# Exploring the Reliability of an Objective Severity Tool to Classify Severe Problem Behaviour

Marie-Chanel Morgan

Bachelor of Science

Ontario College Graduate Certificate

Applied Disability Studies, Applied Behaviour Analysis

Submitted in partial fulfillment

of the requirements for the degree of

Master of Arts

Faculty of Social Sciences, Brock University

St. Catharines, Ontario

# Abstract

The term ‘severe’ is a common descriptor for problem behaviour in research and practice. However, it is often applied inconsistently, and at times based on ill-defined or arbitrary criteria. Existing problem behaviour measurement tools often rely solely on caregiver recall (e.g., interviewing primary caregivers). This study explores the reliability of the first iteration of a severity tool employing direct measurement strategies (e.g., response rate, injury severity as evidenced by permanent product) to classify an individual’s problem behaviour severity. Nine Board Certified Behavior Analyst (BCBA) raters were recruited, five novice raters and four expert raters. They each experienced two conditions. In the first condition, raters classified the severity of 20 case scenarios without access to the tool. In the second condition, raters classified the severity of 20 novel scenarios after completing the tool for each case. All items of the tool (*n*=26) had good internal consistency (∝=.831). Intraclass correlations showed a meaningful increase in reliability for both groups when they had access to the tool (novice *r*=0.860, expert *r*=0.912) compared to when they did not have access to the tool to rate case severity (novice *r*=0.781, expert *r*=0.803). Most raters either strongly agreed or agreed that the severity tool had good applicability across research and clinical settings. This suggests that inconsistencies that may exist in the classification of severe problem behaviour could be mitigated with the proposed tool.

*Keywords:* severe problem behaviour, intellectual and developmental disability, reliability, severity scale, research tool

## Acknowledgements

This study and body of work far surpassed what I believed I could have accomplished. The qualities I have learned and applied during the years of developing and drafting my thesis cannot possibly be captured by the words on this page. One thing is for certain, I would not have dared to type the first letter without the specific series of mentors, friends, and colleagues that I had the honor of knowing. First, to my father, who refused to see me limit my future and always followed my “I can not’s” with “why not’s?”. Over the last year of his life, he asked me each day I saw him when I would be done school. The answer is now Daddy. I am fortunate that he now exists in a place to watch me from the sky and attend every last graduation. To my mother who loves me as best she can and reminds me not to lose my smile in the chaos of growth. To my big sisters, the women who do not bat an eye when I turn to them for help. The women that challenge me to grow for more than just the present moment, but to grow for all the moments to come. To my grown nephew who made “I am so proud of you” part of his regular greeting. I am eternally blessed to have a family who found every achievement a celebration and supported me through my persistence of this research. Its finally here! To my friends, a unit of people I have carefully handpicked to call my family. A foundation of strong, independent, hilarious, and daring people who have built the escapes in nooks of the mountain climbed in a master’s degree. Lastly, to the woman who made this all possible. She saw the potential of this thesis from an eager student who whispered to her classmate “I want to be her” on the first day of her first class, watching Dr. Alison Cox teach. It is through her passion, humour, and humility that I learned to grow into the student and researcher I am today. I am not sure what sparked her to ask me to join this journey, however, this paper is the most spectacular view from mountains I have climbed. I

stand tall and proud to present this body of work and acknowledge that I stand on the backs of giants.

## Table of Contents

## List of Tables

## List of Figures

## List of Abbreviations

| | |
|---|---|
| ABA | Applied Behaviour Analysis |
| IDD | Intellectual or Developmental Disability |
| ASD | Autism Spectrum Disorder |
| ADHD | Attention-deficit Hyperactivity Disorder |
| SIB | Self-injurious Behaviour |
| ABC | Aberrant Behaviour Checklist |
| BPI | Behaviour Problems Inventory |
| TIRF | Treatment Intensity Rating Form |
| CGI | Clinical Global Impression Scale |
| SDAS | Social Dysfunction and Aggression Scale |
| BCBA | Board Certified Behaviour Analyst |
| BCBA-D | Board Certified Behaviour Analyst – Doctorate |
| RBT | Registered Behaviour Technician |
| BCaBA | Board Certified Assistant Behaviour Analyst |
| SPSS | Statistical Package for Social Sciences |
| ICC | Intraclass Correlation |
| RBS-R | Repetitive Behavior Scale – Revised |

Exploring the Reliability of an Objective Severity Tool to Classify Severe Problem Behaviour

# Introduction

Applied Behaviour Analysis (ABA) is a scientific approach to understanding behaviour and involves the application of learning principles to address socially significant behaviours (Cooper et al., 2020). Regardless of the targeted behaviours (e.g., skill acquisition or behaviour reduction), behaviour analysis emphasizes the importance of relying on objective measurement strategies to accurately report research and clinical outcomes (Morris et al., 2013). One of the most influential articles by Baer et al. (1987) established the seven dimensions of ABA as guiding tenants in research and practice. Arguably, objectivity is an underlying theme across each dimension. Adhering to the *technological* dimension means providing written objective descriptions of the intervention, setting and participants. More recently, Morris et al. (2013) reviewed important components of research in ABA and identified specific criteria required to meet the standards of behaviour analytic literature. They identify that inclusion criteria and participant descriptions should rely on quantitative and direct observational means. This suggests that the use of specific terminology (i.e., *severe* problem behaviour) to describe participants should adhere to the same expectations of relying on quantitative and direct measures of the target behaviour. Direct measurement is the collection of information about a behaviour through observation or permanent products such as property damage (Cooper et al., 2020). Direct measurement may be considered more objective than indirect measurement because direct measurement relies on observed events, whereas indirect measurements can be inaccurately reported or bias (Floyd et al., 2005).

Broadly, problem behaviour can be operationalized as any behaviour which threatens the safety and quality of life for an individual and/or their caregivers (Evers & Pilling, 2012;

Hanratty et al., 2015; Lowe et al., 2007). Problem behaviour topography and the magnitude of interference on an individual's quality of life varies (Lowe et al., 2007). Typically, behaviour analytic researchers and clinicians create specific operational definitions allowing for individualized descriptions of topographies for observation (such as collecting data) or reporting purposes (such as article descriptions) (Cooper et al., 2020).

For the purpose of clarity, I will use the term severe to discuss how authors use this term as a descriptor of an individual's problem behaviour. Although severe is a common term used to describe the risk or magnitude of an individual's problem behaviour, there is no agreed upon standard for the use of this term in behaviour analysis. In fact, there are many examples of the inconsistent use of this term across the literature (Hanratty et al., 2015). There have been indirect measures developed to rate problem behavior (e.g., Aberrant Behaviour Checklist (ABC); Aman, 2013, Behaviour Problems Inventory (BPI); Rojahn et al., 2012); however, they cannot address the variability and subjectivity of the term severe because indirect measurements are secondary sources of data such as caregiver report, questionnaires, and interviews (Cooper et al., 2020). Unfortunately, there are many drawbacks with these tools stemming primarily from their reliance on indirect measures (Evers & Pilling, 2012; Lowe et al., 2007). Other researchers have attempted to objectively define and classify problem behaviour by designing tools that rely on observational data (Hanratty et al., 2015). However, many of these tools are designed for very specific diagnoses (e.g., anxiety, depression, or autism) (e.g., The Baby and Infant Screen for Children with autism Traits, Part 3; Hanratty et al. 2015), problem behaviour topographies (e.g., self-injurious, repetitive or aggressive behaviours) (Repetitive Body Focused Behavior Scale; Selles et al., 2018), or ages (e.g., Children, Adolescents, or Adults) (The Child Behaviour Checklist; Hanratty et al. 2015). Additionally, no tool in the literature relies entirely on objective means such as data collection, medical reports, or legal documentation to classify problem behaviour severity.

Developing and empirically evaluating a severity tool that relies entirely on objective means to classify problem behavior severity may be one way to facilitate consistent use of the term severe across the literature. By making the use of this term accountable to quantitative methods such as direct measurement and permanent products may help to remove bias as a factor during behavioural assessment. Following the development of such a tool, researchers should test their tool's reliability given reliability is recommended as the first step before application or validation (Koo & Li, 2016). Further, the development of a severity tool may have many collateral implications for practitioners, researchers and stakeholders.

## Intellectual and Developmental Disability

An Intellectual and Developmental Disability (IDD) is characterized by impairments affecting adaptive functioning including social, conceptual, and practical domains (Baker, 2017). Language deficits and communication disorders are common with this diagnosis and may bring about a wide range of problem behaviour (American Psychiatric Association, 2013). These impairments are first noticed during the developmental period and may affect both physical and cognitive development (American Psychiatric Association, 2013). The prevalence of IDD is approximately 1% of the general population with varying statistics across age groups and ethnicities (American Psychiatric Association, 2013; Friedman et al., 2018). Lifetime costs of health and social care resources for these individuals are estimated to be 2.4 million (Emerson, 2010). The degree of impairment is typically measured by adaptive functioning rather than standardized intelligence testing, and presents differently for each individual (Baker, 2017).

Neurodevelopmental disorders such as Autism Spectrum Disorder (ASD) and Attention-deficit/Hyperactivity Disorder (ADHD) are included in this group.

## Problem Behaviour in Persons with Intellectual Developmental Disabilities

Common broad topographies of problem behaviour include aggression, self-injury, property destruction, inappropriate social or sexual conduct, non-compliance, motor or vocal stereotypy, and inappropriate consumption of non-edible objects (pica) (Emerson, 2001; Poppes et al., 2010). Problem behaviours often have many negative effects on the individual and their caregivers such as physical injury, chronic pain, social isolation, and physical abuse (Evers & Pilling, 2012; Poppes et al., 2010). Lasting consequences such as a restriction to personal development and poor community integration often coincide with problem behaviour (Poppes et al., 2010).

Poppes et al. (2010) report that individuals with IDD are three to five times more likely to engage in severe problem behaviours. Further, the prevalence of severe problem behaviour in persons with an IDD have been estimated at 10%-15%. However, statistical accuracy may be questionable due to ambiguous and subjective descriptions (Evers & Pilling, 2012; Lowe et al., 2007). For example, Lowe et al. (2007) reviewed three studies that claimed different statistics of severe problem behaviour. The authors attributed differences to the generality of definitions used in the literature and the lack of objective definitions to adequately identify the inclusion or exclusion of specific topographies.

## Limitations of Qualitative and Quantitative Participant Descriptions in the Literature

### *Qualitative*

Behaviour analytic journals include participants described as engaging in severe problem behaviour. However, authors often used this term in situations featuring vastly different participant profiles (Foxx, 2003; Hausman et al., 2009; Lowe et al., 2007). Authors commonly associate severe with terms such as *dangerous* but fail to provide any specific objective definition or behavioural anchor for this term (Bonner & Borrero, 2019). Other authors rely on caregiver report alone to establish severity. For example, both Knight et al. (2019) and Fritz et al. (2013) attributed the severity of participant's problem behaviour with teacher testimonials and *observations* but did not provide any measures, quantitative or qualitative, to support their claim of labeling participant's behaviour as severe. Although often used, reliance on caregiver recall can be inaccurate (Fahmie & Iwata, 2011). Therefore, observations should be supplemented with quantitative measurement such as data collection on the rate of problem behaviour occurrences. However, there are problems with this solution if it is carried out in isolation. That is, when only one behavioural dimension is used to classify severe problem behaviour.

Some researchers identify severity solely on the topography of problem behaviour. Bonner and Borrero (2019) distinguished severe problem behaviour by identifying the topography of *aggression* only, with no description of specific behaviours or an operational definition (e.g., hitting, scratching, punching). By contrast, Poppes et al. (2010) identified that severe problem behaviours included aggression, self-injurious behaviour (SIB), and stereotypic behaviour. Fahmie and Iwata (2011) then contradicted this when they conducted a review to determine the frequency with which precursors have been reported in the severe problem behaviour literature. Specifically, they excluded stereotypic behaviours because they claimed this topography was not considered severe due to the limitation of bodily harm to the individual and caregivers. With no agreed upon understanding of problem behaviour severity, the term severe cannot serve as an effective descriptor of problem behaviour.

### ***Quantitative***

I use the term *cut-off* to describe a specific quantitative value to determine or classify a problem behaviour as severe or not severe. Although some researchers have applied various quantitative techniques to justify a participant's severe label (e.g., Emerson et al., 2001; Roscoe et al., 2013); descriptions informing the quantitative techniques may be vague (Jessel et al., 2018; Lowe et al., 2007), ambiguous (Bonner & Borrero, 2019; Lowe et al., 2007), ill-defined or reference arbitrary cut-offs. For example, Oropeza et al. (2018) used injury to caregivers as a measure of severity. They demonstrated that when rates of aggression increased, the likelihood of caregiver injury also increased, and therefore linked behaviour response rate to caregiver injury. However, they also assigned a vague cut-off, such that the paper did not reveal what the cut-off was or provide rationale. Specifically, they discussed that there could be an increase in severity with an "increased intensity of aggression" (Oropeza et al., 2018, p. 685) but had not assigned a specific value of increase in frequency or magnitude. Similarly, Schmidt et al. (2014) claimed to define problem behaviour severity by the most frequently occurring behaviour. Along with no empirical validation, they did not provide any information on the frequency of occurrence or cut-off that informed their classification.

Physical trauma has also been referenced as a measure for severity of SIB. Roscoe et al. (2013) developed an objective measurement system to classify hand-mouthing severity by referring to physical trauma such as swelling, bruising, or a requirement of protective equipment to prevent injury. Their tool functioned within the context of their study however, their goal was problem behavior reduction rather than developing a severity tool. Although they did not assess their measures' validity, Roscoe et al. (2013) is one of the few studies that attempted to assess reliability with one independent observer.

Lowe et al. (2007) conducted a study to determine problem behaviour prevalence to classify severity. They developed a five-point scale applied across 1802 participants' problem behaviour. In this scale *1=serious,* (i.e. occurring daily, and/or the client is usually excluded from activities, and requires physical intervention by caregiver, and/or leads to major injury or damage to caregiver, self or others, *2=serious but controlled*, (i.e., there are existing planned procedures), *3= moderate* (i.e., frequency is serval times weekly, and/or the person is excluded from activities, requires occasional physical intervention by caregiver, and/or results in injury to caregiver, self or others, *4=lesser*, (i.e., it does not occur often, and/or is not usually excluded from activities, and/or caregiver is usually not required for physical intervention, and/or does not lead to major injury to caregiver, self, or others, and *5=None*, (i.e., not the person's typical behaviour). Although their means of categorizing problem behaviour included observable reporting, their five-point scale was not evaluated for reliability or validity. Similarly, their scale is subjective in how a rater may conceptualize "often" or "major injury". This scale also does not address other areas of restriction in an individual's life such as living arrangements or pending legal charges related to their problem behaviour (e.g., assault). Evers and Pilling (2012) evaluated the severity of 45 participants' problem behaviour on a five-point Likert scale with *1= indicating a behaviour causing 'minimal problems'* to *5= indicating a behaviour causing 'serious problems'*. The authors did not require respondents to provide quantitative information (e.g., information based on daily data collection) to inform participant behaviour rating and, thus discussed accuracy limitations resulting from reliance on respondent knowledge.

A study by Emerson et al. (2001) features one of the most comprehensive definitions of severe problem behaviour as a behaviour that:

> Caused more than minor injury to themselves or others or destroyed their immediate living or working environment; or showed behaviours at least once a week that required the intervention of more than one member of caregiver to control, or placed them in danger, or caused damage which could not be rectified by care caregiver or caused more than one hour's disruption; or showed behaviours at least daily that caused more than a few minutes disruption. (p. 78)

Unfortunately, this definition has no validity or reliability in their established cut-offs. Many articles using the term severe to describe participant problem behaviour lack in-depth detail, making articles that do include this information exceptionally noteworthy. Foxx (2003) offered comprehensive descriptions and included specific examples of how the problem behaviour impacted the environment, noting damages as consequences of the target behaviour, lifestyle restrictions, and incurred bodily harm. Although important and helpful, comprehensive definitions highlighting many components associated with problem behaviour severity may fall short of including measurable dimensions, use vague terminology (e.g., "minor injury"), and cannot eliminate subjectivity or promote consistent use of the term across the field.

## Existing Tools and Subjective Reporting Limitations

As mentioned above, an operational definition of severity may not suffice given the range of behaviour topographies, engagement intensity and magnitude. Researchers have designed and validated reliable scales for problem behaviour (Aman, 2013; Hanratty et al., 2015). However, most existing measures of problem behaviour rely on Likert scales and indirect measurements (Rojahn et al., 2003; Rojahn et al., 2012). One of the most widely used tools in research and clinical settings is the ABC (Aman, 2013; Hanratty et al., 2015). This tool was originally developed to evaluate the treatment effects of psychotropic medication on problem behaviour

(Newton & Sturmey, 1988). Although it continues to be used for this purpose, it has also shown to be useful across other applications including examining the psychometric characteristics of other comparable measures, environmental variables to problem behaviours, and evaluating an individual's quality of life (Aman, 2013). The tool consists of 58 items where caregivers score a range of *0= not at all a problem* to *3=the problem is severe in degree* and include five subscales: (1) irritability, (2) lethargy/social withdrawal, (3) stereotypic behaviour, (4) hyperactivity/ noncompliance, and (5) inappropriate speech (Newton & Sturmey, 1988). The ABC has demonstrated strong construct validity and applicability across ages 6 and older (Newton & Sturmey, 1988). Unfortunately, this scale relies solely on indirect systems of measurement (i.e., Likert rating scale) and, thus the interrater reliability and test-retest reliability ranges from low .50s to high .90s (Aman, 2013). It is possible that these ranging values may result from the tool's reliance on Likert scale responding. Although rating scales represent a valid method for many tools across disciplines, the test-retest and interrater reliability are commonly affected by subjective responding and a reliance on caregiver recall (Lowe et al., 2007). This may ultimately limit the utility of the ABC.

Another standardized rating form developed by Zarcone and her colleagues in 2016 is called the Treatment Intensity Rating Form (TIRF). The purpose of the TIRF was to measure, and ultimately inform respondents, of the intensity or complexity of either behavioural or psychotropic interventions in the context of problem behaviour crisis management. This scale consists of 10 items with three subscales: (1) pharmacological interventions, (2) behavior supports, and (3) protective equipment. This scale also emphasizes the distinction of intensity and intrusiveness as the main measure. Each of the items are scored with a 5-point rating scale *0 = less intrusiveness and low intensity,* while a score of 5= *more intrusive and generally higher* *intensity*. Although Zarcone et al. (2016) found that this scale had good face validity, the behavioural anchors and objective reporting for the behaviour specifically is not addressed. That is, the scale inquiries about the methods of the intervention (e.g., if the intervention features restraints), rather than the behaviour itself. Additionally, the reliability of this scale was only evaluated by calculating percentages of agreement across two raters. Thus, agreement was likely to produce reliability above zero since the sample size of raters were small (*N*=2) and both raters were of similar training (i.e., both master level raters) (Hallgren, 2012). Finally, the scale was not developed to evaluate the client or the severity of problem behaviour specifically, instead it evaluates the intervention. Although they do not aim to define severe problem behaviour specifically, they do discuss the relation of their responses which may attribute to severity (Zarcone et al., 2016). For example, they discuss that a treatment plan with lower staffing ratios may be a good behavioural anchor of considering a behaviour less severe.

Rojahn et al. (2012) developed a similar 32-item behaviour rating scale, the BPI, to evaluate the occurrence of problem behaviour. This scale was primarily developed for individuals with IDD 2 years of age or older, and consists of three subscales: SIB, stereotyped behaviour, and aggressive/destructive behaviour. The BPI alludes to the importance of objective reporting with the inclusion of a five-point rating scale (*0= never* to *5=hourly*) for each item. This scale also includes a four-point severity scale of *0=not a problem* to *3=a severe problem*. Some researchers have evaluated the effectiveness of specific subsections such as SIB and found acceptable reliability and validity (Sturmey et al., 1995). Drawbacks to this tool are the reliance on caregiver report. Although authors use some quantitative aspects such as the rate of occurrence, they do not require raters to collect data. The tool also focuses on the rate of problem behaviour and does not address the wide range of dimensions that may contribute to problem behaviour severity (e.g., criminal charges, housing restrictions, restraints). Third, this tool was not developed to classify severity given they did not establish a reliable or validated cut-off. Kim et al. (2018) attempted to address the BPI's shortcomings by completing the Child Behavior Checklist in addition the BPI. Kim et al. (2018) argued that these scales independently are unable to adequately measure severity of problem behaviour because they rely too heavily on the frequency of the problem behaviour rather than more diverse aspects. The authors suggest the use of data collection to improve accuracy of severity rating.

A fourth established scale for rating problem behaviour is the Clinical Global Impression (CGI) scale designed for children, adolescence and adults to quantify and track treatment progress in clinical settings (Busner & Targum, 2007). The CGI scale features two subscales of severity and improvement. The severity subscale utilizes a seven-point Likert scale (*1= normal, not at all ill, 2= borderline mentally ill, 3= mildly ill, 4= moderately ill, 5= markedly ill, 6= severely ill,* and *7= among the most extremely ill patients*) to measure the individual's responding. Coincidentally three of the four scales, ABC, BPI, and CGI, include the term severe in their rating scales but neglect to objectively define the term or assign a behavioural anchor for rationale of scoring. Additionally, they all encourage direct observation before or during the completion of the scale, while no literature reports whether raters are trained in data collection (Aman, 2013; Rojahn et al., 2012). Lastly, they all utilize Likert scale responding which can confound a tool's dimension of content and intensity while increasing measurement error (Hodge & Gillespie, 2003).

Various tools have been developed to quantify and understand the extent of physical trauma and injury as a result of problem behaviour. The Abbreviated Injury Scale (Committee on the Medical Aspects of Automotive Safety, 1971), and the Injury Severity Score (Baker et al., 1974; Barancik & Chatterjee, 1981) attempted to determine cut-offs for severity through outcomes of trauma an individual experienced by referencing mortality rates. However, these scales do not address severity from a multidimensional lens; specifically, the behaviour's impact on treatment outcomes, existing social restrictions, or legal history. Therefore, its application in clinical practice may be low (Palmer et al., 2016). Alternatively, these scales can offer rationale in determining cut-offs to aid in designing tools.

Risk is commonly associated with determining severity. As described above, Ricciardi and Rothschild (2017) designed a behavioural risk assessment they also refer to as a screening tool for behaviours of concern to objectively identify the risk of problem behaviour for individuals with IDD. The tool features problem behaviour components such as topography, intensity, settings, and health complications as a result of problem behaviour. They feature yes and no questions rather than the typical Likert scale responding and incorporate observational data. This scale identifies whether the behaviour has occurred or not but fails to prompt raters to report on specific problem behaviour characteristics such as the extent of injury to inform problem behaviour intensity. For example, they feature questions associated with risk to the caregiver or individual as rationale to establish problem behaviour severity but do not note the specific extent of injury. They also feature an ordinal scale for respondents to rate whether the behaviour occurs presently, in the past, never before, or no sufficient information exists. This could result in overestimating the occurrence or topographies of behaviour because these classifications do not include quantitative behavioural anchors or specific times to keep their responding consistent. Therefore, this scale cannot facilitate classifying severity. Finally, the scale's reliability and validity has not been evaluated.

### *The Influence of Rater Characteristics on Reliability*

Rater consistency in assessment tools have been noted as an issue in clinical settings (Chen et al., 2016). In fact, rater characteristics such as the degree of experience in content related to the assessment have been shown to impact the reliability of results (Chen et al., 2016; Zarkada & Regan, 2018). Chen et al. (2016) compared assessment reliability between expert and novice rater groups and found significantly higher inter-rater reliability within the expert rater group. Specifically, researchers have reported that raters with more experience tend to give lower and more critical scores (Gulgin & Hoogenboom, 2014; Villalobos et al., 2014). Some have attributed this to experienced individuals' tendency to pay attention to a wider scope of aspects (Villalobos, et al., 2014). Differences in collecting information to inform assessments can further decrease reliability (Chen et al., 2016).

Randsborg and Sivertsen (2012) evaluated reliability in three groups of doctors: (1) junior registrars, (2) senior registrars and (3) orthopedic consultants. They found that clinicians with more experience achieved higher reliability scores and more consistency. A triaging tool for spinal cord cases was evaluated in hospital settings and found similar results of experts compared to non-experts having more reliable scores (Lwu et al., 2010). Similarly, when Dracobly et al. (2017) evaluated the reliability of caregivers versus experts' completion of an indirect assessment regarding problem behaviour; the Functional Assessment Screening Tool (FAST), they found that experts had much stronger reliability and accuracy for determining the correct function of the problem behaviour. Further confirming that experience of an assessor may strongly impact the results.

Incorporating objectivity within a scale may decrease these sources of variability. Notably, the medical field is where most of the research has been conducted assessing tools while also evaluating and identifying rater inconsistency as a product of rater experience. To my knowledge, the only behavioural literature to assess rater consistency used an indirect tool compared experts to caregivers (Dracobly et al., 2017). Further, it seems as though no behavioural research has evaluated the effects of rater experience on reliability of a(n) direct/objective behaviour analytic scoring tool. Developing measurement systems for clinical use should prioritize assessing reliability (Koo & Li, 2016). Although reliability does not equate validity, a valid tool must first demonstrate reliability (Tavakol & Dennick, 2011). Given the field of behaviour analysis is largely comprised of certified clinicians with less than 5 years of experience (Deochand & Fuqua, 2016) this may be especially relevant.

Interestingly, not all of the *established* tools for measuring problem behaviour mention statistically evaluating internal consistency. Cronbach's alpha is a measure of internal consistency and a generally accepted test for evaluating the contribution of each item in a tool (Tavakol & Dennick, 2011). Kim et al. (2018) was the one of the few authors to employ Cronbach's alpha to evaluate a novel tool assessing problem behaviours in individuals with ASD. Generating alpha coefficients is considered an acceptable way to understand how each item in a tool contributes to the overall reliability of the tool and informs which items may need to be re-evaluated or removed in future iterations.

**Rationale for a Behaviourally Anchored Tool**

The term severe should be reserved for only the most high-risk, complex cases. Problem behaviours are comprised of many facets individually contributing to the behaviour's impact or quality of life interference. Thus, as previously mentioned an operational definition alone would not suffice for classifying severity because it cannot communicate all facets of interference or generate a means to standardize comparison across cases (Donenberg & Baker, 1993; Lach et al., 2009; McIntyre et al., 2002). The absence of a behaviourally anchored tool may also interfere with researchers' and clinicians' capacity to make data driven decisions, such as using evidence to triage individuals and/or justification for applying specific intervention strategies. This gap in the literature may in turn devalue the use of the term severe as a problem behavior descriptor, or even mislead clinicians reviewing the literature in an effort to inform their intervention practices. For example, when no objective, standardized way to classify severe aggression exists, a clinician may rely on a research-base suggesting a specific course of action that is ill-suited for the issues their client is currently presenting with. The consistent application of this term in the literature may help clinicians find and employ relevant, evidence-based techniques that align with clientele who present with similar participant profiles as those presented in research.

### *Clinical Applications*

Another potential benefit of developing an objective severity tool may be exemplified by Sival et al. (2000). The authors had the head nurse of each ward complete an objective behavioural rating tool for aggression (the Social Dysfunction and Aggression Scale; SDAS) weekly for one month. Medication was administered at the discretion of the head nurse in relation to the behavioural occurrences. Following the regular assessment of patient's target behaviour, the researchers observed an overall decrease in prescribed psychotropic medications in 63 patients with problem behaviour. They reported that although participant's score on the SDAS did not change, prescriptions for psychotropic medications significantly decreased from 0.8 to 0.6. Authors concluded that changes in the nurses' management of the behaviours when regularly using the rating tool may have led to a decrease in medicating patients. Silva et al. (2000) demonstrated the potential added value of a small and specialized scale which highlights the possibility of greater impact of a more generally applicable severity tool.

### *Improving Triaging*

Reliably classifying problem behaviour may pave the way to better triaging systems for admission to treatment beds that are in high demand. Additionally, this may justify the use of more intrusive procedures at the outset of programming. For example, high-risk, dangerous behaviours (like those described in Foxx, 2003) may require restrictive (e.g., mechanical restraints), or aversive (punishment) protocols at the outset of treatment to gain sufficient control over the problem behaviour. Similarly, individuals exhibiting severe problem behaviour are recommended to have more hours allotted for behavioural treatment (Busch et al., 2019). An empirically evaluated, and generally agreed upon cut-off, afforded by a severity tool informed by direct observation methods may support clinicians in expediently gaining permissions to apply more restrictive measures rather than prolonging permissions which may place individuals and caregivers at greater risk.

Finally, individuals with severe problem behaviours are more likely to be prescribed a higher dosage of psychotropic medications for behavioural sedation (Deb et al., 2015; Evers & Pilling, 2012). Common treatments for problem behaviour in persons with IDD are either behaviour analytic, psychopharmacological or the concurrent application of both (Baker, 2017; Deb et al., 2007; Deb et al., 2015). Although behavioural interventions have been shown to be successful and cost effective for problem behaviour (Sturmey & Didden, 2014), Deb et al. (2007) reported that between 20%-47% of individuals with IDD were prescribed psychotropic medication, while 14%-30% of those individuals were receiving this medication for reduction in problem behaviour. Despite subjecting severe problem behaviour cases to the most intrusive and restrictive procedures (Deb et al., 2015; Hanratty et al., 2015), only vague and subjective descriptions classifying problem behaviour severity exist (Emerson, 2001; Lowe et al., 2007).

Although problem behavior severity should not be the sole rationale for psychopharmacological treatment, it may be reasonable to use a fully developed severity tool informed by direct measurement as one element determining eligibility for polypharmacy. The continued inconsistent use of this term establishes a risk of under, or overestimating participant problem behaviour (Murphy et al., 2009).

### *Improving Analysis and Replication in Research*

Along with the gap for a reliable measure in practice, researchers also have an obligation to apply this term objectively to adequately describe participant problem behaviours, as well as reliably reporting severe problem behavior prevalence promoting research replication. In addition, a tool will eliminate the need for researchers to rely on different methods, as well as on caregiver report alone, to classify severity when conducting severe problem behaviour reviews or meta-analysis (Cox et al., 2020; Fahmie & Iwata, 2011; Fritz et al., 2013; Poppes et al., 2010).

Some authors have claimed a problem behaviour is severe based on author consideration alone. For example, the researcher would consider a participant's problem behaviour severe without describing their behaviour or identifying any dimension of the behaviour (e.g., frequency or physical trauma) to justify the severe label (Berg et al., 2016). Authors have even published articles claiming severe problem behaviour in the title while making no mention of the word in the target behaviour descriptions (Hausman et al., 2009). Applied behavior analytic research needs to embrace the consistent use of terminology describing participant profiles. Continually neglecting to do so may undermine bridging the gap between research and practice.

## Study Purpose and Research Objectives

The term severe is used inconsistently across behavioural literature (Lowe et al., 2007). Many researchers have highlighted the need for an objective tool to classify problem behaviour for both research and clinical purposes (Evers & Pilling, 2012; Lowe et al., 2007). Therefore, the purpose of this research project was to explore the utility of an objective severity tool. Specifically, I explored:

## Research Objectives

1. Which tool items meaningfully contributed to Cronbach's alpha coefficient?
2. If the severity tool impacted interrater reliability across novice and expert clinician raters?
3. If the severity tool impacted interrater reliability within novice and expert clinician rater groups?
4. If raters reported the tool as acceptable for clinical and research settings?

## Hypothesis

1. I expected all items to contribute well to Cronbach's alpha with values between .75 to .90.
2. I expected severity tool access would improve interrater reliability across novice and expert clinician raters.
3. I expected severity tool access would improve interrater reliability within novice and expert clinician raters.
4. I expected participants would rate the tool as moderate to highly acceptable and recognize the added value it could have in relation to its potential uses.

# Method

## Participants and Recruitment

I recruited nine BCBAs. There were five participants in the novice group, and four participants in the expert group. Participant inclusion criterion was a BCBA, or Board Certified

Behavior Analysts – Doctorate (BCBA-D) certification in good standing. Participants were asked to provide their BCBA certificant number, which I checked against the Behavior Analyst Certification Board Inc. (BACB) database. Participants with an 'inactive' status were not eligible to participate. Other certifications through the BACB such as Registered Behavioral Technicians (RBT) and Board Certified assistant Behavior Analysts (BCaBA) were excluded. Participant age and gender was irrelevant and therefore not documented.

I recruited participants by (1) circulating e-mail invitations distributed via professional behaviour analytic groups (e.g., Ontario Association for Behaviour Analysis, Association for Behaviour Analysis International), (2) distributing flyers at dissemination events, and (3) posting electronic flyers through professional development activities such as annual conferences (e.g., Ontario Association for Behaviour Analysis Annual Convention, November 2019; Association for Behavior Analysis International, May 2020).

## Materials

All communication between the researcher and participants were through email and all study materials were distributed digitally, accessible via a computer or mobile device.

I used two sets of 20 equally complex case scenarios (40 total case scenarios) (see Appendix A for select examples). To enable randomization, I developed 60 case scenarios (30 scenarios with unspecified severity and 30 with behaviours coined as severe). I used an online number generator to randomly assign 40 scenarios that would be used in the study. To ensure scenarios were equal in complexity, all scenarios were informed by published behaviour analytic problem behaviour literature with a purpose of decreasing behavioural occurrences. That is, the development of scenarios relied primarily on the participant profiles provided in the studies; often including age, target behaviour, communication abilities, previous interventions, and living conditions. Infrequently, some hypothetical details were added to ensure equality across both sets and enough data to complete the tool (e.g., adding baseline responding). Participants were instructed to use the information from the scenarios as if the descriptions were all informed by direct observation or permanent products.

Each participant evaluated 20 scenarios during one condition (no tool access), then evaluated 20 novel scenarios during the second condition (tool access). I randomized the order of case appearance within both conditions. Both sets of scenarios had the same ratio of case severity. That is, 50% referenced from articles describing the problem behaviour as severe and 50% referenced from articles wherein authors did not specify problem behaviour severity. Scenarios were also brief in length with paragraphs ranging from 200 to 300 words.

## Measurement

I created two surveys (i.e., demographic survey and social validity questionnaire) and one measurement tool (i.e., severity tool). Participants completed these through a widely used online survey application, Qualtrics. Participants were sent an email with a direct link to the survey and responses were automatically uploaded upon completion.

### *Demographic Survey*

The brief demographic survey (see Appendix B) consisted of seven multiple choice and one short answer question about participants' clinical experience. The survey featured questions about the participant's certification date, employment history, clinical settings, years of experience working with persons with disabilities, and the client population served. Participant survey responses informed group membership (i.e., expert or novice). For example, a BCBA demonstrating a wide range of experience with different client ages, diagnosis, and settings could be considered an expert rater while a recently certified BCBA with limited experience in these areas may have been considered a novice rater. These group assignments did not reflect the BCBAs capacity as a practitioner, rather it served as an indicator of experience with problem behaviour cases.

### *A Proposed Severity Tool*

I created the current iteration of the tool by consulting relevant literature, which informed the inclusion of necessary domains and items to address the many facets that may impact an individual's problem behaviour severity. The severity tool (see Appendix C) is a weighted questionnaire comprised of six multiple choice and 20 yes or no questions (26 items total). Only one response could be selected for each question. This method of close-ended responding was required because of the tool's reliance on permanent products and direct observation (e.g., data collection, past or current tissue damage, billing, or documentation). Therefore, the tool represented an objective measure of problem behaviour, and thereby should avoid ambiguous questions or subjectivity in responses. For example, an individual either does or does not have 2:1 caregiver ratio. Similarly, assault charges have or have not been laid against an individual (pending or otherwise).

#### **Subscales**

The questionnaire had four problem behavior domains. *Frequency*, *chronicity*, *intensity*, *and legal and environmental restrictions* (environmental restrictions, criminal justice system involvement, etc.). These subsections aimed to evaluate the dimensions, effects, and restrictions imposed because of an individuals' problem behaviour. The frequency domain relied on user responses informed by a systematic data collection process (e.g., daily data collection). Although Aman (2013) recommended observable data collection such as frequency of the problem behaviour as a vital step before and during the completion of behavioural tools, their tool itself did not rely on direct observation to be completed. Chronicity pertained to the effectiveness of past interventions and relied on treatment history documentation. Taylor et al. (2011) discusses chronicity being most often found in well-established and severe behaviour cases. Although these authors were speaking to the topography of SIB, this characteristic may help inform any problem behaviour topographies. This is because it focused on failed treatment plans which could indicate persistent and well-established behaviour. Frequency and chronicity did not contain any subsections.

There were three subsections in the intensity domains, including: functional analysis, physical damage, and property damage. Previous literature used components featured in the tool's items to justify severity of behaviour such as Foxx (2003). These authors referenced behavioural outcomes such as social isolation and body harm. Roscoe et al. (2013) also attempted to justify severity with tissue damage and other effects resulting from the behaviour's occurrence. Similarly, Emerson et al. (2001) based the severity of an individual's problem behaviour by the degree of property damage. These items inquired about the physical trauma resulting from problem behaviour occurrences and risk to the individual or caregivers. Responding across these items would require documentation of the individual's history. For example, past treatment data or reports would note whether a behaviour reduction intervention was successful or unsuccessful, receipts for property damage would inform the value of damages, and medical or police reports would note the impact of behavioural incidents. The case scenarios used mentioned any applicable documentation for participants to address these items.

Finally, there were three subsections within legal and environmental restrictions: (1) legal, (2) residence restrictions, and (3) community and resource access. A history with the judicial system and environmental restrictions have been a commonly mentioned characteristics of severe problem behaviour. Evers and Pilling (2012) discussed that access to services and community facilities are often restricted to individuals with more severe problem behaviour. These authors also discussed higher severity in cases involving the judicial system. Restriction in living accommodations such as locked inpatient settings have also been noted as characteristic of more severe problem behaviour (Pilling et al., 2015). The residence/environmental restrictions items require respondents to consider existing specialized housing requirements (e.g., plexi-glass windows, locked half-doors) or intentional exclusion from the community as a result of problem behaviour and select a response.

### Weighting

I assigned item weights in accordance with the behaviour analytic literature, and other relevant sources (see Appendix D) (e.g., Emerson et al., 2001; Fahmie & Iwata, 2011; Rojahn et al., 2012). For example, Lowe et al. (2007) considered problem behaviour more severe when multiple caregivers were needed to supervise the individual. Rojahn et al.'s (2012) description of mild, moderate and severe destructive behaviour also helped to determine the weighting distribution of items in the property damage subsection. Blenkush and O'Neill (2020) consider a problem behaviour severe if it interferes with social development, skill acquisition, and education. These authors also mentioned that behaviours resulting in placements within long-term residential or hospital-based settings were also likely severe behaviours.

The maximum sum of the severity tool rating is 40. The tool must be completed in its entirety, and there is no option to skip an item. If a question does not apply to the target individual, the behaviour topography, or no explicit answer was provided in the scenario - the rater should default to a response of *no*. This would equate to a score of 0 for that item. Notably, Rojahn et al. (2012) recommended defaulting to 0 as the best method for any nonapplicable or missing data in the established BPI scale.

### *Social Validity*

A brief social validity questionnaire (see Appendix E) included five close-ended Likert-scale questions. Participants responded from 1 – strongly agree to 5 – strongly disagree. The purpose of this questionnaire was to gain insight on the application of the tool as well as the clarity of items. The final question was open-ended to provide an opportunity for participants to suggest ideas on how to refine or improve the tool.

## Procedures

### Determining Sample Size

To determine an appropriate sample size for a modified pre-post group design, I reviewed the literature and conducted a power analysis. Of note, generalization was sacrificed due to the nonrandomized rater feature. I conducted a power analysis using a statistical application called g*power. I assigned a significance level of 0.05, power of .80, with a correlation value of .90. The outcome suggested this study's design required approximately four participants per group to conduct a meaningful analysis.

### Demographic Survey

Figure 1 depicts the study procedure. When participants returned a completed consent form (see Appendix F), I distributed the demographic survey through an online questionnaire. Once participants, henceforth referred to as *raters*, completed this survey, I assigned them to one of the two groups (*novice* or *expert*). An expert rater must self-report a minimum of 5 years regularly working experience with severe problem behaviour (item 2) (see Appendix A) and have been certified for 5 years or more (Behavior Analyst Certification Board, 2017; Chen et al., 2016; Villalobos, et al., 2014). They must also report experience with more than one age group (e.g., under 6 years old, and between 6 to 12 years old), or one age group over the age of 12 years (item 4; Appendix A). Lastly, they must report experience with clients across two or more settings as a BCBA (Item 5; Appendix A). Raters who did not meet the criteria of an expert were automatically assigned to the novice group.

## Rating Case Scenarios

I distributed the first group of 20 scenarios (see Appendix A) through a new (second) link. Raters were asked to review each case and rate scenario severity (i.e., *severe* or *not severe)* using only their clinical judgement to guide responses. After raters completed the first 20 scenarios, they received a third link to the second group of 20 scenarios. *Rater*s were invited to complete the severity tool (see Appendix C) for each scenario; after completing the entire tool for each scenario they were immediately prompted with the question "Is this case severe or not severe?" and expected to select their response. Each of the cases were displayed on the same page as the questions for each subsection of the tool and appeared once again when asked to rate the severity. The rater was not provided with the final score of the severity tool for any scenarios. The surveys were designed with a three-week time limit after starting and no option to leave items unanswered.

### *Socially Validity Questionnaire*

Once raters had completed their second group of case scenarios, they received the fourth and final electronic link to the social validity measure (Appendix E).

# Results

Fulfilling the current study's purpose required multiple components of analyses, including: (1) Mann-Whitney, (2) Cronbach's alpha, and (3) four intraclass correlations. Statistical package for social sciences (SPSS) was used to conduct each analysis.

## Demographic Survey

Results of the demographic survey (see Table 1) were used to assign *(N*=9) rater membership to either the expert or novice group. Expert raters (*n*=4) were certified for a minimum of 5 years (*M*= 108.25 months (9.02 years), range= 78 to 156 months (6.5 – 13 years), *SD*=33.96). They also self-reported a minimum of 5 years' experience (range= 5 – 18 years) regularly working with severe problem behaviour (*n*=4) (item 2, item 3, item 4; Appendix B). Experts reported experience with more than 20 clients (*n*=4) and an average of 2.2 different age groups (item 6; Appendix B). Finally, experts reported experience working across an average of 3.5 settings, and with clients across two or more different settings as a BCBA (item 5, item 7; Appendix B). Novice raters (*n*=5) were certified for an average of 19.60 months (1.63 years), ranging from 1 to 43 months (0.08 – 3.58 years) (*SD*= 19.23), and self-reported less than 5 years' experience supporting individuals with severe problem behaviour. Novice raters had a group mean of 4.2 age groups treated and 4.25 work settings. All raters (*N*=9) reported some experience working with severe problem behaviour. None of the raters reported working in a hospital setting or with age groups below 12 months old.

### *Mann-Whitney Test*

The Mann-Whitney test was conducted across select items featured in the demographic survey to identify whether differences between group demographics existed. Short answer questions were excluded from this analysis.

The Mann-Whitney is a nonparametric test statistic that explores whether two groups differ from one another across specified variables (Field, 2018). In sum, the statistic ranks the scores of each variable within each group from lowest to highest then adds up the sum of those ranks for each group and compares these sums. The Mann-Whitney is also noted for its ability to compare two groups of unequal size (Bürkner et al., 2016). If there is a significant difference (i.e., critical *p*-value lower than .05) between the two groups' medians, it would suggest the samples could potentially represent different entities (Field, 2018). In this case, arguably, two different clinician samples. Notably, the minimum sample size to discern whether a significant difference exists between two groups is eight (Fay & Proschan, 2010). Thus, a large *U* value may not be atypical or unexpected because this study's sample size exceeds the minimum by only one (Nachar, 2008). Of note, two survey items featured multiple selections; (1) age groups treated, and (2) current or previous work settings. Therefore, I calculated the sum of selections for these items before running the analysis.

Results of the Mann-Whitney test using data from the demographic survey are displayed in Table 2. Results suggest that expert and novice groups were significantly different in the number of years certified, and years of experience supporting individuals with severe problem behaviour. The group medians did not appear to differ significantly across the following four demographics, (1) number of clients they supported post-certification, (2) regularity of working with clients who engaged in severe problem behaviour, (3) number of age groups treated, and (4) number of current or previous work settings.

## Cronbach's Alpha

Cronbach's alpha is a measure of internal consistency (i.e., error of the tool). The coefficient is expressed as a number between 0 and 1. Acceptable values range from .75 to .95

(Field, 2018; Tavakol & Dennick, 2011). A small alpha coefficient (under .70) may indicate too few items in the tool, while a large alpha coefficient (above .90) may indicate that some items are redundant and measure the same dimension (Tavakol & Dennick, 2011). Statistical package for social sciences also generates a coefficient that determines the item-total correlation for each tool item. This coefficient can help inform researchers of items that may not be applicable and therefore, should be removed in future iterations. Items with a correlation of less than .30 may indicate that the item should be removed. Finally, SPSS generates a revised Cronbach's alpha for each item if that item is deleted from the analysis (i.e., the tool). The item-total correlation informs researchers of the contribution of each item to the overall alpha. An outcome wherein an item, when removed, produces a large increase in the overall alpha may suggest that removing the item would increase the tool's internal consistency (i.e., increases the overall alpha of the tool) (Field, 2018).

Cronbach's alpha for all items of the tool (*N*=26) (Cronbach ∝ = .831) suggested good internal consistency (*M*=12.97, *SD*=7.385). The alpha was also calculated separately for each of the four subscales, frequency (∝=.447), chronicity (∝=.235), intensity (∝=.771), legal and environmental restrictions (∝=.770). Inspecting the corrected item-total correlation for each item revealed several questions generated a coefficient below .30, including: question 1, 3a, 9, 10, 12, and 14 (see Table 3). However, deleting these items did not result in a meaningful increase in the alpha, which may justify retaining them in the tool. Table 3 reports the corrected item-total correlation and Cronbach's alpha if deleted for each item.

## Reliability

The intraclass correlation coefficient (ICC) is a widely used reliability statistic for dichotomous, ordinal, interval, and ratio variables with more than two raters (Hallgren, 2012;

Portney & Watkins, 2008). Intraclass correlation identifies the extent rater's results of a measure can be replicated by identifying the degree of correlation between raters (Koo & Li, 2016). Although several reliability statistics exist, this statistic provides both the degree of correspondence and agreement between two or more raters (Koo & Li, 2016; Portney & Watkins, 2008). Another relevant statistical feature of this method is that groups do not need to be equal (Portney & Watkins, 2008). This statistic yields a reliability value between 0 and 1; with zero representing chance reliability and one representing perfect reliability. Although there is no standard for acceptable reliability values, it is generally agreed that an ICC value above .90 can indicate excellent reliability (Field, 2018; Koo & Li, 2016).

This study employed the two-way model (commonly referred to as Model 3; Portney & Watkins, 2000) because all raters experienced both conditions. The mixed model accounted for the specific selection of raters in a nonrandomized order. I calculated four ICC values, (1) novice rater no tool, (2) novice rater tool, (3) expert rater no tool, and (4) expert rater tool (see Table 4). There was no formal statistical test of differences comparing the ICC values across no tool and tool conditions within groups. However, Cumming and Finch (2005) describe a visual analysis of confidence intervals (CI) (also called 'inference by eye') that can be helpful in facilitating a fulsome results analysis in the absence of running formal statistical tests. Intraclass correlation values and their corresponding CIs are illustrated by Figure 2, complete with error bars depicting CIs, supporting the results analysis and related discussion content.

The coefficient for novice raters in the no tool condition was *r*=.781, 95% CI [.582, .903]. This value was lower than the coefficient for novice raters in the tool condition (*r*=.860), 95% CI [.733, .938]. Similarly, the coefficient of expert raters in the no tool condition (*r*=.803), 95% CI [.610, .913] was lower than the coefficient of expert raters in the tool condition (*r*=.912), 95% CI

[.825, .961]. In reviewing these coefficients, the value associated with novice raters in the no tool condition (*r*=.781), 95% CI [.582, .903] was lower than the coefficient of expert raters in the no tool condition (*r*=.803) 95% CI [610, .913]. Although the tool condition was associated with higher coefficients for both groups, there appeared to be a larger increase across the no tool and tool conditions for the expert raters (*r*=.803 to *r*=.912), 95% CI [610, .913], [.825, .961] respectively; compared to novice raters (*r*=.781 to *r*=.860) 95% CI [.582, .903], [.733, .938]. Figure 2 provides a visual of this difference. With regards to CIs, the range was markedly smaller in the tool access condition for both groups (novice 95% CI [.733, .938]; experts 95% CI [.825, .961]) compared to the no tool condition for novice and expert groups (95% CI [.582, .903], 95% CI [.610, .913] respectively).

## Social Validity

Results of the social validity questionnaire are depicted in Figure 3 (see legend on Table 5). Mean and median values for each rater group are displayed in Table 6. Five of the nine raters *agreed* that the tool was easy to use. When comparing group responses, expert rater responses reflected better social validity compared to novice group responses. The final question was open ended and asked raters if there was some case scenario characteristics missing which may have improved the tool. Raters suggested adding items in the tool to address, (1) the timeline of problem behaviour, and (2) more information about previous interventions that were implemented (if applicable).

# Discussion

The results appear to support the value that my severity tool may have in terms of classifying severe problem behaviour. However, given this is the first iteration, I interpreted my results with caution while generating discussion content and tentative conclusions. I will discuss the results of my study in the order corresponding to my hypotheses. I will begin with discussing the internal consistency of the tool, followed by reliability, group membership outcomes, and finally, social validity outcomes.

## Cronbach's Alpha

Overall, the alpha coefficient indicated good internal consistency. A fulsome discussion of the tool's internal consistency requires consideration of the result in relation to, (1) the sample size, (2) the consistency of Cronbach's alpha across the subscales and (3) the complexity of the term severe.

First, the number of raters may have impacted the calculation of Cronbach's alpha (Bujang et al., 2018). Namely, Bonett's formula, commonly used to determine the ideal rater sample size required to produce an accurate alpha, indicated 12 as the optimal number of total raters. This calculation was conducted post hoc because I had prioritized adhering to an optimal sample size for the ICC analysis; given evaluating reliability was the main purpose of my study. Despite the fact that the sample size for the alpha statistic was less than optimal, the tool still produced an acceptable alpha coefficient. It may, therefore, be reasonable to conclude that the items in this tool may contribute meaningfully. Thus, the tool may hold promise for the classification of severe problem behavior. I would suggest future researchers consider recruiting a larger sample size to achieve optimal power.

### *Alphas by Subscale*

Although I observed an acceptable alpha, it is possible the outcomes were an artifact of the larger number of items in the tool (26). That is, a large number of items in a tool can increase the likelihood of generating an acceptable alpha (Field, 2018). To explore this consideration, I calculated Cronbach's alpha for each of the subscales. The subscales *Intensity* and *Legal and* *environmental restrictions*, which were comprised of 11 items each, revealed moderately lower values (∝=.771 and ∝=.770 respectively) than the overall alpha. This could suggest that the tool's internal consistency was not an artifact of the number of items. That is, a moderate alpha was generated for these two subscales, even though they were comprised of less than half of the tool items (i.e., 11 items vs. 26 items).

Of note, two of the four subscales, comprised of only two items each, (*Frequency* and *Chronicity*) releveled noticeably lower alpha values (∝=.447 and ∝.235 respectively). Many authors have described problems with evaluating the alpha of a two-item tool, specifically that it may underestimate (sometimes substantially) internal consistency. This could lead to misinterpretations (Eisinga et al., 2012; Sijtsma, 2009). Therefore, it may have been inappropriate to analyze and report two-item scale with this statistic. However, I did this because it is common to report subscale alpha values when evaluating first iterations of tools (Aman & Singh, 1985; Rojahn et al., 2001). Further, the well-established ABC initially reported subscales with alpha values ranging from *r*= –.13 to *r*=.52 (Aman & Singh 1985). They opted to use a Pearson correlation for their analysis and have since demonstrated values that are much higher than those obtained in their original assessment of the ABC. This could suggest that my tool could be very promising; considering it, arguably, outperformed the first iteration of the ABC – which went on to be a well-established, commonly used problem behaviour rating tool.

Analyzing the tool by generating alpha values for each subscale may have produced outcomes that corroborate the idea of severity as a complex construct comprised of multiple components (Blenkush, & O'Neill, 2020; Emerson, 2001; Lowe et al., 2007). This alleged severity complexity may be another reason the tool's value should be judged based on overall alpha, rather than alpha values obtained by its individual components (i.e., subscales). Further, a good quality tool may be characterized by its capacity to measure different dimensions of the same construct (Field, 2018). So, the overall alpha may indicate that the tool is measuring many separate, but complimentary, dimensions. In short, I may tentatively conclude that the overall alpha suggests all tool items are valuable and relevant for their collective purpose (i.e., classifying severity).

Finally, it is important to reiterate that the weighted items were informed by the literature (see Appendix E). I did not explore or analyze the weighting specifically. Given the main purpose of this paper was to evaluate reliability, it may be helpful for future research to explore the weighting of the tool.

**Reliability**

There were several interesting outcomes. First, the reliability coefficients observed in the tool condition of the current study exceeded those of established scales such as the ABC (*r*=.63) (Karabekiroglu & Aman, 2009), BPI (*r*=.74) (Rojahn et al., 2013), and the Repetitive Behavior Scale – Revised (RBS-R) (*r*=.67) (Lam & Aman., 2016). These authors conducted a comparison on the ICC values of the severity components of these scales. The results showed reliability values varied from *r*=.65 to *r*=.80. The current severity tool may have facilitated rater reliability (across both groups) that exceeded those observed across other scales. This could suggest that it could offer a relatively reliable classification of severity. Of note, the raters featured in the studies listed above (Karabekiroglu & Aman, 2009; Lam & Aman., 2016; Rojahn et al., 2013) typically used a wider range of clinician raters (without a specific certification or education level) which may have been partially responsible for producing more modest reliability values.

Second, I observed a difference between ICC values generated in the novice no tool condition compared to the expert no tool condition. It is possible that expert raters were slightly more reliable when reporting severity in the no tool condition. These results align with existing literature evaluating a competence assessment tool (Weck et al., 2011). That is, the authors found that expert rater responses were initially more reliable than novice raters before the tool. Danov and Symons (2008) also found that experts were considerably more reliable compared to the novice group when visually inspecting functional analysis graphs.

Third, I had expected access to the tool to minimize reliability discrepancy across groups. However, the difference in reliability values using visual analysis across the groups got minimally larger in the tool condition (no tool *r*=.02 versus tool access *r*=.05). Regardless, according to raw coefficients I observed a modest increase in reliability for both groups when raters were given access to the tool to classify severity. Without the tool, the novice group just met the required value for *good* reliability (.75 – .90; Liljequist et al., 2019), accompanied by a wide CI (see Figure 2). By contrast, with the tool, I observed a score that fell well within the *good* range (.86), accompanied by narrower CI which could suggest I can be more confident that the CI includes the ICC coefficient compared to the no tool condition (Cumming & Finch, 2005). Tool access for the expert group resulted in a categorical shift from good to *excellent* reliability (> .90; Liljequist et al., 2019), alongside a narrower CI (see Figure 2). These results align with previous research. Specifically, Carballo-Fazanes et al. (2021) evaluated an objective test to report fundamental motor skills and found expert rater reliability improved more so than novices. Similarly, O'Hara and Rehm (1983) evaluated the reliability of reporting depression using a rating scale. They found that the expert rater group was more reliable compared to novice undergraduate raters. It is possible that expert BCBAs may have had more experience using structured assessment tools (Jensen-Doss & Hawley, 2010), and this contributed to the larger increase in reliability across conditions.

Another possible explanation for the lower reliability of novice raters may have been that they needed more detail (e.g., examples, definitions) across items to fully benefit from tool access. Specifically, items 2, 5, 5a, 6, 6b, 12, and 18 included examples, while all other items on the tool did not. Although, it is common among other similar tools and scales (e.g., ABC) to include items with and without examples, most of the items in my tool with examples were associated with higher item-total correlations ranging from *r*=.286 to *r*=.851. Thus, future research should investigate the effects on reliability when examples are included across all items of the tool. Doing so may generate further improvement in reliability within novice raters such that final reliability values align more closely with expert raters' reliability.

### Discussing Confidence Intervals Outcomes

No tool and tool access conditions for the expert group showed a commonly accepted standard of more than 40% overlap across error bars (see Figure 2) (Cumming & Finch, 2005). Therefore, further visual analysis was used to compare groups and conditions given statistical models to compare ICC are not readily available. Visual analysis of Figure 2 depict some similarities across groups (novice and expert) and conditions (no tool and tool access). That is, the narrower CIs observed in the tool condition could suggest that members of each group were responding more similarly compared to responding in the no tool condition. The upper and lower CI values may also be noteworthy. Specifically, the range of CIs for the novice raters (95% CI [.582, .903]) and expert raters' 95% CI [.610, .913] in the no tool condition indicated that the genuine ICC value could have fallen between the *moderate* to *good* category. By contrast, the upper and lower bounds of the CIs were all within the *good* reliability category for both groups (novice 95% CI [.733, .938]; experts 95% CI [.825, .961]). This could suggest that the difference in rater performance across groups was relatively minor, while also suggesting the improved performance within groups may have been genuine. It is important to note that these interpretations are tentative given narrower CIs could be partially explained by correlations between the paired scores. Future research should conduct more rigorous designs (i.e., control vs. treatment group; cross-over) to corroborate the patterns observed in the current project.

## Demographic Survey

My results are only as valid as the precision with which the demographic survey classified raters. That is, incorrectly assigning group membership would be extremely problematic in terms of the entire study. Thus, a fulsome discussion requires reflecting on relevant considerations related to the demographic survey and ultimately group assignment.

Currently, there is limited published literature addressing the criteria for classifying a behaviour analyst as an expert with severe problem behaviour. Therefore, I used content from The Behavior Analyst Certification Board (2017) and information from other related fields to develop the demographic survey (Chen et al., 2016; Villalobos, et al., 2014).

It is likely that I observed a significant difference in years' experience and years' certified because these were heavily emphasized (Chen et al., 2016; Villalobos, et al., 2014). The Behavior Analyst Certification Board (2017) suggests that newly certified BCBAs should receive supervision from a BCBA with a minimum of 5 years post certification. This recommendation, combined with information from other related disciplines meant that rater responses to all featured questions informed group membership; however, responses to years of experience and years certified ultimately determined rater placement (see e.g., Hmelo-Silver & Pfeffer, 2004; Saini et al., 2017; Westerman, 1991).

***Possible Rationale for Non-Significant Differences.***

All raters self-reported having worked in more than two settings. Interestingly, when the medians of groups were compared, novice raters reported experience in more work settings than experts. However, the difference was not significant. One possible explanation may be the growing demand for behaviour analysts *across settings* (Behaviour Analyst Certification Board, 2020). That is, behavior analysts with more than 5 years' experience may have had fewer setting options when looking for work, compared to newer behavior analysts. Annual job postings for behaviour analysts have increased from 798 to 33, 996 over the past 10 years (from 2010-2020) (Behavior Analyst Certification Board, 2021). Namely, healthcare, educational services, social assistance, insurance carriers, and public administration are among the settings with a growth in demand for behaviour analysts (Burning Glass Technologies, 2015) (Industry classifications followed the North American Industry Classification System; NAICS) (Statistics Canada, 2017). Hiring practice literature on this topic suggests the demand for behaviour analyst positions have more than doubled between 2012 and 2014 (Burning Glass Technologies, 2015). Considering that the expert groups' median of years certified was 8.2 years (i.e., certified by 2012), while the novice median was 1.5 years (certified by 2019), it may not be surprising that novice raters, certified more recently, had opportunities to work in more diverse settings.

The self-report feature of the demographic survey may partially explain some of the non-significant outcomes. That is, items inquired about *general* experiences (i.e., Which of the following best describes your work experience working to decrease severe problem behaviour?) rather than *specific* clinical experiences (i.e., Which of the following best describes how frequently you support clients that engage in persistent, life threatening and dangerous problem behaviour?). Katowa-Mukwato et al. (2014) found that the correlation between medical students' actual experience and their *overall* self-perceived confidence was low (.55). By contrast, perceived confidence of specific medical procedures and actual experience with those medical procedures were noticeably higher (.82). Thus, it is possible that adjusting the survey questions to inquire about specific clinical experiences, as opposed to general experience with severe problem behaviour may help to improve the demographic survey. Future research should explore whether adjusting the items in this way produces more distinct (or significant) differences across groups on currently non-significant items.

To better understand rater's experience with severe problem behaviour I had to use the term severe in survey item 3. This may have introduced additional biases (see *Existing Tools and Subjective Reporting Limitations*). Biases associated with self-report (Donaldson & Grant-Vallone, 2002; Finney et al., 1998; Floyd et al., 2005; Hajiaghamohseni et al., 2020) may have negatively affected the likelihood of obtaining significant results on this item. One possible solution could be to provide several examples of severe cases alongside this item. Another possibility is to replace the term sever with other synonyms or commonly used terms to describe *hard to serve* cases (e.g., highly dangerous, life threating, excessive).

Given the goal of my severity tool is to address inconsistencies associated with the term severe, it is possible that the final tool (after several iterations) could help clinicians more accurately reflect on and report their own experience with severe problem behaviour. Specifically, future researchers may consider investigating whether raters' self-report changes before and after using the severity tool. If changes in self-report are observed this could suggest improved accuracy in self-identifying their scope. In practice, using this tool could possibly encourage clinicians who are not regularly working with *truly* severe cases to seek appropriate supervision when presented with severe client cases that exceeds their skill set.

### *Limitations of the Mann-Whitney Analysis*

The number of raters for this study minimally exceeded the required sample size (8) to complete a Mann-Whitney analysis. In addition, distributions were not normal (Fagerland & Sandvik, 2009; Nachar, 2008), therefore, the results should be interpreted with caution. It is possible that these limitations impacted the likelihood of identifying a significant difference between two groups. Alternatively, this could mean that any observed significant differences (e.g., years serving clients with severe problem behaviour) may indicate the groups are considerably different. That is, the significant outcomes may be quite meaningful.

Thorough analysis of the demographic survey went beyond the scope of the current study. However, future researchers may consider featuring a larger number of raters to explore the value of each demographic survey item, as well as the complete survey in its capacity to reliably classify expert versus novice clinicians. Given the field of behavior analysis is growing at a rapid pace to meet the demand for behaviour analysts, experience and mentorship is often overlooked in an effort address this deficit (LeBlanc, 2015). Research projects targeting the establishment of firmer criteria for distinguishing expert and novice clinicians may be important in retaining the integrity of the field. It may be especially helpful to identify opportunities for mentorship when a clinician is operating outside of their scope of practice.

## Social Validity

Overall raters scores suggest they felt the tool was a good measure to classify problem behaviour. Raters across both groups uniformly rated usability and applicability highest. The fact that expert raters generally appraised the tool more positively across all items compared to novice raters seems to align with existing literature. Specifically, Jensen-Doss and Howley (2010) measured clinician attitude ratings across three evidence-based assessment tools. They evaluated the ratings of doctoral-level clinicians and psychologists, master's-level clinicians, and non-psychologists. Authors found that doctoral-level raters expressed more positive ratings across all social validity categories. Bjaastad et al. (2019) also reported that more experience using standardized tools may lead to more positive attitudes towards similar tools. With regards to practicality, it is possible that experts rated the tool more positively on this item because they have more *lived experience* in relation to the variability that exists in reporting severity. Or experts may have more experience with comparable tools, which lead them to rate similar tools more positively (Bjaastad et al., 2019)

## Practical Applications

The novelty of this tool limits immediate application in the field, and more investigation regarding reliability and validity analyses are warranted. However, there are some areas wherein this tool may be most applicable. First, the tool could be used to improve triaging. That is, it may offer a way to insert objectivity to justify treatment allocation. Lwu et al. (2010) highlight that clinical stakeholders and other nonexpert staff may be responsible for allocating treatment funding and awarding treatment beds. Given these individuals may not understand the complexity or needs of each client, a more objective severity measure, such as the one proposed, could help clarify the expectations. This severity tool may offer a method of qualifying an individual's need or suitability for a program. A second application may be having clinicians complete the tool to better understand their scope of practise. That is, whether client problem behaviour severity warrants additional support or supervisory training. Brodhead et al. (2018) discusses that although clinicians may have similar coursework or designation (e.g., BCBA) their coursework does not guarantee competence in all course content. The authors go on to discuss that experience with different severities of problem behaviour may also be considered different scopes of practice. When a client case does exceed the scope of competence for the clinician, it is recommended that the clinician seek additional support or supervision to intervene effectively. This may provide quantitative evidence that a client's problem behaviour exceeds their scope and may serve as support for the clinician to recruit a supervisor with more expertise supporting individuals with severe problem behaviour. Finally, researchers may use this tool and reference the severity score in their published articles. This may help others to identify relevant participants for the replication of successful procedures. That is, participants with similar problem behaviour severity profiles.

## Limitations of a Pre-post Design

The current research design is relatively weak and vulnerable to several threats to internal validity, including testing, maturation, and history. Although these threats can make it difficult to establish a causal response, I felt common interval validity threats may not have unduly influenced the results for several reasons.

First, given the large number of scenarios raters were exposed to, testing effects could have confounded the results. That is, reliable responding observed in the tool access condition may have been an artifact of having evaluated so many scenarios in the previous (no tool) condition. However, the scenarios were presented in random order for each condition. Additionally, content featured in scenarios was drawn directly from behavioural literature. Raters would have had exposure to this type of material on an ongoing basis by simply accessing journal articles. Moreover, raters did not receive feedback regarding whether a scenario was informed by a paper that designated the case as severe or not. Therefore, it may be unlikely that repeated exposure to the scenarios featured in the study could have somehow confounded reliability across conditions. Another consideration of testing effect was the lack of counterbalancing. Although I randomized the presentation of the scenarios within each condition, I could not counterbalance scenarios across the two conditions within the context of a pre-post design. That is, rater 1 could not see different scenarios than rater 2 across conditions because then reliability could not be evaluated. Further, counterbalancing conditions was not possible. That is, if I conducted the tool condition first for some raters followed by the no tool condition, I could not control if information from the tool would influence their responses in the no tool condition that followed. Finally, other validated tools in the literature such as the ABC used similar methodology during tool development (Lehotkay et al., 2015). Although it is unlikely that testing effects unduly influenced my outcomes, future research could conduct two no tool conditions, prior to a tool condition, to determine whether reliability remains stable across the two conditions with no access to the tool (i.e., baseline).

Second, although maturation can be difficult to control for in the context of a pre-post design study, median time to study completion was three weeks. Therefore, it is unlikely that a period this brief could have unduly influenced the results. Fourth, regarding history as a confound – all raters were Ontario-based clinicians and completed the study during one of the five months of COVID-19 lock down. Therefore, it may be reasonable to conclude that any influence historical events associated with COVID-19 could have had were minimized. However, this event may further limit generality given conducting future research evaluating updated tool iterations will likely not recruit raters who will be completing the study during a pandemic lock down.

Finally, although designs incorporating randomization are often considered the *gold standard* for establishing causal evidence (Gopalan et al., 2020) a pre-post design was likely more appropriate in the current context. This is in part because they are commonly employed across tool evaluation research; especially for early iterations of tools (Carballo-Fazanes et al., 2021; O'Hara & Rehm, 1983; Wagner et al., 2007)

There have been reported limitations with the ICC analysis. Specifically, that it is not a universal analysis, the statistic assumes normality, and the estimates are dependent on the range of the scale (i.e., tool; Müller & Büttner, 1994). Authors highlight that the careful selection of the ideal method for an analysis will help to avoid and address these limitations. The selection of the ICC was well suited for the sample size, dichotomous variables, and methods. I carefully followed a selection procedure outlined by Portney and Watkins (2008) which features a flow chart identifying specific components of the study elements to determine the appropriate reliability analysis. With this considered, ICC was the best suited method to compare the consistency of rater agreement. Notably, various authors evaluating standardized tools have employed ICC as the initial analysis to inform reliability of raters and planning of future iterations (Carballo-Fazanes et al., 2021; O'Hara & Rehm, 1983).

**General Strengths**

Despite the limitations mentioned above, there are several noteworthy strengths. First, I balanced and randomized the scenarios in each condition by generating 30 severe and 30 non severe scenarios and randomly assigning 10 severe scenarios (from a possible 30) to each condition and 10 non severe scenarios (from a possible 30) to each condition. This meant that scenarios were not presented uniformly and ensured raters saw equal number of *severe* and not *severe* scenarios.

Second, I generated a large number of scenarios for raters to review (i.e., offered a large number of *observations*). This circumvented the need to recruit large numbers of raters for the purpose of evaluating the *first* iteration of this tool. I established maximal power for statistical reporting by incorporating a large number of observations (20) per participant, thus justifying a smaller sample size (Bujang & Baharum, 2017).

Finally, raters were recruited from across Ontario and the content in each of the scenarios came directly from participant profiles in behaviour analytic publications. At times, existing literature evaluating a tool's reliability has featured a group of clinician raters from the *same agency* while using existing client cases as testing material (case scenarios) (Rojahn et al., 2001; Paclawskyj et al., 1997). These study features may introduce a host of confounds that were not a problem for the current study.

## Future Research

The results of this study were promising, and therefore there are several opportunities for future research. First, a more rigorous research design should be implemented to corroborate the outcomes of the current project. For example, a group design with a control group or a grouped multiple baseline design such that one group experiences the no tool and tool condition, while the other group experiences two no tool conditions before a tool access condition. These designs may help to control for the threat of testing. A second avenue for further research would be establishing a reliable *cut-off* to determine which raw score on the severity tool would need to be observed for a case to be classified as severe. Third, recruiting raters from different professions that allocate funding and triaging (e.g., insurance providers or casework managers) may help to establish the utility of this tool in justifying treatment and funding allocation. Finally, research is warranted to explore the validity of the tool to ensure it is evaluating the right behaviour dimensions to achieve an objective severity report.

## Conclusion

The results of my study suggest the current iteration of the tool achieved good internal consistency.  Although expert raters showed higher reliability in the no tool and tool conditions, both rater groups showed a meaningful improvement in reliability upon accessing the tool. Finally, all raters evaluated the tool with high social validity. While the study showcases several important strengths, including, balanced and novel stimuli, a high number of observations and a comparison of well classified rater groups; study limitations were primarily small sample size and a pre-post design. Although there is further work to be done to establish the final iteration of this tool, these preliminary results (i.e., my study results) are encouraging.

# References

Aman, M. G. (2013) Aberrant Behavior Checklist. In: Volkmar F.R. (eds) *Encyclopedia of Autism Spectrum Disorders.* Springer, New York, NY. https://doi.org/10.1007/978-1-4419-1698-3

Aman, M. G., & Singh, N. N. (1985). The aberrant behavior checklist: A behavior rating scale for the assessment of treatment effects. *American Journal of Mental Deficiency, 89*(5), 485-491. https://www.researchgate.net/publication/19167636_The_Aberrant_Behavior_Checklist_A_behavior_rating_scale_for_the_assessment_of_treatment_effects

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders, DSM-V, fifth edition.* Washington, DC: Author.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-current dimensions of applied behaviour analysis. *Journal of Applied Behavior Analysis, 20*(4), 313-327. https://doi.org/10.1901/jaba.1987.20-313

Baker, B. L. (2017). Intellectual disability and developmental risk: Promoting intervention to improve child and family well-being. *Child Development, 88*(2), 436-445. https://doi.org/10.1111/cdev.12740

Baker, S. P., O'Neill, B., Haddon, W., & Long, W. B. (1974). The injury severity score: A method for describing patients with multiple injuries and evaluating emergency care. *Journal of Trauma, 14,* 187-196. https://journals.lww.com/jtrauma/Citation/1974/03000/THE_INJURY_SEVERITY_SCORE__A_METHOD_FOR_DESCRIBING.1.aspx

Barancik ,J. I., & Chatterjee, B. F. (1981). Methodological considerations in the use of the Abbreviated Injury Scale in trauma epidemiology. *Journal of Trauma, 21*, 627- 631. https://doi.org/10.1097/00005373-198108000-00006

Behavior Analyst Certification Board. (2017). *BACB October 2017 newsletter: Special edition on experience and supervision requirements.* https://www.bacb.com/wpcontent/uploads/BACB_Newsletter_101317.pdf

Behavior Analyst Certification Board. (2021). *US employment demand for behavior analysts: 2010-2020*. Littleton, CO: Author https://www.bacb.com/wp-content/uploads/2021/01/BurningGlass2021_210126.pdf

Berg, W. K., Wacker, D. P., Ringdahl, J. E., Stricker, J., Vinquist, K., Salil Kumar Dutt, A., ... & Mews, J. (2016). An integrated model for guiding the selection of treatment components for problem behavior maintained by automatic reinforcement. *Journal of Applied Behavior Analysis, 49(3),* 617-638. https://doi.org/10.1002/jaba.303

Bjaastad, J. F., Jensen-Doss, A., Moltu, C., Jakobsen, P., Hagenberg, H., & Joa, I. (2019). Attitudes toward standardized assessment tools and their use among clinicans in a public mental health service. *Nordic Journal of Psychiatry, 73*(7), 387-396. https://doi.org/10.1080/08039488.2019.1642383

Blenkush, N. A., & O'Neill, J. (2020). Contingent skin-shock treatment in 173 cases of severe problem behaviour. *International Journal of Psychology and Behavior Analysis, 6*(167), 1-10. https://doi.org/10.15344/2455-3867/2020/167

Bonner, A. C., & Borrero, J. C. (2019). Differential reinforcement of low rate schedules reduce severe problem behavior. *Behavior Modification, 42*(5), 747-764. https://doi.org/10.1177/0145445517731723

Brodhead, M. T., Quigley, S. P., & Wilczynski, S. M. (2018). A call for discussion about scope of competence in behaviour analysis. *Behavior Analysis in Practice, 11*, 424-435. https://doi.org/10.1007/s40617-018-00303-8

Bujang, M. A., & Baharum, N. (2017). A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. *The Journal of the School of Dental Sciences, 12*(1), 1-11. https://www.researchgate.net/publication/318788161_A_simplified_guide_to_determination_of_sample_size_requirements_for_estimating_the_value_of_intraclass_correlation_coefficient_A_review

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A review on sample size determination for Cronbach's alpha test: A simple guide for researchers. *Malays Journal of Medical Science, 25*(6), 85-99. https://doi.org/10.21315/mjms2018.25.6.9

Bürkner, P., Doebler, P., & Holling, H. (2016). Optimal design of the Wilcoxon-Mann-Whitney-test. *Biometrical Journal, 59*(2017), 25-40. https://doi.org/10.1002/bimj.201600022

Burning Glass Technologies. (2015). *US behavior analyst workforce: understanding the national demand for behavior analysts.* https://www.simmons.edu/sites/default/files/2019-03/Behavior%20Analysis%20Careers.pdf

Busner, J., & Targum, S. D. (2007). The clinical global impressions scale: applying a research tool in clinical practice. *Psychiatry, 4*(7), 28-37. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2880930/pdf/PE_4_7_28.pdf

Carballo-Fazanes, A., Rey, E., Valentini, N. C., Rodríguez-Fernández, J. E., Varela-Casal, C., Rico-Díaz, J., Barcala-Furelos, R., & Abelairas-Gómez, C. (2021). Intra-rater (live vs. video assessment) and inter-rater (expert vs. novice) reliability of the test of gross motor development – third edition. *International Journal of Environmental Research and Public Health, 18*(4), 1-12. https://doi.org/10.3390/ijerph18041652

Chen, S., Tao, T., Tao, Q., Fang, Y., Zhou, X., Chen, H., Chen, Z., Huang, J., Chen, L., & Chan, C. C. H. (2016). Rater experience influences reliability and validity of the brief international classification of functioning, disability, and health core set for stroke. *Journal of Rehabilitation Medicine, 48*(3), 265–272. https://doi.org/10.2340/16501977-2063

Committee on the Medical Aspects of Automotive Safety. (1971). Rating the severity of tissue damage: I. The abbreviated scale. *Journal of the American Medical Association, 215*, 277-280. https://doi.org/10.1001/jama.1971.03180150059012

Cooper, J. O., Heron, T. E., & Heward, W. L. (2020). *Applied Behaviour Analysis (3rd Ed)*. Upper Saddle River, NJ: Pearson Education

Cox, A. D., Leung, J., Anderson, B. M., & Morgan, M-C. (2020, December 7). Examining research patterns in the treatment of adults with problem behavior and intellectual disability: A brief review. *Behavioral Development.* Advanced online publication. http://doi.org/10.1037/bdb0000100

Cumming, G., & Finch, S. (2005). Inference by eye confidence intervals and how to read pictures of data. *American Psychologist, 60*(2), 170-180. https://doi.org/10.1037/0003-066X.60.2.170

Busch, L., Cox, A., Cunningham, J. & Saini, V. (2019). *Evidenced-based practices for the treatment of challenging behaviour in intellectual and developmental disabilities: Recommendations for caregivers, practitioners, and policy makers.* The Ontario

Association for Behaviour Analysis, Inc. https://www.ontaba.org/pdf/ONTABA_OSETT-CB_Final_Report_Jan_2019.pdf

Danov, S. E., Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behaviour Modification, 32*(6), 828-839. https://doi.org/10.1177/0145445508318606

Dracobly, J. D., Dozier, C. L., Briggs, A. M., & Juanico, J. F. (2017). Reliability and validity of indirect assessment outcomes: Experts versus caregivers. *Learning and Motivation, 62*(2018), 77-90. https://doi.org/10.1016/j.lmot.2017.02.007

Deb, S., Sohanpal, S., Soni, R., Lentre, L., & Unwin, G. (2007). The effectiveness of antipsychotic medication in the management of behaviour problems in adults with intellectual disabilities. *Journal of Intellectual Disability Research, 51*(10), 766–777. https://doi.org/10.1111/j.1365-2788.2007.00950.x

Deb, S., Unwin, G., & Deb, T. (2015). Characteristics and the trajectory of psychotropic medication use in general and antipsychotics in particular among adults with an intellectual disability who exhibit aggressive behaviour. *Journal of Intellectual Disability Research, 59*, 11-25. https://doi.org/10.1111/jir.12119

Deochand, N., & Fuqua, R. W. (2016). BACB certification trends: State of the states (1999 to 2014). *Behavior Analysis in Practice, 9*(3), 243–252. https://doi.org/10.1007/s40617-016-0118-z

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology, 17*(2), 245-260. https://doi.org/10.1023/A:1019637632584

Donenberg G., & Baker B. L. (1993) The impact of young children with externalizing behaviors on their families. *Journal of Abnormal Child Psychology, 21*(2), 179–98. https://doi.org/10.1007/BF00911315

Eisinga, R., Grotenhuis, M., & Pelzer, B. (2012). The reliability of a two-item scale: Pearson, Cronbach, or spearman-brown? *International Journal of Public Health, 2013*(58), 637-642. https://dio.org/10.1007/s00038-012-0416-3

Emerson, E. (2001). *Challenging behaviour: Analysis and intervention in people with learning difficulties.* Cambridge: University Press.

Emerson, E. (2010). Deprivation, ethnicity and the prevalence of intellectual and developmental disabilities. *Journal of Epidemiol Community Health, 66*, 218-224. https://doi.org/10.1136/jech.2010.111773

Emerson, E., Kiernan, C., Alborz, A., Reeves, D., Mason, H., Swarbrick, R., Mason, L., & Hatton, C. (2001). The prevalence of challenging behaviours: A total population study. *Research in Developmental Disabilities, 22*(1), 77-93. https://doi.org/10.1016/S0891-4222(00)00061-5

Evers, C., & Pilling, N. (2012). Improving outcomes for people with severe challenging behaviour. *Learning Disability Practice, 15*(9), 30-36. https://doi.org/10.7748/ldp2012.11.15.9.30.c9382

Fagerland, M. W., & Sandvik, L. (2009). The Wilcoxon-Mann-Whitney test under scrutiny. *Statistics in Medicine, 28*, 1487-1497. https://doi.org/10.1002/sim.3561

Fahmie, T. A. & Iwata, B. A. (2011). Topographical and functional properties of precursors to severe problem behavior. *Journal of Applied Behavior Analysis, 44*(4), 993-997. https://doi.org/10.1901/jaba.2011.44-993

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys, 4*(2010), 1-39. https://doi.org/10.1214/09-SS051

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics (5th Ed.)*. California: SAGE Publications.

Finney, J. W., Putnam, D. E., & Boyd, C. M. (1998). Improving the accuracy of self-reports of adherence. *Journal of Applied Behavior Analysis, 31*(3), 485-488. https://doi.org/10.1901/jaba.1998.31-485

Floyd, R. G., Phaneuf, R. L., & Wilczynski, S. M. (2005). Measurement properties of indirect assessment methods for functional behavioral assessment: A review of research. *School Psychology Review, 34*(1), 58-73. https://doi.org/10.1080/02796015.2005.12086275

Foxx, R. M. (2003). The treatment of dangerous behavior. *Behavioral Interventions, 18*(1)*,* 1-21. https://doi.org/10.1002/bin.127

Friedman, D. J., Parrish, G. R., & Fox, M. H. (2018). A review of global literature on using administrative data to estimate prevalence of intellectual and developmental disabilities. *Journal of Policy and Practice in Intellectual Disabilities, 15,* 43-62. https://doi-org/10.1111/jppi.12220

Fritz, J. N., Iwata, B. A., Hammond, J. L., & Bloom, S. E. (2013). Experimental analysis of precursors to severe problem behavior. *Journal of Applied Behavior Analysis, 46*(1), 101-129. https://doi.org/10.1002/jaba.27

Gopalan, M., Rosinger, K., & Ahn, J. B. (2020). Use of quasi-experimental research designs in education research: Growth, promise, and challenges. *Review of Research in Education 44*, 218-243. https://doi.org/10.3102/0091732X20903302

Gulgin, H., & Hoogenboom, B. (2014). The functional movement screening (fms)TM: An inter-rater reliability study between raters of varied experience. *International Journal of Sports Physical Therapy, 9*(1), 14–20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3924604/pdf/ijspt-02-014.pdf

Hajiaghamohseni, Z., Drasgow, E., & Wolfe, K. (2020). Supervision behaviors of board certified behavior analysts with trainees. *Behavior Analysis in Practice, 2021*(14), 97-109. https://doi.org/10.1007/s40617-020-00492-1

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34. https://doi.org/10.20982/tqmp.08.1.p023

Hanratty, J., Livingstone, N., Robalino, S., Terwee, C. B., Glod, M., Oono, I. P., Rodgers, J., Macdonald, G., & McConachie, H. (2015). Systematic review of the measurement properties of tools used to measure behaviour problems in young children with autism. *PLoS One, 10*(12), 1-21. https://doi.org/10.1371/journal.pone.0144649

Hausman, N., Kahng, S., Farrell, E., & Mongeon, C. (2009). Idiosyncratic functions: Severe problem behavior maintained by access to ritualistic behaviors. *Education and Treatment of Children, 32*(1), 77-87. https://doi.org/10.1353/etc.0.0051

Hmelo-Silver, C. E., Pfeffer, M. G. (2003). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviours and functions. *Cognitive Science, 28*(2004), 127-138. https://doi.org/10.1207/s15516709cog2801_7

Hodge, D., & Gillespie, D. (2003). Phrase completions: an alternative to Likert scales. (Note on research methodology). *Social Work Research, 27*(1), 45–55. https://doi.org/10.1093/swr/27.1.45

Iwata, B. A., Pace, G. M., Kissel, R. C., Nau, P. A., & Farber, J. M. (1990). The self-injury trauma (SIT) scale: A method for quantifying surface tissue damage caused by self-injurious behavior. *Journal of Applied Behavior Analysis. 23*(1): 99-110. https://doi.org/10.1901/jaba.1990.23-99

Jensen-Doss, A., Hawley, K. M. (2010). Understanding barriers to evidence-based assessment: Clinical attitudes toward standardizes assessment tools. *Journal of Clinical Child and Adolescent Psychology, 39*(6), 885-896. https://doi.org/10.1080/15374416.2010.517169

Jessel, J., Ingvarsson, E. T., Metras, R., Kirk, H., & Whipple, R. (2018). Achieving socially significant reductions in problem behavior following the interview-informed synthesized contingency analysis: A summary of 25 outpatient applications. *Journal of Applied Behavior Analysis, 51*(1), 130-157. https://doi.org/10.1002/jaba.436

Jurs, S. G., & Glass, G. V. (1971). The effect of experimental morality on the internal and external validity of the randomized comparative experiment. *The Journal of Experimental Education, 40*(1), 62-66. https://www.jstor.org/stable/20157241

Karabekiroglu, K., & Aman, M. G. (2009). Validity of the aberrant behaviour checklist in a sample of toddlers. *Child Psychiatry and Human Development, 2009*(40), 99-110. https://doi.org/10.1007/s10578-008-0108-7

Katowa-Mukwato, P., Andrews, B., Maimbolwa, M., Lakhi, S., Michelo, C., Mulla, Y., & Banda, S. S. (2014). Medical students' clerkship experiences and self-perceived competence in clinical skills. *Journal of Health and Professional Education, 6*(2), 155-160. https://doi.org/10.7196/ajhpe.358

Kim, J. I., Shin, M., Lee, Y., Lee, H., Yoo, H. J., Kim, S., Kim, H., Kim, S., & Kim, B. (2018). Reliability and validity of a new comprehensive tool for assessing challenging behaviors in autism spectrum disorder. *Psychiatry Investigation, 15*(1), 54-61. https://doi.org/10.4306/pi.2018.15.1.54

Knight, V. F., Wright, J., & DeFreese, A. (2019). Teaching robotics to a student with ASD and severe problem behaviour. *Journal of Autism and Developmental Disorders, 49*, 2632-2636. https://doi.org/10.1007/s10803-019-03888-3

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.2016.02.0121556-3707

Lach, L. M., Kohen, D. E., Garner, R. E., Brehaut, J. C., Miller, A. R., Klassen, A. F., & Rosenbaum, P. L. (2009). The health and psychosocial functioning of caregivers of children with neurodevelopmental disorders. *Disability and Rehabilitation, 31*(8), 741–752. https://doi.org/10.1080/09638280802242163

Lam, K. A. L., & Aman, M. G. (2006). The repetitive behaviour scale – revised: Independent validation in individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 2007*(37), 855-866. https://doi-org.proxy.library.brocku.ca/10.1007/s10803-006-0213-z

Lehotkay, R. Devi, S., Raju, M. V. R., Bada, P. K., Nuti, S., Kempf, N., & Carminati, G. G. (2015). Factor validity and reliability of the aberrant behavior checklist-community (ABC-C) in an Indian population with intellectual disability. *Journal of Intellectual Disability Research, 59*, 208-214. https://doi.org/10.1111/jir.12128

Liljequist, D., Elfving, B., & Roaldsen, K. S. (2019). Intraclass correlation – A discussion and demonstration of basic features. PLoS ONE 14(7) 1-35 https://doi.org/10.1371/journal.pone.0219854

Lowe, K., Allen, D., Jones, E., Brophy, S., Moore, K., & James, W. (2007). Challenging behaviours: prevalence and topographies. *Journal of Intellectual Disability Research, 51*, 625-636. https://doi.org/10.1111/j.1365-2788.2006.00948.x

Lwu, S., Paolucci, E. O., Hurlbert, R. J., & Thomas, K. C. (2010). A scoring system for elective triage of referrals: spine severity score. *The Spine Journal, 10*(8), 697-703. https://doi.org/10.1016/j.spinee.2010.05.011

Mayer, T., Matlak, M. E., Johnson, D. G., & Walker, M. L. (1980). The modified injury severity scale in pediatric multiple trauma patients. *Journal of Pediatric Surgery, 15*, 719-726. https://doi.org/10.1016/s0022-3468(80)80271-5

McIntyre, L. L., Blacher, J., & Baker, B. L. (2002). Behaviour/mental health problems in young adults with intellectual disability: The impact on families. *Journal of Intellectual Disability Research, 46*(3), 239-249. https://doi.org/10.1046/j.1365-2788.2002.00371.x

Morris, E. K., Altus, D. E., & Smith, N. G. (2013). A study in the founding of applied behavior analysis through its publications. *The Behavior Analyst, 36*(1), 73–107. https://doi.org/10.1007/BF03392293

Müller, R., & Büttner, P. (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine, 13*(23-24), 2465-2476. https://doi.org/10.1002/sim.4780132310

Murphy, O., Healy, O., & Leader, G. (2009). Risk factors for challenging behaviours among 157 children with autism spectrum disorder in Ireland. *Research in Autism Spectrum Disorders, 3*(2): 474-482. https://doi.org/10.1016/j.rasd.2008.09.008

Newton, J. T., & Sturmey, P. (1988). The aberrant behavior checklist: A british replication and extension of its psychometric properties. *Journal of Mental Deficiency Research, 32*, 87-92. https://doi.org/10.1111/j.1365-2788.1988.tb01394.x

O'Hara, M. W., & Rehm, L. P. (1983). Hamilton rating scale for depression: Reliability and validity of judgements of novice raters. *Journal of Consulting and Clinical Psychology, 51*(2), 318-319. https://doi.org/10.1037//0022-006x.51.2.318

Oropeza, M. E., Fritz, J. N., Nissen, M. A., Terrell, A. S., & Phillips, L. A. (2018). Effects of therapist-worn protective equipment during functional analysis of aggression. *Journal of Applied Behavior Analysis, 51*(3), 681-686. https://doi.org/10.1002/jaba.457

Paclawskyj, T. R., Matson, J. L., Bamburg, J. W., & Baglio, C. S. (1997). A comparison of the diagnostic assessment for the severely handicapped-II (DASH-II) and the aberrant behavior checklist (ABC). *Research in Developmental Disabilities, 18*(4), 289-298. https://doi.org/10.1016/S0891-4222(97)00010-3

Palmer, C., Gabbe, B., & Cameron, P. (2016). Defining major trauma using the 2008 abbreviated injury scale. *Injury, 47*(1), 109–115. https://doi.org/10.1016/j.injury.2015.07.003

Poppes, P., Putten, A. J. J. V., & Vlaskamp, C. (2010). Frequency and severity of challenging behaviour in people with profound intellectual and multiple disabilities. *Research in Developmental Disabilities, 31*, 1269-1275. https://doi.org/10.1016/j.ridd.2010.07.017

Portney, L. G., & Watkins, M. P. (2008). *Foundations of Clinical Research: Applications to Practice* (3rd ed.). Pearson.

Randsborg, P., & Sivertsen, E. A. (2012). Classification of distal radius fractures in children: good inter- and intraobserver reliability, which improves with clinical experience. *BMC Musculoskelet Disorder, 13*(6), 1-8. https://doi.org/10.1186/1471-2474-13-6

Ricciardi, J. N., & Rothschild, A. W. (2017). 5-Behavioral Risk Assessment. In J. K. Luiselli (Ed.), Applied behavior analysis advanced guidebook: A manual for professional practice. (pp.93-116). Academic Press. https://doi.org/10.1016/B978-0-12-811122-2.00005-X

Rojahn, J., Aman, M. G., Matson J. L., & Mayville, E. (2003). The aberrant behavior checklist and the behavior problems inventory: Convergent and divergent validity. *Research in Developmental Disabilities, 24*(5), 391-404. https://doi.org/10.1016/S0891-4222(03)00055-6

Rojahn, J., Matson, J. L., Lott, D., Esbensen, A. J., & Smalls, Y. (2001). The behavior problems inventory: An instrument for assessment of self-injury, stereotyped behavior and aggression/destruction in individuals with developmental disabilities. *Journal of Autism and Developmental Disorders, 31*(6), 577-588. https://doi.org/10.1023/a:1013299028321

Rojahn, J. Rowe, E. W., Sharber, A. C., Hastings, R., Matson, J. L., Didden, R., Kroes, D. B. H., & Dumont, E. L. M. (2012). The behavior problems inventory-short form for individuals with intellectual disabilities: Part I: Development and provisional clinical reference data. *Journal of Intellectual Disability Research, 56*(5), 527-545. https://doi.org/10.1111/j.1365-2788.2011.01507.x

Rojahn, J., Schroeder, S. R., Mayo-Ortega, L., Oyama-Ganiko, R., LeBlanc, J., Marquis, J., & Berke, E. (2013). Validity and reliability of the behaviour problems inventory, the aberrant behavior checklist, and the repetitive behavior scale – revised among infants and toddlers at risk for intellectual or developmental disabilities: A multi-method assessment approach. *Research in Developmental Disabilities, 34*(2013), 1804-1814. http://dx.doi.org/10.1016/j.ridd.2013.02.024

Roscoe, E. M., Iwata, B. A., & Zhou, L. (2013). Assessment and treatment of chronic hand mouthing. *Journal of Applied Behavior Analysis, 46*(1), 181-198. https://doi.org/10.1002/jaba.14

Saini, V., Betz, A. M., Gregory, M. K., Leon, Y., & Fernandez, N. (2017). A survey-based method to evaluate optimal treatment selection for escape-maintained problem behavior. *Journal of Applied Behavior Analysis, 2017*(10), 214-227. https://doi.org/10.1007/s40617-017-0180-1

Saini, V., Sullivan, W. E., Baxter, E. L., DeRosa, N. M., & Roane, H. S. (2018). Renewal during functional communication training. *Journal of Applied Behavior Analysis*. *51*(3), 603-619. https://doi.org/10.1002/jaba.471

Schmidt, J. D., Drasgow, E., Halle, J. W., Martin, C. A., & Bliss, S. A. (2014). Discrete-trial functional analysis and functional communication training with three individuals with autism and severe problem behavior. *Journal of Positive Behavior Interventions, 16*(1), 44-55. https://doi.org/10.1177/1098300712470519

Selles, R., Mcbride, N., Dammann, J., Whiteside, S., & Storch, E. (2018). Initial psychometrics, outcomes, and correlates of the repetitive body focused behavior scale: Examination in a sample of youth with anxiety and/or obsessive-compulsive disorder. *Comprehensive Psychiatry, 81*, 10–17. https://doi.org/10.1016/j.comppsych.2017.11.001

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika, 74*(1), 107-120. https://doi.org/10.1007/s11336-008-9101-0

Sival, R., Albronda, T., Haffmans, P., Saltet, M., & Schellekens, C. (2000). Is aggressive behaviour influenced by the use of a behaviour rating scale in patients in a psychogeriatric nursing home? *International Journal of Geriatric Psychiatry, 15*(2), 108–

111. https://doi.org/10.1002/(SICI)1099-1166(200002)15:2<108::AID-GPS80>3.0.CO;2-H

Slocum, T. A., Detrich, R., Wilczynski, S. M., Spencer, T. D., Lewis, T., & Wolfe, K. (2014). The evidence-based practice of applied behavior analysis. *The Behavior analyst, 37*(1), 41–56. https://doi.org/10.1007/s40614-014-0005-2

Statistics Canada. (2017). *North American Industry Classification System (NAICS) Canada.* https://www150.statcan.gc.ca/n1/pub/12-501-x/12-501-x2016003-eng.pdf

Sturmey, P., & Didden, R. (2014). *Evidence-based practice and intellectual disabilities.* Chichester, England: Wiley-Blackwell.

Targum, S. D., Hassman, H., Pinho, M., & Fava, M. (2011). Development of a clinical global impression scale for fatigue. *Journal of Psychiatric Research, 46*(2012), 370-374. https://doi.org/10.1016/j.jpsychires.2011.12.001

Taylor, L., Oliver, C., & Murphy, G. (2011). The chronicity of self-injurious behaviour: A long-term follow-up of a total population study. *Journal of Applied Research in Intellectual Disabilities, 24*(2), 105-117. https://doi.org/10.1111/j.1468-3148.2010.00579.x

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education 2011*(2), 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Wagner, A., Lecavalier, L., Arnold, E., Aman, M. G., Scahill, L., Dtigler, K. A., Johnson, C. R., McDougle, C. J., & Vitiello, B. (2007). Developmental disabiltiies modification of children's global assessment scale. *Biological Psychiatry, 61*(4), 504-511. https://doi.org/10.1016/j.biopsych.2007.01.001

Weck, F., Hilling, C., Schermelleh-Engel, K., Rudari, V., & Stangier, U. (2011). Reliability of adherence and competence assessment in cognitive behavioral therapy. *The Journal of*

*Nervous and Mental Disease. 199*(4), 276-279. https://doi.org/10.1097/NMD.0b013e3182124617

Westerman, D. A. (1991). Expert and novice teacher decision making. *Cognitive Science, 42*(4), 292-305. https://doi.org/10.1177/002248719104200407

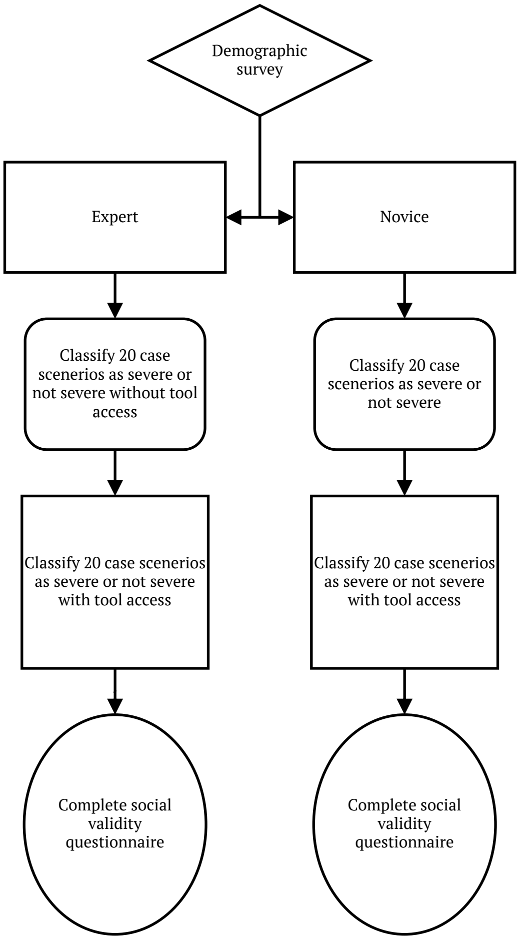
Villalobos, I. C. B., Ramírez, G. C., & Acosta, J. R. (2014). Intra-rater reliability and the role of experience: A comparative case. *Revista de Lenguas Modernas, 20,* 295-307. Retrieved from https://doaj.org/article/f84f2c7fe43e42a6a8f6d468ee46093d

Zarcone, J. R., Hagopian, L., Ninci, J., McKay, C., Bonner, A., Dillon, C., & Hausman, N. (2016). Measuring the complexity of treatment for challenging behaviour using the treatment intensity rating form. *International Journal of Developmental Disabilities, 62*(3), 183-191. https://doi.org/10.1080/20473869.2016.1173316

Zarkada, A., & Regan, J. (2018). Inter-rater reliability of the dysphagia outcome and severity scale (DOSS): Effects of clinical experience, audio-recording and training. *Dysphagia, 33*(3), 329–336. https://doi.org/10.1007/s00455-017-9857-4

**Figure 1**

*Flowchart of Procedures*


Demographic survey
Expert
Novice
Classify 20 case scenerios as severe or not severe without tool access
Classify 20 case scenerios as severe or not severe
Classify 20 case scenerios as severe or not severe with tool access
Classify 20 case scenerios as severe or not severe with tool access
Complete social validity questionnaire
Complete social validity questionnaire

*Note.* First set of case scenarios will be different scenarios than the second set.

**Table 1**

*Rater Demographic Results*

| Rater demographic | Ratio of raters (*N*=9) | | | |
|---|---|---|---|---|
| | Novice (*n*=5) | Novice *Mdn* | Experts (*n*=4) | Expert *Mdn* |
| Years Certified | | | | |
| < 1 year | 2 | 18 months | - | 99.5 months |
| ≥ 1 year but <3 years | 2 | | - | |
| ≥ 3 years but <5 years | 1 | | - | |
| ≥ 5 years but < 7 years | - | | 1 | |
| ≥ 7 years but <10 years | - | | 2 | |
| ≥ 10 years | - | | 1 | |
| Regularity working with severe problem behaviour | | | | |
| Never | 1 | Regularly (more than 5 client cases) | - | Regularly (more than 5 client cases) |
| Limited (less than 5 client cases) | 1 | | - | |
| Regularly (more than 5 client cases) | 3 | | 4 | |
| Years of experience supporting severe problem behaviour | | | | |
| < 1 year | 1 | ≥ 5 years but < 10 years | - | ≥ 10 years but < 15 years |
| ≥ 1 year but < 2 years | - | | - | |
| ≥ 2 years but < 5 years | 3 | | - | |
| ≥ 5 years but < 10 years | 1 | | 1 | |
| ≥ 10 years but < 15 years | - | | 2 | |
| ≥ 15 years but < 18 years | - | | - | |
| ≥ 18 years | - | | 1 | |
| Number of clients supported throughout their BCBA | | | | |
| < 5 clients | 2 | ≥ 15 clients but < 20 clients | - | >20 clients |
| ≥ 5 clients but < 10 clients | - | | - | |
| ≥ 10 clients but < 15 clients | - | | - | |
| ≥ 15 clients but < 20 clients | 1 | | - | |
| ≥ 20 clients | 2 | | 4 | |
| Age groups treated [a] | | | | |
| < 1 year | - | *Mdn*= 5.00<br>*M*= 4.20 | - | *Mdn*= 4.50<br>*M*= 2.20 |
| ≥ 1 year but < 2 years | 2 | | - | |
| ≥ 2 years but < 5 years | 5 | | 3 | |
| ≥ 5 years but < 10 years | 5 | | 3 | |
| ≥ 10 years but < 15 years | 4 | | 3 | |
| ≥ 15 years but < 18 years | 3 | | 4 | |
| ≥ 18 years but < 61 years | 2 | | 3 | |
| ≥ 61 years | - | | 1 | |
| Current or previous work settings [a] | | | | |
| Intervention center (i.e., 1:1 treatment room, playroom, lunchroom) | 5 | *Mdn*= 2.00<br>*M*= 4.25 | 3 | *Mdn*= 3.00<br>*M*= 3.50 |
| In client home (Private home) | 3 | | 3 | |
| Group Day program (i.e., large room with more than 5 clients, community center) | 1 | | 1 | |
| Group Home (i.e., Common areas with individual overnight bedrooms) | - | | 3 | |
| Large, supported accommodation residence (e.g., institution) | - | | 1 | |
| School classroom (this also includes segregated classrooms if it is considered school property) | 2 | | 2 | |
| Hospital (i.e., Hospital beds, nursing, and medical staffing on site, integrated with other typically developing patients) | - | | - | |
| Secured unit (e.g., dual diagnosis treatment facility, correctional facilities) | - | | 1 | |

*Note.* The dash (-) indicates that none of the raters were assigned to that demographic.

[a] Raters could select multiple selections for this demographic.

*Mdn* = the median

*M* = the mean

**Table 2**

*Results of Mann-Whitney*

| Rater Demographic | Mann-Whitney *U* | *P*-value |
|---|---|---|
| Years Certified | 20.00 | .016 |
| Years of experience supporting severe problem behaviour | 19.50 | .016 |
| Number of clients supported throughout their BCBA | 16.00 | .109 |
| Regularity working with severe problem behaviour | 14.00 | .413 |
| Age groups treated | 9.50 | .905 |
| Current or previous work settings | 15.00 | .286 |

**Table 3**

*Corrected Item-total Correlation and Cronbach's Alpha by Item*

| Items | Corrected item-total correlation | Cronbach's alpha if item deleted | Tool's Cronbach ∝ |
|---|---|---|---|
| 1. Direct observation data exists indicating problem behaviour(s) occur… | .248 | .830 | .831 |
| 2. Direct observation data exists indicating problem behaviour(s) last, uninterrupted (e.g., tantrums) … [a] | .411 | .824 | .831 |
| 3. Previous treatment(s) has decreased problem behaviour(s)… | .424 | .823 | .831 |
| 3a. Have more than two different interventions been applied with no decreases in problem behaviour(s)? | .209 | .836 | .831 |
| 4. Does the intensity of problem behaviour(s) render conducting an FA extremely unsafe? | .384 | .826 | .831 |
| 5. Client engages in problem behaviour(s) that result (has resulted) in tissue damage? (e.g., swelling, bleeding, bruising, redness) [a] | .613 | .818 | .831 |
| 5a. Has problem behaviour(s) resulted in tissue damage lasting more than 30 seconds? (e.g., redness on skin remains longer than 30-seconds) [a] | .651 | .817 | .831 |
| 6. Is there permanent tissue damage due to problem behaviour(s)? (e.g., 'cauliflower' ear; damaged vision) [a] | .427 | .823 | .831 |
| 6a. Has there been documented physical trauma or observed consistent damage within the last two months? | .593 | .814 | .831 |
| 6b. Is there physical evidence of healed injuries/lacerations as a result of engaging in problem behaviour(s) (this can include outward aggression/ property destruction resulting in an injury, i.e., breaking a window causing lacerations; in addition to self-injury) (e.g., scarring, toughened and calloused skin) [a] | .561 | .816 | .831 |
| 6c. Has problem behaviour(s) resulted in documented staff injury that required medical attention (excluding first aid). | .309 | .830 | .831 |
| 7a. Is protective gear worn by staff? | .473 | .821 | .831 |
| 7b. Is protective gear worn by the client? | .347 | .826 | .831 |
| 8. Has as problem behaviour(s) resulted in property damage? | .206 | .832 | .831 |
| 8a. If yes, what is the estimated value of the destroyed property? | .339 | .827 | .831 |
| 9. Is there evidence of past legal repercussions due to problem behaviour(s)? (i.e., court document or restraining orders for arrest, | -.101 | .839 | .831 |

| | | | |
|---|---|---|---|
| freedoms significantly limited due to sexualized behaviour targeting vulnerable populations) | | | |
| 10. Is there evidence of previous assault charges because of engaging in problem behaviour(s) (pending or otherwise)? | .018 | .833 | .831 |
| 11a. Is the client's living space confined to home or designated space by locks (either to outside or other areas of the home) | .484 | .822 | .831 |
| 11b. Do others live in the home, however, they lock their own living spaces to ensure their own safety? | .319 | .828 | .831 |
| 12. Are speciality building materials required in client residence? (e.g., reinforced drywall, lexan/safety glass; floor standing stainless steel toilets) [a] | .286 | .829 | .831 |
| 13. Is 2:1 or more staffing complement required for behaviour management? | .348 | .826 | .831 |
| 14. Is there evidence of past expulsions from one or more learning or residential placements due to problem behaviour? | .183 | .831 | .831 |
| 15. Has access to day programs been denied due to problem behaviour? | .435 | .824 | .831 |
| 16. Do healthcare consultations occur at the residence to avoid having to go into the community? | .511 | .822 | .831 |
| 17. Do community outings occur less than once per month due to the probability of problem behaviour(s) placing community members at risk? (not including riding in a vehicle wherein the client does not exit the vehicle to enter a community setting) | .638 | .827 | .831 |
| 18. Are family members or caregivers unable to engage in typical community activities due to the client's problem behaviour(s)? (e.g., vacations, swimming pool, visiting the library, going out for dinner) [a] | .612 | .818 | .831 |

*Note.* [a] Item included an example.

**Table 4**

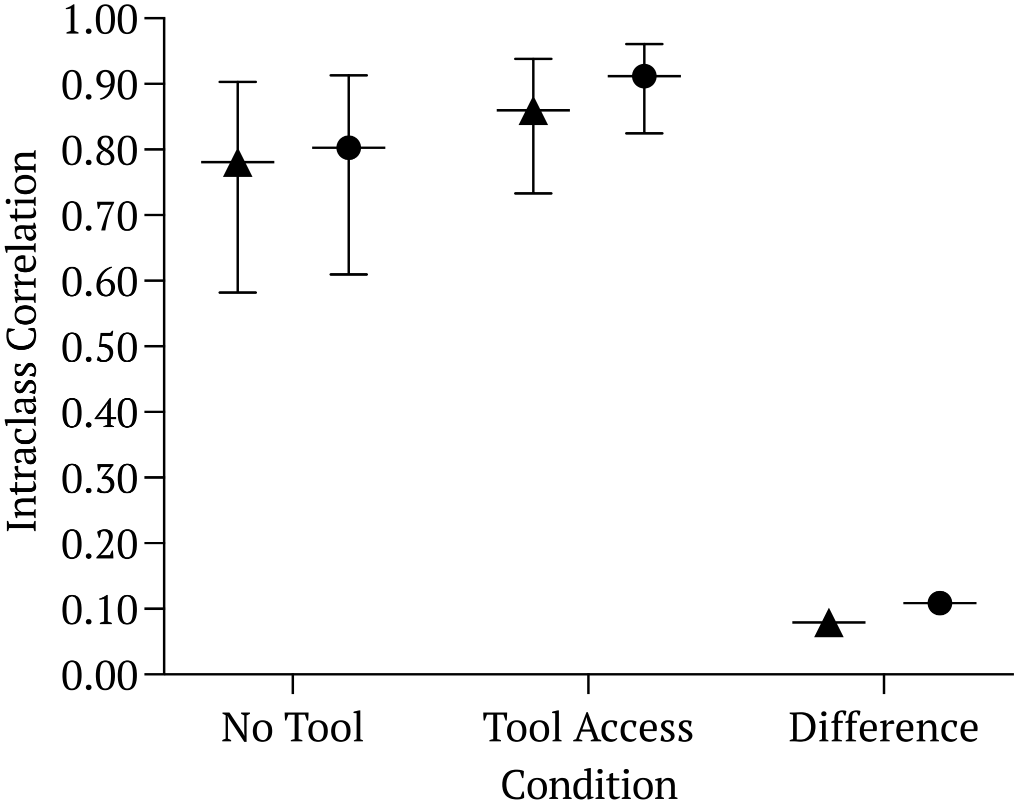
*Intraclass Correlation Results*

| Condition | | ICC | 90% Confidence interval | |
|---|---|---|---|---|
| | | | Lower bound | Upper bound |
| No Tool | Novice | .781 | .582 | .903 |
| | Expert | .803 | .610 | .913 |
| Tool Access | Novice | .860 | .733 | .938 |
| | Expert | .912 | .825 | .961 |

*Note.* Results of the ICC across all conditions and rater groups

**Figure 2**

*Intraclass Correlations with Confidence Intervals*



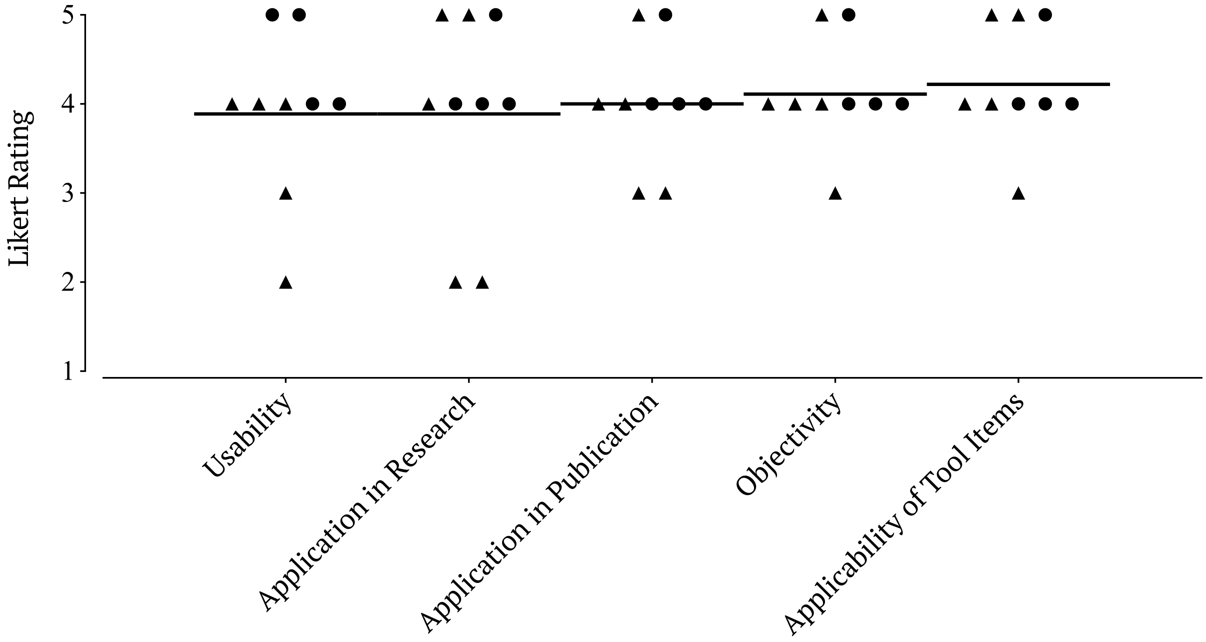
*Note.* The error bars represent the upper and lower bounds of the 95% confidence intervals for each ICC calculation. Both conditions are represented on the x-axis while the difference of each group is plotted by the datapoints on the far-right.

The line at the centre of the data point more accurately identifies the ICC value.

The closed circles represent expert raters, while the closed triangles represent novice raters.

**Figure 3**

*Results of Social Validity Questionnaire*



*Note.* Table 5 outlines the question associated with each category name on the x axis.

Strongly disagree =1, disagree = 2, neutral = 3, agree = 4, and strongly agree = 5.

Closed circles represent expert raters and closed triangles represent novice raters.

The solid horizontal lines represent the grand mean across all raters.

**Table 5**

*Social Validity Questionnaire Legend*

| Question | Category Name |
|---|---|
| This tool was easy to use. | Usability |
| This tool could be easily applied to categorize/classify severity using information provided in published participant descriptions. | Application in Research |
| This tool could be incorporated into publication practices to promote the consistent application of the term 'severe' across articles. | Application in Publication |
| This tool would be useful to achieve an objective measurement of severity. | Objectivity |
| The questions in this tool are applicable to the severity of problem behaviour. | Applicability of tool Items |

*Note.* A five-point Likert scale ranging from strongly disagree to strongly agree was used for raters to indicate their experience with the tool.

**Table 6**

*Novice and Expert Means for Social Validity Items*

| Category Name | Novice | | Expert | |
|---|---|---|---|---|
| | *M* | *Mdn* | *M* | *Mdn* |
| Usability | 3.4 | 4 | 4.5 | 4.5 |
| Application in Research | 3.6 | 4 | 4.25 | 4 |
| Application in Publication | 3.8 | 4 | 4.25 | 4 |
| Objectivity | 4 | 4 | 4.25 | 4 |
| Applicability of Tool Items | 4.2 | 4 | 4.25 | 4 |

*Note.* Strongly disagree =1, disagree = 2, neutral = 3, agree = 4, and strongly agree = 5.

## Appendix A

### Case scenarios

**Case 1**
Amy is a 15-year-old girl admitted to an institutional care facility for severe aggression and property destruction. Amy is diagnosed with Autism and Severe Developmental Delay. She engages in property destruction that has previously resulted in damages to walls, furniture and windows. Property destruction usually includes forceful contact between body parts (i.e., kicking and punching) and surfaces, as well as throwing objects. She also engages in aggression towards others, including open handed hits, pushing, spitting, and kicking. She was expelled from her last two schools where she consistently engaged in property destruction (at least two instances per minute) for over 2 hours. Her problem behaviour has resulted in peers and staff requiring medical assistance (e.g., emergency room visits). Staff use protective equipment (e.g., helmets) at the onset of Amy's behaviours. Amy requires a 2:1 staffing ratio. Staff are trained to restrain Amy when her behaviours result in severe harm to others. Amy's family reported problem behaviour in the home result in periods of confinement to the basement to protect her parents and siblings. Amy participated in two previous functional analysis. FA outcomes were undifferentiated and resulted in minor staff injury because applying designated consequences did not reliably stop the behaviour. Amy's does not participate in any social or community programs, and her skill acquisition has been stagnant because she is often too aggressive to engage in skills training. She communicates using 2-3 word sentences and often engages in repetitive vocal stereotypy.

**Case 2**
Joey is a 7-year-old boy diagnosed with Autism Spectrum Disorder. He can vocally imitate some sounds but communicates primary with modified sign language. Joey attends intensive behaviour intervention for 30 hours. He is currently working on early programming targeting imitation and receptive identification skills. Therapists are also working to reduce instances of problem behaviour, which occur approximately 4 times/hour but are described as quite intense. Joey also engages in motor stereotypy, as well as restrictive behaviours around routine. When others interrupt his routine, he will engage in aggression towards staff or family members so that he can complete the routine. Sometimes his routines can take up to 15 minutes to complete. For example, he will remove every book from the shelf and touch the first page before returning it to the shelf. Joey's aggression includes kicking, scratching and biting. Staff have made requests for additional support during behaviour reduction protocol to prevent injury. Joey's aggression has resulted in staff injury in the form of bruises and scratches requiring first aid. Some staff have reported they have permanent scarring as a result of being scratched by Joey. Staff report that Joey will participate in session for long periods without engaging in aggression. They report that functional communication training has resulted in some improvement in instances of aggression, however every instances of aggression lasts at least 15 seconds and at times, appears to 'come out of nowhere'. Joey's parents have reported that he seldom engages in aggression at home, however when he does – he typically targets his siblings.

## Appendix B

## Demographic Survey

**Are you a Board Certified Behaviour Analyst?**
☐ Yes ☐ No

**Do you have experience working with severe problem behaviour?**
☐ Yes ☐ No

If you answered yes,

Which of the following best describes your experience working to decrease severe problem behaviour?
☐ Never ☐ Limited (less than 5 client cases) ☐ Regularly (more than 5 client cases)
Please describe your average case of severe problem behaviour:

How many years of experience do you have supporting individuals with problem behaviour?
☐ less than one year ☐ 1-2 years ☐ 2-5 years ☐ 5-10 years ☐ 10-15 years ☐ 15-18 years ☐ over 18 years
What has been your involvement in supporting clients with challenging behaviour? (e.g., frontline implementer, assisting behaviour analyst, supervising behaviour analyst)

**How many clients have you supported throughout your tenure as a BCBA?**
☐ 5 or less ☐ 6-10 ☐ 11-15 ☐ 16-20 ☐ 21+

**What age groups have you treated? Check all that apply**
☐ 12 months or younger ☐ 1-2 years ☐ 2-5 years ☐ 5-10 years ☐ 10-15 years ☐ 15-18 years ☐ over 18 years – 60 years ☐ 61+

**What setting best describes your current or past workplace(s) as a BCBA?**
☐ Intervention center (i.e. 1:1 treatment room, play room, lunch room)
☐ In client home (Private home)
☐ Group Day program (i.e. Large room with more than 5 clients, community center)
☐ Group Home (i.e. Common areas with individual overnight bedrooms)
☐ Large supported accommodation residence (e.g., institution)
☐ School classroom (this also includes segregated classrooms if it is considered school property)
☐ Hospital (i.e. Hospital beds, nursing and medical staffing on site, integrated with other typically developing patients)
☐ Psychiatric hospital or ward
☐ Secured unit (e.g., dual diagnosis treatment facility, correctional facilities)
☐ Other (Please specify):

## Appendix C

## Severity Tool

**Client name/ Case number (from case scenario):**

**Instructions: To complete the tool, read each question and select whether the client meets the criteria for options provided. Only select one score for each question. Each question must be answered to complete the tool. Pending the completion of all questions researchers will calculate the overall sum value which will determine the client's rating of severity for their problem behaviour.**

### A. Frequency

| | | | |
|---|---|---|---|
| 1. Direct observation data exists indicating the target behavior occurs…. | 1/day to 1/hour<br>**(0)** | 1/min to 3/min<br>**(1)** | More than 3/min (or > 20% conditions)<br>**(2)** |
| 2. Direct observation data exists indicating the target behavior lasts, uninterrupted (e.g., tantrums) …. | Less than 30 min<br>**(0)** | | More than 30 minutes<br>**(1)** |
| **Frequency section Total:** | | | **/3** |

### B. Chronicity

| | | |
|---|---|---|
| 3. Previous treatment has decreased the target behaviour…. | More than 80%<br>**(0)** | Less than 80%<br>**(2)** |
| 3a. Have more than two different interventions been applied with no decreases in target behaviour? | No<br>**(0)** | Yes<br>**(2)** |
| **Chronicity section Total:** | | **/4** |

### C. Intensity

| | | |
|---|---|---|
| Functional Analysis | | |
| 4. Does the intensity of the target behaviour render conducting an FA extremely unsafe? | No<br>**(0)** | Yes<br>**(2)** |

| Physical Damage | **No** | **Yes** |
|---|---|---|
| 5. Client engages in behaviours that result (has resulted) in tissue damage?<br>*(e.g., swelling, bleeding, bruising, redness)* | **(0)** | **(1)** |
| 5a. Has the target behaviour resulted in tissue damage lasting more than 30 seconds? (e.g., redness on skin remains longer than 30-seconds | **(0)** | **(1)** |
| 6. Is there permanent tissue damage due to target behaviour? | **(0)** | **(2)** |

| | | | |
|---|---|---|---|
| *(e.g., 'cauliflower' ear; damaged vision, scarring)* | | | |
| 6a. Has there been documented physical trauma or observed consistent damage within the last two months? | **(0)** | **(2)** | |
| 6b. Is there physical evidence of healed injuries/lacerations as a result of engaging in the behavior (this can include outward aggression/ property destruction resulting in an injury, i.e., breaking a window causing lacerations; in addition to self-injury)?<br>*(e.g., scarring, toughened and calloused skin)* | **(0)** | **(2)** | |
| 6c. Has the target behaviour resulted in <u>documented</u> staff injury that required time away from work? | **(0)** | **(2)** | |
| 7. Is protective gear worn… | No | When the bx occurs | Ongoing |
| 7a. …by staff? | **(0)** | **(1)** | **(2)** |
| 7b. …by the client? | **(0)** | **(1)** | **(2)** |
| Physical damage subsection total: | | | /4 |

| | | | |
|---|---|---|---|
| Property Damage | | | |
| 8. Has the target behaviour resulted in property damage? | No<br>**(0)** | Yes<br>**(2)** | |
| 8a. **If yes**, what is the estimated value of the destroyed property? | $0 to $750<br>**(0)** | $751-$1500<br>**(1)** | $1500 +<br>**(2)** |
| Property damage subsection total: | | | /4 |
| **Intensity section total:** | | | **/20** |

### D. Legal and Environment Restrictions

| Legal | <u>**No**</u> | <u>**Yes**</u> |
|---|---|---|
| 9. Is there <u>evidence</u> of past legal repercussions due to target behaviour?<br>*(i.e., court document or restraining orders for arrest, freedoms significantly limited due to sexualized behaviour targeting vulnerable populations)* | **(0)** | **(2)** |
| 10. <u>Evidence</u> of previous assault charges because of engaging in problem behaviour (pending or otherwise) | **(0)** | **(1)** |

| | |
|---|---|
| Legal subsection total: | /3 |

| Residence Restrictions | **No** | **Yes** |
|---|---|---|
| 11. Is the client's living space confined... | | |
| 11a. to home or designated space by locks (either to outside or other areas of the home) OR... | **(0)** | **(1)** |
| 11b. others live in the home, however they lock their own living spaces to ensure their own safety? | **(0)** | **(1)** |
| 12. Are speciality building materials required in client residence? *(e.g., reinforced drywall, lexan/safety glass; floor standing stainless steel toilets)* | **(0)** | **(1)** |
| 13. Is 2:1 or more staffing complement required for behaviour management? | **(0)** | **(2)** |
| Residence Restrictions subsection total score: | | /5 |

| Community and Resource Access | **No** | **Yes** |
|---|---|---|
| 14. Is there evidence of past expulsions from one or more learning or residential placements due to challenging behaviour? | **(0)** | **(1)** |
| 15. Has access to day programs been denied due to challenging behaviour? | **(0)** | **(1)** |
| 16. Do healthcare consultations occur at the residence to avoid having to go out into the community? | **(0)** | **(1)** |
| 17. Do community outings occur less than once/month due to probability of challenging behaviour placing community members at risk? (*not including riding in a vehicle wherein the client does not exit the vehicle to enter a community setting*) | **(0)** | **(1)** |
| 18. Are family members or caregivers unable to engage in typical community activities due to the client's problem behaviour? *(i.e., vacations, swimming pool, visiting the library, going out for dinner)* | **(0)** | **(1)** |

| | |
|---|---|
| Community and Resource Access subsection total: | /5 |
| **Legal and Environment Restrictions section total:** | **/13** |

Summaries of Totals

| | |
|---|---|
| Frequency | /3 |
| Chronicity | /4 |
| Intensity | /20 |
| Functional Analysis | /2 |
| Physical Damage | /14 |
| Property Damage | /4 |
| Legal and environmental restrictions | /13 |
| Legal | /3 |
| Residence restrictions | /4 |
| Community and Resource Access | /5 |
| Total | /40 |

## Appendix D

Literature to Support Weighting Distributions

| Tool Item | Relevant Literature |
|---|---|
| 1 | Emerson, E., Kiernan, C., Alborz, A., Reeves, D., Mason, H., Swarbrick, R., Mason, L., & Hatton, C. (2001). The prevalence of challenging behaviours: A total population study. *Research in Developmental Disabilities, 22*(1), 77-93. https://doi.org/10.1016/S0891-4222(00)00061-5<br>Roscoe, E. M., Iwata, B. A., & Zhou, L. (2013). Assessment and treatment of chronic hand mouthing. *Journal of Applied Behavior Analysis, 24*(1), 181-198. https://doi.org/10.1002/jaba.14<br>Banda, D. R., McAfee, J. K., & Hart, S. L. (2012). Decreasing self-injurious behavior and fading self-restraint in a student with autism and Tourette syndrome. *Behavioral Interventions, 27*(3), 164-174. https://doi.org/10.1002/bin.1344 |
| 2 | Emerson, E., Kiernan, C., Alborz, A., Reeves, D., Mason, H., Swarbrick, R., Mason, L., & Hatton, C. (2001). The prevalence of challenging behaviours: A total population study. *Research in Developmental Disabilities, 22(*1), 77-93. https://doi.org/10.1016/S0891-4222(00)00061-5 |
| 3 | Hagopian, L. P., Fisher, W. W., Sullivan, M. T., Acquisto, J., & LeBlanc, L. A. (1998). Effectiveness of functional communication training with and without extinction and punishment: a summary of 21 inpatient cases. *Journal of Applied Behavior Analysis, 31*(2), 211-235. https://doi.org/10.1901/jaba.1998.31-211 |
| 3a | Salvy, S., Mulick, J. A., Butter, E., Bartlett, R. K., & Linscheid, T. R. (2004). Contingent Electric Shock (SIBIS) and a Conditioned Punisher Eliminate Severe Head Banging in a Preschool Child. *Behavioral Interventions, 19*(2), 59–72. https://doi.org/info:doi/<br>Roscoe, E. M., Iwata, B. A., & Zhou, L. (2013). Assessment and treatment of chronic hand mouthing. *Journal of Applied Behavior Analysis, 46*(1), 181-198. https://doi.org/10.1002/jaba.14 |
| 4 | Fahmie, T. A. & Iwata, B. A. (2011). Topographical and functional properties of precursors to severe problem behavior. Journal of Applied Behavior Analysis, 44(4), 993-997. https://doi.org/10.1901/jaba.2011.44-993<br>Roscoe, E. M., Iwata, B. A., & Zhou, L. (2013). Assessment and treatment of chronic hand mouthing. *Journal of Applied Behavior Analysis, 46*(1), 181-198. https://doi.org/10.1002/jaba.14 |
| 5 | Rojahn, J. Rowe, E. W., Sharber, A. C., Hastings, R., Matson, J. L., Didden, R., Kroes, D. B. H., & Dumont, E. L. M. (2012). The behavior problems inventory-short form for individuals with intellectual disabilities: part I: development and provisional clinical reference data. *Journal of Intellectual Disability Research, 56*(5), 527-545. https://doi.org/10.1111/j.1365-2788.2011.01507.x<br>Roscoe, E. M., Iwata, B. A., & Zhou, L. (2013). Assessment and treatment of chronic hand mouthing. *Journal of Applied Behavior Analysis, 46*(1), 181-198. https://doi.org/10.1002/jaba.14 |

| | |
|---|---|
| 5a | Rojahn, J. Rowe, E. W., Sharber, A. C., Hastings, R., Matson, J. L., Didden, R., Kroes, D. B. H., & Dumont, E. L. M. (2012). The behavior problems inventory-short form for individuals with intellectual disabilities: part I: development and provisional clinical reference data. *Journal of Intellectual Disability Research, 56*(5), 527-545. https://doi.org/10.1111/j.1365-2788.2011.01507.x<br>Roscoe, E. M., Iwata, B. A., & Zhou, L. (2013). Assessment and treatment of chronic hand mouthing. *Journal of Applied Behavior Analysis, 46*(1), 181-198. https://doi.org/10.1002/jaba.14 |
| 6 | Rojahn, J. Rowe, E. W., Sharber, A. C., Hastings, R., Matson, J. L., Didden, R., Kroes, D. B. H., & Dumont, E. L. M. (2012). The behavior problems inventory-short form for individuals with intellectual disabilities: part I: development and provisional clinical reference data. *Journal of Intellectual Disability Research, 56(*5), 527-545. https://doi.org/10.1111/j.1365-2788.2011.01507.x<br>Roscoe, E. M., Iwata, B. A., & Zhou, L. (2013). Assessment and treatment of chronic hand mouthing. *Journal of Applied Behavior Analysis, 46(*1), 181-198. https://doi.org/10.1002/jaba.14 |
| 6a | Luiselli, J. K. (2009). Physical restraint of people with intellectual disability: A review of implementation reduction and elimination procedures. *Journal of Applied Research in Intellectual Disabilities, 22*(2), 126-134. https://doi.org/10.1111/j.1468-3148.2008.00479.x |
| 6b | Foxx, R. M. (2003). The treatment of dangerous behavior. *Behavioral Interventions, 18*(1), 1-21. https://doi.org/10.1002/bin.127 |
| 6c | Evers, C., & Pilling, N. (2012). Improving outcomes for people with severe challenging behaviour. *Learning Disability Practice, 15*(9), 30-36. https://doi.org/10.7748/ldp2012.11.15.9.30.c9382 |
| 7a | Newcomb, E. T., & Hagopian, L. P. (2018). Treatment of severe problem behaviour in children with autism spectrum disorder and intellectual disabilities. *International Review of Psychiatry, 30*(1), 96-109. https://doi.org/10.1080/09540261.2018.1435513 |
| 7b | Banda, D. R., McAfee, J. K., & Hart, S. L. (2012). Decreasing self-injurious behavior and fading self-restraint in a student with autism and Tourette syndrome. Behavioral Interventions, 27(3), 164-174. https://doi.org/10.1002/bin.1344 |
| 8 | Rojahn, J. Rowe, E. W., Sharber, A. C., Hastings, R., Matson, J. L., Didden, R., Kroes, D. B. H., & Dumont, E. L. M. (2012). The behavior problems inventory-short form for individuals with intellectual disabilities: part I: development and provisional clinical reference data. Journal of Intellectual Disability Research, 56(5), 527-545. https://doi.org/10.1111/j.1365-2788.2011.01507.x |
| 8a | Rojahn, J. Rowe, E. W., Sharber, A. C., Hastings, R., Matson, J. L., Didden, R., Kroes, D. B. H., & Dumont, E. L. M. (2012). The behavior problems inventory-short form for individuals with intellectual disabilities: part I: development and provisional clinical reference data. Journal of Intellectual Disability Research, 56(5), 527-545. https://doi.org/10.1111/j.1365-2788.2011.01507.x |

| | |
|---|---|
| 9 | Evers, C., & Pilling, N. (2012). Improving outcomes for people with severe challenging behaviour. *Learning Disability Practice, 15*(9), 30-36. https://doi.org/10.7748/ldp2012.11.15.9.30.c9382 |
| 10 | Murphy, G. (2009). Challenging behavior: A barrier to inclusion? *Journal of Policy and Practice in Intellectual Disabilities, 6*(2), 89-90. https://doi.org/10.1111/j.1741-1130.2009.00216.x |
| 11a | Murphy, G. (2009). Challenging behavior: A barrier to inclusion? *Journal of Policy and Practice in Intellectual Disabilities, 6*(2), 89-90. https://doi.org/10.1111/j.1741-1130.2009.00216.x |
| 11b | Murphy, G. (2009). Challenging behavior: A barrier to inclusion? *Journal of Policy and Practice in Intellectual Disabilities, 6*(2), 89-90. https://doi.org/10.1111/j.1741-1130.2009.00216.x |
| 12 | Murphy, G. (2009). Challenging behavior: A barrier to inclusion? *Journal of Policy and Practice in Intellectual Disabilities, 6*(2), 89-90. https://doi.org/10.1111/j.1741-1130.2009.00216.x |
| 13 | Lowe, K., Allen, D., Jones, E., Brophy, S., Moore, K., & James, W. (2007). Challenging behaviours: prevalence and topographies. Journal of Intellectual Disability Research, 51, 625-636. https://doi.org/10.1111/j.1365-2788.2006.00948.x<br>Emerson, E., Kiernan, C., Alborz, A., Reeves, D., Mason, H., Swarbrick, R., Mason, L., & Hatton, C. (2001). The prevalence of challenging behaviours: A total population study. Research in Developmental Disabilities, 22(1), 77-93. https://doi.org/10.1016/S0891-4222(00)00061-5 |
| 14 | Foxx, R. M. (2003). The treatment of dangerous behavior. *Behavioral Interventions, 18*(1), 1-21. https://doi.org/10.1002/bin.127 |
| 15 | Murphy, G. (2009). Challenging behavior: A barrier to inclusion? *Journal of Policy and Practice in Intellectual Disabilities, 6*(2), 89-90. https://doi.org/10.1111/j.1741-1130.2009.00216.x |
| 16 | Newcomb, E. T., & Hagopian, L. P. (2018). Treatment of severe problem behaviour in children with autism spectrum disorder and intellectual disabilities. *International Review of Psychiatry, 30*(1), 96-109. https://doi.org/10.1080/09540261.2018.1435513<br>White, D. A., & Dodder, R. A. (2000). The relationship of adaptive and maladaptive behavior to social outcomes for individuals with developmental disabilities. *Disability and Society, 15*(6), 897-908. https://doi.org/10.1080/713662014 |
| 17 | Murphy, G. (2009). Challenging behavior: A barrier to inclusion? *Journal of Policy and Practice in Intellectual Disabilities, 6*(2), 89-90. https://doi.org/10.1111/j.1741-1130.2009.00216.x<br>Newcomb, E. T., & Hagopian, L. P. (2018). Treatment of severe problem behaviour in children with autism spectrum disorder and intellectual disabilities. *International Review of Psychiatry, 30*(1), 96-109. https://doi.org/10.1080/09540261.2018.1435513 |

| | |
|---|---|
| 18 | Murphy, G. (2009). Challenging behavior: A barrier to inclusion? *Journal of Policy and Practice in Intellectual Disabilities, 6*(2), 89-90. https://doi.org/10.1111/j.1741-1130.2009.00216.x<br>Newcomb, E. T., & Hagopian, L. P. (2018). Treatment of severe problem behaviour in children with autism spectrum disorder and intellectual disabilities. *International Review of Psychiatry, 30*(1), 96-109. https://doi.org/10.1080/09540261.2018.1435513 |

## Appendix E

## Social Validity Questionnaire

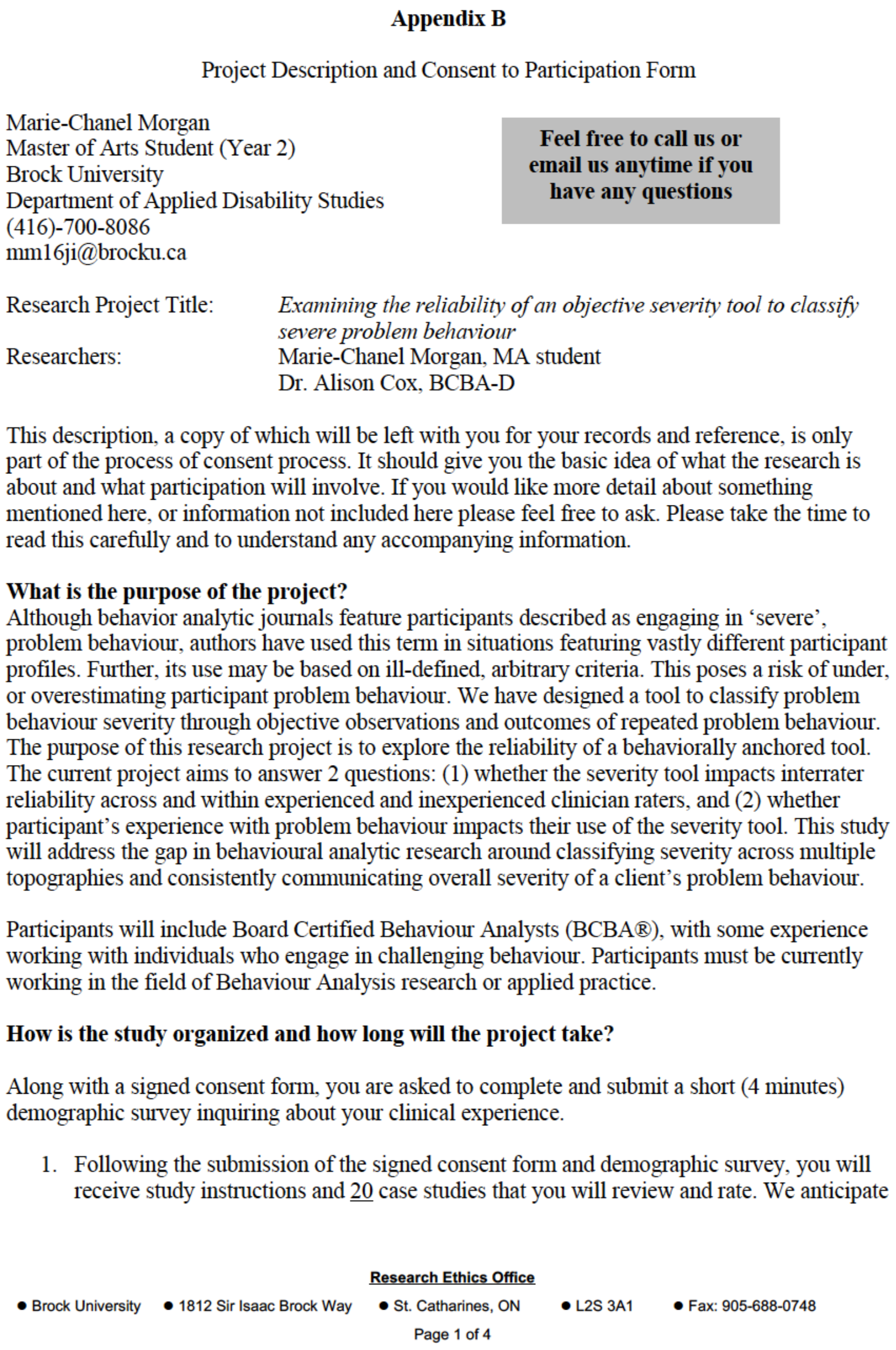
Participant number: __________________

Severity Tool Evaluation Form

(Post-Likert 1-5 & open ended)

| | |
|---|---|
| 1) This tool was easy to use. | ☐ Strongly agree 1<br>☐ Agree 2<br>☐ Neutral 3<br>☐ Disagree 4<br>☐ Strongly disagree 5 |
| 2) This tool could be easily applied to categorize/classify severity using information provided in published participant descriptions. | ☐ Strongly agree 1<br>☐ Agree 2<br>☐ Neutral 3<br>☐ Disagree 4<br>☐ Strongly disagree 5 |
| 3) This tool could be incorporated into publication practices to promote the consistent application of the term ‘severe’ across articles. | ☐ Strongly agree 1<br>☐ Agree 2<br>☐ Neutral 3<br>☐ Disagree 4<br>☐ Strongly disagree 5 |
| 4) This tool would be useful to achieve an objective measurement of severity. | ☐ Strongly agree 1<br>☐ Agree 2<br>☐ Neutral 3<br>☐ Disagree 4<br>☐ Strongly disagree 5 |
| 5) The questions in this tool are applicable to the severity of problem behaviour. | ☐ Strongly agree 1<br>☐ Agree 2<br>☐ Neutral 3<br>☐ Disagree 4<br>☐ Strongly disagree 5 |
| 6) Are there some participant characteristics that you feel are missing, that would improve the tool? | |

# Appendix F



## Appendix B

Project Description and Consent to Participation Form

Marie-Chanel Morgan
Master of Arts Student (Year 2)
Brock University
Department of Applied Disability Studies
(416)-700-8086
mm16ji@brocku.ca

**Feel free to call us or email us anytime if you have any questions**

Research Project Title: *Examining the reliability of an objective severity tool to classify severe problem behaviour*

Researchers: Marie-Chanel Morgan, MA student
Dr. Alison Cox, BCBA-D

This description, a copy of which will be left with you for your records and reference, is only part of the process of consent process. It should give you the basic idea of what the research is about and what participation will involve. If you would like more detail about something mentioned here, or information not included here please feel free to ask. Please take the time to read this carefully and to understand any accompanying information.

**What is the purpose of the project?**

Although behavior analytic journals feature participants described as engaging in 'severe', problem behaviour, authors have used this term in situations featuring vastly different participant profiles. Further, its use may be based on ill-defined, arbitrary criteria. This poses a risk of under, or overestimating participant problem behaviour. We have designed a tool to classify problem behaviour severity through objective observations and outcomes of repeated problem behaviour. The purpose of this research project is to explore the reliability of a behaviorally anchored tool. The current project aims to answer 2 questions: (1) whether the severity tool impacts interrater reliability across and within experienced and inexperienced clinician raters, and (2) whether participant's experience with problem behaviour impacts their use of the severity tool. This study will address the gap in behavioural analytic research around classifying severity across multiple topographies and consistently communicating overall severity of a client's problem behaviour.

Participants will include Board Certified Behaviour Analysts (BCBA®), with some experience working with individuals who engage in challenging behaviour. Participants must be currently working in the field of Behaviour Analysis research or applied practice.

**How is the study organized and how long will the project take?**

Along with a signed consent form, you are asked to complete and submit a short (4 minutes) demographic survey inquiring about your clinical experience.

1. Following the submission of the signed consent form and demographic survey, you will receive study instructions and <u>20</u> case studies that you will review and rate. We anticipate

this task will take less than 1 hour. We ask participants to submit their completed cases within 3 weeks of receiving them.

2. Following the return of the first 20 cases, you will receive the severity tool and an additional 20 case studies to review and rate. We anticipate this task will take approximately 2 hours. We ask participants to submit their completed cases within 3 weeks of receiving the email.
3. When you return the final 30 cases, we will send you a link to complete a short social validity questionnaire (6 questions) which will take approximately 3 minutes to complete.

Total participation time is approximately 3.25 hours over six weeks.

**Will my personal information be kept confidential?**

All information that we obtain about you including: 1.) relevant participant demographics; 2.) contact information; 3.) scenario ratings; 4.) BACB® certification number; and 5.) completed social validity questionnaire, will be kept confidential and stored in a locked office or on an encrypted electronic database. The research staff directly involved in the study will have access to study data. Any presentation, reports, or publications about the project will reflect group performance, or anonymized individual results – if required, and conclusions drawn about the utility of the tool.

**What are the risks taking part in this study?**

There are no risks for participating in this study. We encourage you to provide feedback when completing the participate evaluation form on the tools' utility.

**What are the potential benefits of participating in this study?**

Your participation in this project will help us begin exploring the utility of a severity tool informed by objective client characteristics. Your participation may lead to greater consistency in use of the term 'severe' across the behaviour analytic literature.

**Is there any cost for participating?**

There is no cost for participating.

**Is participation voluntary?**

Participation is voluntary. You can decline to participate at any time. Please note that there is no penalty for withdrawing from the study. If you choose to withdraw from the study, you can notify the researchers via e-mail. We will assume you have withdrawn from the study if: 1.) you do not respond to three consecutive reminder emails issued two weeks apart, or 2) you do not return rated scenarios, after being issued emailed reminders. We ask that you let us know if you will need more than three weeks to complete the scenarios. If you do choose to withdraw, we may ask you for permission to use any data previously collected, at which point, you may choose to consent or decline to the use of your data.

**How and to whom will the research results be shared?**

Summary results will be used to meet the requirements of a master's level thesis as well as disseminated at conferences, other presentations for educational purposes, and in a scientific journal. A nontechnical summary report will be shared on the researchers Brock website and emailed to each participant upon study completion. No disseminated results will contain any of your identifying information.

**When will I receive the results of this project?**

Within four months of the completion of this study (approx. January 2021), we will email all participants a summary of the results as well as a link to results posted on the researcher's website. You can also contact Dr. Cox or Marie-Chanel for more information about this project at any time.

**Signing the Consent Forms**

Signing the following pages of this *Project Description and Consent Form* indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate. You are free to withdraw from the project at any time, and/or refrain from answering any questions you prefer to omit, without prejudice or consequence. Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation.
Student Investigator: Marie-Chanel Morgan 416-700-8086
Faculty Advisor: Dr. Alison Cox: 905-668-5550 Ext: 3949

The Social Science Research Ethics Board at Brock University has cleared this research project (file # 19-083). If you have any concerns or complaints about this project, you may contact any of the above-named persons, or directly contact the Research Ethics Office at Brock University reb@brocku.ca, or by phone: (905) 668-5550 Ext: 3035. A copy of this Project Description and Consent Form has been given to you to keep for your records and reference.

I, ____________________________________________ (print and sign your name)

hereby consent to participate in the project, entitled "*Examining the reliability of an objective severity tool to classify severe problem behaviour*"

By giving consent I allow the research project staff to:

- Score my submissions and include the results in publications, reports, and talks, so that others may learn from this project. My identity, however, <u>will not</u> be disclosed.
- Use my demographic survey and social validity responses for study purposes
- Retain my name, professional email address and BACB certificant number to verify certification.
- I understand that I can revoke or amend this consent at any time for any reason.

| *Please check YES or NO for the following items:* | **YES** | **NO** |
|---|---|---|
| I would like to receive the results of this project after it is completed (approx. Jan 2021). I would prefer that the researchers contact me by (check):<br>□Email ________________ □Phone ________________ | □ | □ |
| I give permission for the researchers to retain my name, professional email address, and BACB certificant number until the submission of the study, | □ | □ |
| I give permission for the researchers to retain my data, without personal identifiers, up to 5 years after study completion. | □ | □ |
| I give permission for the researchers to use my data for other types of related studies (secondary use of data).<br>Participant Signature: ______________________________ | □ | □ |
| The researchers may contact me directly for possible future related studies. | □ | □ |

**BACB Certificant #**______________________________________________

**Email:** ______________________________________________

**Phone 1:** ______________________________ **Preferable hours:**__________

**Date:** ______________________________

____________________ ____________________ __________

Name of researcher/delegate Signature of researcher/delegate Date