the attention and validation they appear to deeply desire. They denied other mental health issues, but I suspect they may be withholding depressed feelings as they conflict with their pride in their exuberance.

Examples of Low Scoring Conceptualizations

Example 1: This client has been experiencing difficulties in their interpersonal relationships, these difficulties seem to stem from a tendency to displace feelings and expectations from one person to another, a process called transference. Based off negative experiences with an early caretaker they now expect most people will have a critical and angry attitude towards them, which is not necessarily true. Entering social situations with this expectation leads to a self-fulfilling prophecy, as the client is already ready to defend themselves from attack and is more prone to seeing aggression when it is not actually occurring. Most unfortunate is how this client's own mind is the source of their most painful insults, they have a very well developed self-critic, overdeveloped in fact. This altogether has led to their isolation, low mood, lack of romantic relationships, and tendency towards feeling unsafe and anxious most of the time.

Example 2: Greg has been having a lot of difficulty at work. His boss is always yelling at him, his co-workers are mean to him, and his output and productivity are low. Greg thinks he is the problem, because he doesn't know how to stand up for himself. Greg should take some boxing classes to improve his self-esteem. He should also not work at such a toxic workplace. Greg should also consider getting some advice on his finances, as he described spending a lot of money on food, which he seems to eat a lot of. He seemed pretty uncomfortable in the video and could have been more open with the therapist.

Example 3: The patient's name is Ava-Mai, she is a sixty-five-year-old woman who has been experiencing a great amount of emotional distress recently. She is hoping therapy will help her find happiness again. She presents as someone with much less maturity than others her age, for example she enjoys speeding in her car, she complains with a teenage tone about other people who live in her apartment building, and she smokes and drinks somewhat excessively without much recognition of the health impacts these pose. She appears to project her own personality onto her son, as he is described by her as being quite immature, making poor decisions, and being too lazy to make anything of himself, this is despite him working towards an apprenticeship as an electrician, having made the honour roll in high school, and having a clean legal and academic record. Ava-Mai had a turbulent childhood, noting a severe and unkind upbringing from her parents which likely has influenced her personality and approach to life. Ava-Mai also has lost her husband recently, he died of a heart-attack. She appears to be struggling to process her grief well.

Appendix 3 | Self-Report Items and Measures from Studies 1 and 2

Need for Cognition Inventory

1) I prefer simple to complex problems.

2) I like having the responsibility of handling a situation that requires a lot of thinking.

3) Thinking is not my idea of fun.

4) I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.

5) I try to anticipate and avoid situations where there is a chance I will have to think in depth about something.

6) I find satisfaction in deliberating hard and for long hours.

7) I only think as hard as I have to.

8) I prefer to think about small daily projects to long term ones.

9) I like tasks that require little thought once I've learned them.

10) The idea of relying on thought to make my way to the top appeals to me.

11) I really enjoy a task that involves coming up with new solutions to problems.

12) Learning new ways to think doesn't excite me very much

13) I prefer my life to be filled with puzzles I must solve.

14) The notion of thinking abstractly is appealing to me.

15) I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

16) I feel relief rather than satisfaction after completing a task that requires a lot of mental effort.

17) It's enough for me that something gets the job done: I don't care how or why it works.

18) I usually end up deliberating about issues when they do not affect me personally.

Interpersonal Reactivity Index

1) I daydream and fantasize, with some regularity, about things that might happen to me.

2) I often have tender, concerned feelings for people less fortunate than me.

3) I sometimes find it difficult to see things from the "other guys" point of view.

4) Sometimes I don't feel very sorry for other people when they are having problems.

5) I really get involved with the feelings of the characters in a novel

6) In emergency situations, I feel apprehensive and ill-at-ease

7) I am usually objective when I watch a movie or play, and I don't often get completely caught up in it.

8) I try to look at everybody's side of a disagreement before I make a decision.

9) When I see someone being taken advantage of, I feel kind of protective towards them.

10) I sometimes feel helpless when I am in the middle of a very emotional situation.

11) I sometimes try to understand my friends better by imagining how things look from their perspective.

12) Becoming extremely involved in a good book or movie is somewhat rare for me.

13) When I see someone get hurt, I tend to remain calm.

14) Other people's misfortunes do not usually disturb me a great deal.

15) If I'm sure I'm right about something, I don't waste much time listening to other people's arguments.

16) After seeing a play or movie, I have felt as though I were one of the characters.

17) Being in a tense emotional situation scares me.

18) When I see someone being treated unfairly, I sometimes don't feel very much pity for them.

19) I am usually pretty effective in dealing with emergencies.

20) I am often quite touched by things that I see happen.

21) I believe that there are two sides to every question and try to look at them both.

22) I would describe myself as a pretty soft-hearted person

23) When I watch a good movie, I can very easily put myself in the place of a leading character.

24) I tend to lose control during emergencies.

25) When I'm upset at someone, I usually try to "put myself in his shoes" for a while.

26) When I am reading an interesting story of novel, I imagine how I would feel if the events in the story were happening to me.

27) When I see someone who badly needs help in an emergency, I go to pieces.

28) Before criticizing somebody, I try to imagine how I would feel if I were in their place.

Big Five Inventory - Short Form -2

I am someone who...

1) Tends to be quiet.

2) Is compassionate, has a soft heart.

3) Tends to be disorganized.

4) Worries a lot.

5) Is fascinated by art, music, or literature.

6) Is dominant, acts as a leader.

7) Is sometimes rude to others.

8) Has difficulty getting started on tasks.

9) Tends to feel depressed, blue.

10) Has little interest in abstract ideas.

11) Is full of energy.

12) Assumes the best about people.

13) Is reliable, can always be counted on.

14) Is emotionally stable, not easily upset.

15) Is original, comes up with new ideas.

16) Is outgoing, sociable.

17) Can be cold and uncaring.

18) Keeps things neat and tidy.

19) Is relaxed, handles stress well.

20) Has few artistic interests.

21) Prefers to have others take charge.

22) Is respectful, treats others with respect.

23) Is persistent, works until the task is finished.

24) Feels secure, comfortable with self.

25) Is complex, a deep thinker.

26) Is less active than other people.

27) Tends to find fault with others.

28) Can be somewhat careless.

29) Is temperamental, gets emotional easily.

30) Has little creativity.

Mental Health Skill and Interest Items

1) I am familiar with psychotherapy (either personally, through courses/reading, through training)

2) I am interested in eventually working in the mental health field, where I might provide psychotherapy.

3) I have a good understanding of what causes people to experience mental health problems

Empathy for Larry Scale

1) I would care about him and his difficulties

2) I would want to understand what he was going through

3) I would be impacted by his feelings

4) I would think he deserves compassion for the problems he is facing

5) I could see things from his point of view

6) I would feel protective of him

7) I would want to hear more about his story and experiences

8) I could imagine myself "being in his shoes"

9) I would be sympathetic to even the problems he seems to be causing for himself

Abbreviated Mental Health Stigma Scale

1) If someone has experiences severe mental health illness, they will suffer for the rest of their life.

2) People with severe mental illness are failures.

3) In spite of any efforts they are making, people with severe mental illness will never be like other people

4) Severe mental illness is easily recognizable

5) People with severe mental illness are dangerous

6) Severe mental illness is caused by bad luck

Data Quality Items

1) I was significantly distracted by off task thoughts (unrelated to the study) during the video

2) I personally know the actor in the video

3) I put forwards my best effort on the tasks in this study

4) The written description I made was accurate to my internal thoughts and understanding about Larry and his problems

5) If I hadn't been told ahead of time I might have thought that was a real video of a therapy session

6) I am confident that my understanding of Larry is correct

Appendix 4: Key Features from "Master Conceptualization" of Larry

1) Social Anxiety

2) Anxiety

3) Avoidance (behavioural)

4) Shy/introverted (personality)

5) Judgment/sense of scrutiny (worries about, fear of)

6) Poor self-concept

7) Physical symptoms of anxiety (sweating, shortness of breath, feeling hot)

8) Depression

9) Sadness

10) Low mood

11) Low motivation (due to low mood)

12) Self-worth tied to achievement

13) Anger, resentment/negative view of others (jealous, feel left out)

14) Poor/problems with communication skills/assertiveness

15) Lack of social support, close friendships

16) Unhealthy Coping

17) Alcohol

18) Video games

19) Smoking

20) Parents, upbringing (conflict), poor role modeling re: relationships

21) Grandmother’s death

22) Move (from US to Canada)

23) Excluded, bullied

24) Procrastination

25) Transition to university difficult

26) Academic difficulties because of mental health

27) Poor emotional awareness/expression

28) Has succeeded academically before

29) Has a few friends currently

30) Willing to attend therapy

Appendix 5 | Study 1 Results (Influential Outliers Not Excluded)

Table 20

BAQCS Correlations to Study Scales and Variables, Influential Outliers Not Excluded (Written Conceptualizations)

Variable	*n*	*r*
Criterion Validity Variable		
Number of Key Features	68	.65**
Convergent Validity Variables		
Participant GPA	41	.45**
Mental health skill and interest	70	.32**
Need For Cognition total	72	.24*
Empathy for Larry	71	.23
Interpersonal Reactivity Index		
Fantasy subscale	72	.29**
Perspective taking subscale	72	-.08
Empathic concern subscale	72	.14
Personal distress subscale	71	-.15
Openness	72	.15
Agreeableness	72	-.01
Mental health stigma items	71	-.17
Discriminant Validity Variables		
Conceptualization word count	72	.27*
Flesch-Kincaid grade level	71	.26*
Other Big Five Traits		
Conscientiousness	72	-.07
Extraversion	72	-.23*
Neuroticism	72	.06

note: *** p < .001, ** p < .01, * p < .05

Table 21

Regression Predicting BACQS Total Scores, Influential Outliers Not Excluded (Written Conceptualizations)

Variable	*β*	*t*
Warm Empathy	.04	.30
Affective Empathy/Distress	.12	.77
Detached Openness	.37*	2.55
Introspection	.28	1.98
Number of Words	.27*	2.23

note: * = significant at the *p* < .05 level, ** = significant at the *p* < .01 level.

Regression result: R^2 = .22, *SE* = 10.44, *F*(5, 61) = 3.51, *p* < .01

Appendix 6 | Examples of High and Low Scoring Conceptualizations from Study 1

High Scoring Conceptualizations

Example 1: I think the main stem of the clients academic struggles stem from a fear of social judgement. This fear of having "all eyes on him" causes him to not go to class, not participate in class and inables him to pay attention to what the professor is saying. I think this fear could have started in elementary school when most of his classmates seemed to bully him. When he transitioned into High School, the bullying seemed to disapate shortly after the start but then but then no one seemed "interested in being his friend" causing him to believe he was not good enough. This doubt can create a fear of messing up which in return allows for a developing fear of having people notice your mistakes, which all leads back to a fear of being judge by others. The client may also have trouble developing close relationships with others as growing up his parents failed to demonstrate a close family bond or marital relationship. Having the only memories of parental connect be arguing, can make a child feel that no interaction is better than any, resulting in the ideas of "me against the world". This weak connection now with his parents, can also be a cause a feeling isolation which may also result in him not developing a support system with his fellow peers. An important part to keep in mind is that the transition into university, and importantly into a comptative program that he is in, will always be difficult. These academic troubles he is facing is not steming from the inability to understand the material but more from the inability to focus and his anxiety around attending class.

Example 2: larry is struggling from a low performance in school, as identified form his professor and poor returning marks. A lot of what seems to be causing the academic difficulty isn't the content, as he states that he feels capable, but that there is a lot of anxiety around social evaluation in University for him. Larry has trouble gettin gup for 8:30 am lectures, as many students do, but he won't even go if he's late because he feels that that is drawing more attention to himself, and he can "feel the eyes"; by not going, he misses a whole lecture, not just part of it, and by consistently missing this lecture, it may be wearing on him in the back of his mind that he's skipping, and the traits associated with it (lazy, not high performing, etc.) He's also anxious about answering questions in class (and likely asking questions), again based off of the anxiety around social evaluation that he has. This anxiety comes back to play into how he struggles to get work done while he is in his dorm, as his peers are loud on the floor, and he doesn't want to talk to them because he feels that they wouldn't stop being loud just because a roomate asked them to. He appears to have a low self-esteem, with some of his self-describing words like being "uncool" in high school, and how " girls don't want to get to know me"; this self-esteem might be playing into axiety, as being uncomfortable with yourself might be spurring on the idea that others are uncomfortable/judging you as much as you are judging yourself, when it's often not the case. Larry also seems to have a weak support system; he has stated already that he's not comfortable communicating with roomates, feeling that they don't care for his concerns, so he liekly doesn't communicate with them much as it is. He also doesn't keep touch with his parents so much, aside form the odd call and when they pay for some stuff for him; his relationship was described as never having been close with them, so that's already a weak foundation. He's in first year, so the transition is just adding to it. When asked about his friends, he details a few people who he "smokes with"; no mention of any other activities done with them, and doesn't mention any peers that he is close with through academics. The "friends" that he has seems to be a negative infleunce, depending on drugs as a mediator and not related to any element of well-being. he doesn't even seem to be very close to them as it is. Overall, he just seems very isolated. And being isolated will influence his self-esteem as well, because again not having people who seem to genuinely care about your well-being and performance will influence how much value you feel towards yourself. uses video games potentially as that distractor, to fill in for a social support system, and what he turns to when he doesn't want to do academic work (it's a bit of a stretch. This poor support system would also add to anxiety, if impplicitly, because if you don't feel like you have a good net to catch you, you're more worried about negative social evaluation being detrimental to your social standing

Example 3: Larry is very self critical of himself. For example, he sees himself as uncool and a loner. This nature of his self-critical thoughts is worsened by the University environment. The thought of being noticed or judged by his peers makes him very anxious. Because of his anxious thoughts regarding being judged, Larry is skipping class and avoiding his school work. This is causing Larry to be even more self-critical, and contributes to his overall low mood. His anxiety over being judged has also prevented him from making many close relationships to the people around him in University. The fact that he does not havy many friends also contributes to his anxiety and low mood. To cope with his anxious thoughts and low mood, Larry drinks alcohol and plays videogames, which is a way of avoiding his classes, school work, and creating relationships. Again, because of his low grades and lack of relationships, he is self-critical, anxious, and has a generall low mood.

Low Scoring Conceptualizations

Example 1: The client's difficulties stems from the fact that currently he is failing his test and midterms, which will eventually result in a failing of the courses. There are a couple factors that play a role for the client being in this position. This results from the client not going to class, not being able to concentrate, having no support system and finally having the feeling of being axious all the time. This in turn has some reprecussions in his performance in school, even without him knowing. The client needs to have specific aspects of his life turned around to see some results in making a difference in his school performance.

Example 2: Larry is feeling disinterested, unmotivated and 'off'. These feelings are a combination of things that he felt as discussed in his session. No one is holding larry accountable and as a result he is getting bad grades. When he tries to do work, he can't focus, which is something that has begun happening recently since coming to university. To help him focus, larry describes having 1-2 drinks to 'take the edge off' when studying. He says this helps to relieve the tension he feels. Larry has disclosed he does not have a good relationship with his family and as a result does not speak to them often. He explained that whenever he sees his parents they are often fighting and they do not 'share dinners together' like most families do. He also explained that he has made 3 friends since coming to university. This leads him to feel like noone cares, as no one is ever checking in on him to see how he is doing. I believe that larry needs to devote time to studying in silent spots, reach out to his family more, and attend more organized social events on campus such as clubs/dances/game nights. As a result, I think these 3 steps will help boost Larry's sense of belonging and help him to have a higher mood.

Example 3: Based on the guy's talking of himself's life so far, he has to two problem relative to his life- acadomic problem and little infamtile actism. First, his major problem in collage life is falling in class. The cause of this problem is from two parts: hard to be concentrated in study and not learn enough lecutr knowleage. He was skipping his morning class often, so he didn't know enough leture in order for him to pass the test; moreover, he always didn't get his homewrok done due to that he had a hard time to concetrated in study, becasue of change in to a new enviornemnt , feel people's eye on him and care what others are thinking about him. The reson why he cares about other so much is because of he thinks that himself is not cool enough to let people want to make friends with him due to past experience that he didn't has close friends to talk to and not girl want to date him. In addition, his childhood was not a good memory for him, which the family moved to another place after his grandma was died and his parents was flighting all the time. All those unhappy experience cause him to scared to talk to people, then he is not having enough social life to have someone to talk to. To this end, all those situiton letting him have infantile autism.

Appendix 7 | Study 2 Results (Influential Outliers Not Excluded)

Table 22

BACQS Correlations to Study Scales and Variables, Influential Outliers Not Excluded (Case Conceptualization Diagrams)

Variable	*n*	*r*
Convergent Validity Variables		
Participant GPA	36	.32
Mental health skill and interest	64	.13
Need For Cognition total	66	.12
Empathy for Larry	65	.17
Written conceptualization scores	67	.50**
Interpersonal Reactivity Index: scale total		
Fantasy subscale	66	.31*
Perspective taking subscale	66	-.01
Empathic concern subscale	66	.11
Personal distress subscale	66	.00
Openness	66	.03
Agreeableness	66	-.01
Mental health stigma items	65	-.19
Discriminant Validity Variables		
Number of Elements	66	.31**
Familiar with concept maps	65	.11
Other Big Five Traits		
Conscientiousness	66	-.05
Extraversion	66	-.07
Neuroticism	66	.04

note: *** *p* < .001, ** *p* < .01, * *p* < .05

Table 23

Regression Predicting BACQS Total Score, Influential Outliers Not Excluded (Case Conceptualization Diagrams)

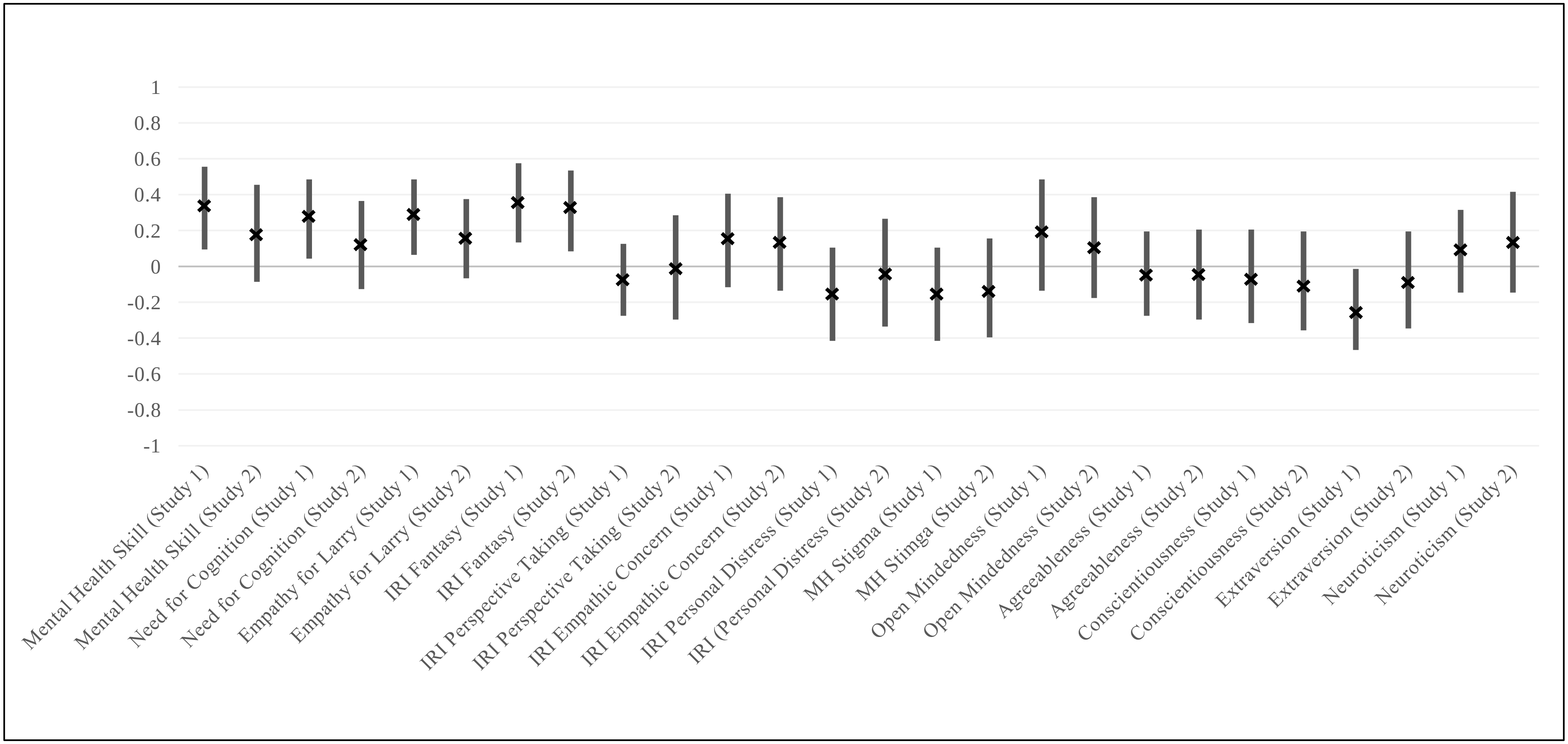
Variable	*β*	*t*
Warm Empathy	.14	1.13
Affective Empathy/Distress	.16	0.97
Detached Openness	.06	.42
Introspection	.15	1.03
Number of Elements	.34**	2.86

note: * = significant at the $p < .05$ level, ** = significant at the $p < .01$ level.

Regression result: $R^2 = .19$, $SE = 10.97$, $F\ (5, 60) = 2.85$, $p = .02$

Appendix 8 | Study 2: Correlations Between BACQS Scores and Other Variables for Written and Diagram Conceptualizations

Figure 2: Point Estimates and 95% CI for Correlations Between BACQS Scores and Other Variables Across Studies 1 and 2



note: 95% Confidence intervals obtained via bootstrapping