Computers in Human Behavior 114 (2021) 106534

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: http://www.elsevier.com/locate/comphumbeh

Full length article

The Measure of Online Disinhibition (MOD): Assessing perceptions of reductions in restraint in the online environment

Jaimee Stuart *, Riley Scott

School of Applied Psychology, Griffith University, Australia

ARTICLE INFO

Keywords:
Online disinhibition
Internet
Social media
Assessment
Individual differences
Scale development

ABSTRACT

Online disinhibition, or the experience of diminishing constraints when online, has important influences on behavior, yet theoretically robust, reliable, and valid measures of this construct are lacking. This research developed a new Measure of Online Disinhibition (MOD). In study 1, 403 participants were split into two samples; n = 212 were subject to exploratory factor analysis, and n = 191 to confirmatory factor analysis. The final 12 items loaded onto a single factor with high reliability and construct validity among a range of measures (toxic and benign disinhibition, time online, false self, online self-disclosure, and trolling). In Study 2, using a distinct sample (N = 242), the MOD was again confirmed and the nomological network was extended to examine cyberbullying and well-being. Additionally, in both studies path models were tested to explore the mediation of time online on positive and negative indicators via MOD. Results found that greater time online was associated with increases in both positive and negative cyber behaviors but decreased well-being via increases in MOD. The MOD operationalizes online disinhibition in a theoretically driven fashion, allowing researchers to build upon our understanding of the impacts of the online environment on human behavior in a systematic way.

1. Introduction

Individual behavior in social settings is governed by rules and norms that designate what are acceptable and unacceptable ways of acting (Litt & Stock, 2011; Rimal & Real, 2005). As a result of self-consciousness, concern over self-presentation and social evaluation, and to protect against judgement from others, most people experience some degree of inhibition when in social contexts, where behaviors are constrained to align with perceived social norms (Joinson, 1998). Whereas inhibition is characterized by restraint and control of behavior, disinhibition, in contrast, is characterized by reductions in discipline and control, as well as disregard for, or violation of, social norms (Zuckerman, 1979). Disinhibition may be manifested in behavioral, emotional, and cognitive domains of functioning, and has been found to be likely to occur when individual or environmental factors act to diminish or remove the social and behavioral constraints that are commonly kept in place to ensure norms are upheld (Starkstein & Robinson, 1997). An emerging body of research has found that the internet comprises of several unique features that can promote disinhibition, such that people may act, think, and feel differently online when compared to face-to-face interactions. Suler (2004) labelled this phenomenon the 'Online Disinhibition Effect'.

Suler (2004) proposed that online disinhibition primarily manifests in the dissipation of personal restraints in digital contexts as a result of six core factors; dissociative anonymity, meaning that online environments provide potential for identity concealment; invisibility, meaning it is possible to not be directly seen or observed online; asynchronicity, meaning that online interactions need not be performed on a real-time basis; solipsistic introjection, meaning that interactions with others online may be played out as internalized narratives that have little objective reality; dissociative imagination, meaning that online our subjective characterizations of ourselves and others are distinct from in person interactions; and the minimization of status and authority, meaning that online settings offer the ability for all individuals to be presented equally to one another. Effectively, Suler proposed that these features of the digital environment create distinct interpersonal contexts where personal identity can be concealed, communication does not require real-time, face-to-face interaction, subjectivities can be distorted, and where the social rules, responsibilities, and hierarchies of offline contexts often do not apply.

Suler (2004) proposed that the experience of some or all of these conditions affects online relationships and can foster online disinhibition, which operates in two opposing ways. Firstly, 'benign' online

* Corresponding author. School of Applied Psychology, Griffith University, Mount Gravatt, QLD, 4111, Australia.
E-mail address: j.stuart@griffith.edu.au (J. Stuart).

https://doi.org/10.1016/j.chb.2020.106534
Received 19 March 2020; Received in revised form 10 August 2020; Accepted 27 August 2020
Available online 31 August 2020

disinhibition results when the effects and outcomes of lowered inhibitions are positive, such as individuals showing uncharacteristic or heightened kindness, support and generosity, or being more likely to self-disclose and share personal information, thoughts, and feelings online. Prosocial behaviors such as providing compliments, defending others, and giving more to charitable organizations are also acts characteristic of benign disinhibition (Lapidot-Lefler & Barak, 2015). Conversely, use of rude and threatening language, making mean comments, and behaviors such as flaming (acting out to damage another's or one's own image) or trolling (malicious online behavior, intended to aggravate, annoy or disrupt others) are typical of 'toxic' online disinhibition (Suler, 2004). Toxic online disinhibition additionally encompasses behaviors such as using derogatory names, making sexually inappropriate comments, or engaging with explicitly violent or sexual online materials (Lapidot-Lefler and Barak, 2012).

Despite original theorizing that online disinhibition can be associated with both positive and negative interpersonal outcomes, the research concerning this effect has predominantly focused on the association of online disinhibition with harmful digital behaviors, specifically cyber aggression (e.g., Varjas, Talley, Meyers, Parris, & Cutts, 2010; Wright, Harper, & Wachs, 2019). The growing body of research concerning online disinhibition and aggression proposes that one of the key reasons antagonistic behaviors are more common online is because of the visual, emotional, and physical distance aggressors have from their victims which may limit the ability to empathize with others (e.g. Lapidot-Lefler & Barak, 2012; Udris, 2014; Varjas et al., 2010; Voggeser, Singh, & Göritz, 2018). Indeed, the research of Wachs and colleagues (Wachs & Wright, 2018; 2019; Wachs et al., 2019) found that online disinhibition not only has direct negative effects on harmful online behaviors in the form of cyber-hate, but that it also exacerbates the influence of other factors (cyberbullying perpetration and victimization, and exposure to cyber-hate) on perpetration of cyber-hate. In contrast, very little is known about what produces positive outcomes as a result of online disinhibition, although a recent meta-analysis (Clark-Gordon, Bowman, Goodboy, & Wright, 2019) that treats online self-disclosure as analogous to benign disinhibition found that anonymity had a positive average correlation with online self-disclosure. Furthermore, self-disclosing online has been found to have beneficial effects, with research finding that it is associated with higher levels of social well-being (Ko & Kuo, 2009; Varnali & Toker, 2015). Given the importance of emerging results in this area, as well as the growing pervasiveness and associated societal challenges of technology use, the aim of this study is to define and operationalize online disinhibition and to propose a new measure of this construct in order to develop the field in a more systematic way.

1.1. Defining, operationalizing, and measuring online disinhibition

Although the research to date concerning online disinhibition is compelling and has extended our knowledge of the social and psychological impacts of digital contexts, there are notable issues in the operationalization and measurement of the construct itself. Firstly, existing measures often conflate the theorized antecedents with the experience of disinhibition. Specifically, the terms "toxic" and "benign" disinhibition were not originally used by Suler (2004) to assess different kinds of disinhibition, but rather to underscore that the unique attributes of the online environment "can work in two seemingly opposing directions" (p. 321) to produce positive, neutral, negative, and mixed outcomes. However, to date, most research has confused the outcomes of disinhibition with the construct itself by embedding hedonic tone into the measurement of the construct. As a case in point, the Online Disinhibition Scale from Udris (2014), which was developed with the purpose of investigating the role of online disinhibition in predicting cyberbullying, comprises of two factors measuring toxic and benign online disinhibition. This measure is currently the most common method of assessing online disinhibition and has been used with international samples of both adolescents and adults (e.g., Barlett & Helmstetter, 2018; Kim & Chang, 2017; Lai & Tsai, 2016, Wachs & Wright, 2018, 2019; Wachs, Wright, & Vazsonyi, 2019). Research of (Lapidot-Lefler and Barak 2012; 2015) also examined toxic and benign online disinhibition, but instead of self-report, used observational methods among a group of chat room users who were provided with vignettes concerning social dilemmas in online settings.

As stated by Suler (2004, p. 322) "Cultural relativity as well as the complexities of psychological dynamics will blur any simple contrasts between disinhibition that is positive or negative". Indeed, it is argued that online disinhibition itself is not inherently good nor bad, as it represents a phenomenon where individuals experience acting, thinking, or feeling differently online when compared to face-to-face interactions. Therefore, online disinhibition should be understood as an observed or perceived intraindividual difference in online as compared to offline tendencies. Such valence free ways of measuring online disinhibition that distinguish between the experience and outcomes of disinhibition are available in the literature. Examples include a three-item measure by Schouten, Valkenburg, and Peter (2007) which was developed to examine disinhibited self-disclosure within instant messaging contexts (also used by Casale, Fiovaranti, & Caplan, 2015; Weidman et al., 2012) and a five item measure developed by Kurek, Jose, and Stuart (2019). Additionally, the research of Antoniadou and colleagues (Antoniadou, Kokkinos, & Fanti, 2019; Antoniadou, Kokkinos, & Markos, 2019) utilizes the social confidence and socially liberating factors from the Internet Behavior and Attitudes Scale (Morahan-Martin & Schumacher, 2000) to measure the experience of disinhibition online. However, these measures are also not without issues in that they are quite narrow in their focus, have been developed and used predominantly with youth samples, have limited validation, and may conflate online disinhibition with its antecedents, as will be discussed subsequently.

Indeed, another common issue with most previous measures is that they fail to separate the precursors from the experience of online disinhibition. Two examples of this from the Online Disinhibition Scale are; "It is easier to communicate online because you can reply any time you like" (benign disinhibition) and "It is easy to write insulting things online because there are no repercussions" (toxic disinhibition; Udris, 2014). Additional examples include "The anonymity of the online environment influences the way I express myself online" (Kurek et al., 2019) and "The anonymity of being online is liberating" (Morahan-Martin & Schumacher, 2000). Each of these items conflate the experience of disinhibition with the features of the online context which are suggested to produce disinhibitory effects. Suler's main aim was to address the question of "What elements of cyberspace lead to this weakening of the psychological barriers that block hidden feelings and needs?" (2004, p. 322), and in answering this question the six internet attributes (dissociative anonymity; invisibility; asynchronicity; solipsistic introjection; dissociative imagination and the minimization of status and authority) were proposed to be antecedents of online disinhibition. Yet some research has suggested that online disinhibition is a multidimensional latent construct which itself comprises of these six key dimensions (Cheung, Wong, & Chan, 2016; Wu, Lin, & Shih, 2017), whereas most others have simply failed to adequately distinguish the features of the digital context from the experience of disinhibition.

Recent reviews on the impacts of social media on the behavior of young people conducted by Nesi, Choukas-Bradley, and Prinstein (2018a; 2018b) suggests that the internet can be understood as a type of "transformation framework" where relationships and interactions are uniquely constructed in line with the affordances of the digital environment. The transformation framework outlines seven features of social media that influence experiences of young people, namely asynchronicity, permanence, publicness, availability, cue absence, quantifiability, and visualness. Effectively, this recent consolidation of the research suggests, very similarly to early theorizing of Suler (2004), that features of the digital environment create distinct interpersonal

contexts that transform or influence the experiences of people online. We argue that the work of Nesi et al. (2018a; 2018b) supports the notion that the features of the internet are distinct from the psychological phenomena of online disinhibition, and therefore, assessments of disinhibition should be disambiguated from the features of the internet as well as from the potential outcomes of online disinhibition.

1.2. The current research

The prevalence and ubiquitous use of internet connected devices in everyday life means that it is becoming increasingly important to assess how digital contexts influence social and personal outcomes. Research into online disinhibition holds particular promise in helping us to understand the blurred lines between online and offline contexts and how these may impact on psychology in an increasingly digitally connected world. However, to date, measures of the construct have not been well operationalized or broadly validated, which limits our ability to further the research in this area in a robust and generalizable way. The primary objectives of this research are, therefore, to: (1) to define and operationalize online disinhibition, (2) to develop a reliable instrument measuring online disinhibition, (3) to validate this measure, and (4) to begin to explore the relationship between internet use, online disinhibition, and both social and psychological outcomes. The present research draws upon on two studies with two unique samples; a scale development and validation study comprised of exploratory and confirmatory sub-samples, and a validation study with a sample of adults. As such, Study 1 addresses the development and construct validation of the instrument as well as examining the predictors and outcomes of online disinhibition, and Study 2 seeks to further validate the measure, to advance the nomological network of online disinhibition, and to further examine predictors and outcomes of online disinhibition.

1.3. Study 1: Scale construction

Study 1 aims to construct and validate a new assessment tool, the Measure of Online Disinhibition (MOD). Following the past literature and based on Suler's (2004) theorizing, our definition of online disinhibition is *the perception or experience of reductions in restraint in the online environment such that individuals may act, think, and feel differently online when compared to face-to-face interactions*. Thus, we seek to develop a measure that is distinct from its antecedents and potential outcomes, as well as embedded in a comparative paradigm such that intraindividual online are relative to offline (or in-person) tendencies.

In order to validate the MOD, a variety of indicators of were examined. Firstly, the convergent validity of the new measure was tested with the two subscales of the Online Disinhibition Scale, which is the most commonly used measure currently used in the research (Udris, 2014). It was expected that the MOD would have positive and significant associations with both the benign and toxic subscales of the Online Disinhibition Scale. We measured convergent validity of the MOD with presentation of false self online, where it was suggested that greater endorsement of behaving or feeling differently online as compared to offline is likely to be positively associated with self-reported inauthentic displays of identity online. We tested whether the MOD was related to amount of time spent online and frequency of social media use, where it was expected that higher users of the internet and social media have greater opportunities for interaction online, and thus would experience higher levels of online disinhibition. Additionally, predictive validity was assessed through hypothesized associations with what are considered to be positive and negative cyber behaviors associated with online disinhibition; namely the amount of self-disclosure online and engagement in trolling behaviors. As such, the construct validation of the MOD was tested by significant positive associations with benign and toxic online disinhibition, online false self, time spent online, frequency of social media engagement, online self-disclosure, and trolling.

Lastly, we sought to explore whether online disinhibition mediates the effect of time spent online on trolling and online self-disclosure. It has been suggested that online disinhibition is a frequent outcome of high levels of internet use (Niemz, Griffiths, & Banyard, 2005; Peter, Valkenburg, & Schouten, 2007). In fact, a number of studies have found associations between excessive internet use and greater online disinhibition (Armstrong, Phillips, & Saling, 2000; Casale et al., 2015; Morahan-Martin & Schumacher, 2000; Niemz et al., 2005). Further, given that greater time spent online affords greater exposure to contexts of interaction with others online, it is not surprising that greater time spent online is associated with increased engagement in both negative and positive cyber behaviors (e.g., Clark-Gordon et al., 2019; Thacker & Griffiths, 2012). Yet there has been no research to date that has sought to examine whether online disinhibition mediates the effect of time spent online on behavior in the online environment. Indeed, while online disinhibition has been treated as both a mediator (of the effect of personality and false self; Kurek et al., 2019) as well as a moderator (of cyberbullying and cyber-hate; Wachs & Wright, 2018, 2019; Wachs et al., 2019) on negative cyber behaviors, the relatively obvious relationship between time spent online, disinhibition and cyber behaviors have not been examined. We hypothesize that time spent online will be positively associated, both directly and indirectly via online disinhibition, with trolling and online self disclosure.

2. Method

2.1. Construction of the Scale

Following a review of the literature, in order to generate relevant, ecologically valid items for the new Measure of Online Disinhibition (MOD), eight focus group discussions ($N = 20$) were conducted in which participants were asked about their current internet use, how they think, feel and act online, and about their perceptions of the online environment. Because much of the research surrounding online disinhibition has been conducted with adolescent samples, a diverse age range (min = 18 years, max = 39 years; $M = 25.70$, SD = 5.56), and equal gender split was recruited for this component of the research (60%; $n = 12$ females). Example discussion questions included "Do you act or feel differently online compared to how you do in real-life?", and "What makes the internet or online environments different to the offline world?" The focus group discussions were recorded, and content was used to generate a pool of 74 items relating to online disinhibition. Finally, in order to reduce and consolidate the item pool, nine workshops were run with 38 participants in total; 63.2% female, age range from 17 to 67 years, with an average age of 29.44 years ($SD = 11.34$). As part of a broader set of activities, in the workshops, participants were provided with the initial pool of 74 items and were instructed to identify and exclude repetitive or confusing items. Information across the nine workshops was collated and the pool was reduced and clarified to 28 items based on the overall feedback. The use of workshops for item reduction and consolidation, as opposed to item selection by the research team, was important in the development of an ecologically set of items for the scale development study.

2.2. Procedure and participants

Participants were invited to complete a confidential online survey through Mechanical Turk (MTurk). To participate in the study, respondents were required to be 18 years of age or older and be a resident of the United States. Respondents were given a small monetary credit which was distributed via MTurk for completing the entire survey. The survey was prefaced by an information sheet, specifying that: participants could not be identified by the researchers; completion of the questionnaire indicated informed consent; and participants could withdraw from the research at any time until their survey responses had been submitted. The research was approved under the delegated authority of the University's Human Ethics Committee.

In total, 403 individuals who met the inclusion criteria completed the online survey. Participants were aged between 18 and 70 (M = 35.06 years, SD = 11) and 49% were female. The majority were employed either full-time or part-time (77.2%) with fewer respondents self-employed (14.9%), studying (6.2%), or unemployed (5.2%). The sample was also relatively highly educated, with 70.6% having an undergraduate degree or higher. Participants identified predominantly as White (Caucasian; 75.7%), followed by African American (9.7%), Asian (8.2%), and Hispanic (7.4%), with 4.4% identifying with other ethnicities (e.g., Native American, Pacific Islander, Middle Eastern). A random filter was applied in order to identify approximately 50% of the overall dataset for the subsequent sub-sample analyses. This resulted in an exploratory sub-sample (n = 212) and a confirmatory sub-sample (n = 191), both of which shared demographic characteristics analogous to the overall sample.

2.3. Materials

2.3.1. Measure of online disinhibition (MOD)

The pool of 28 items developed and refined from the workshops and focus groups were included in the survey. The items broadly assess self-perceptions of psychological and behavioral change in the online as compared to the offline environment. For example, "My behaviors online are less restricted than in person", "I am more expressive online than I am offline", and "I am more competitive online than I am offline" (see Tables 1 and 2 for the reduced item pool and final scale items respectively). Participants were asked to assess how much each item was representative of themselves on a rating scale from 1 = not at all like me, to 5 = very like me.

2.3.2. Toxic and benign online disinhibition

The subscales of Udris (2014) were included to examine the convergent validity of the new MOD scale with the most commonly used instrument assessing online disinhibition. The Benign Online Disinhibition Scale included seven items that broadly measured respondents' tendency for openness online and were closely aligned to the six antecedents of online disinhibition outlined by Suler (2004). For example, "It is easier to write things online that would be hard to say in real life because you don't see the other's face", and "I have an image of the other person in my head when I read their e-mail or messages online".

Table 1
Study 1. Exploratory factor analyses results and descriptive statistics of items (N = 212).

Item	Factor	M	*SD*
1. I am more confident online than I am offline	0.890	3.14	1.44
2. I am more able to discuss controversial issues online than I am in person	0.854	3.11	1.42
3. I am more expressive online than I am offline	0.850	3.09	1.43
4. My behaviours online are less restricted than in person	0.836	3.00	1.43
5. I am more outgoing online than I am offline	0.825	3.09	1.40
6. I find it easier to express myself on the internet than in a face-to-face conversation*	0.822	3.06	1.45
7. I am more assertive online than I am offline	0.819	3.19	1.47
8. I say things on the internet that I would not say in person	0.794	2.78	1.47
9. I act tougher on the internet than I do face-to-face	0.793	2.75	1.50
10. I make friends more easily online than I do offline	0.785	3.00	1.45
11. I act differently online than I do offline	0.782	2.94	1.37
12. I find communicating with others easier on the internet than in person*	0.763	3.22	1.44
13. I am more competitive online than I am offline	0.760	3.00	1.45
14. I am less cautious about what I say online than about what I say in person	0.750	2.93	1.48
15. The way I talk to people online is different than how I talk offline*	0.750	3.05	1.36

Note * indicates items removed after confirmatory factor analysis.

Four items, such as "Writing insulting things online is not bullying" were included to measure Toxic Online Disinhibition. Responses were recorded along a 5-point agreement scale, from 1 = strongly disagree, to 5 = strongly agree.

2.3.3. Online false self

The false self (deception) four item subscale from The Self-Presentation on Facebook Questionnaire was used in order to assess inauthentic presentations of self online (Michikyan, Subrahmanyam, & Dennis, 2014). Items were adapted in the current study to refer to social networking sites more broadly as compared to the original that focused on Facebook. For example, "I post information about myself on my social networking site profiles that is not true". Respondents were asked to rate their level of agreement with each statement on a 5-point scale ranging from 1 = strongly disagree to 5 = strongly agree.

2.3.4. Internet use

The amount of leisure time spent using the internet was assessed by computing the mean of two items: "How many hours per week day on average do you spend on the internet for leisure?" and "How many hours per weekend day on average do you spend on the internet for leisure?" Respondents were asked to indicate their amount of internet use on a 1 to 24-h scale.

2.3.5. Social media use

One item asked respondents how often they used the internet to use social networking sites (SNSs) including Facebook, Instagram, Snapchat and Twitter. Responses were recorded along a 5-point frequency scale ranging from 1 = never, to 5 = always, with higher scores indicating more frequent social media use.

2.3.6. Online self-disclosure

Online self-disclosure was measured by four items from the amount of self-disclosure scale, which was developed by Gibbs, Ellison, and Heino (2006). This scale assesses the degree to which individuals self-disclose in online settings. For example, "I often discuss my feelings about myself online." Respondents were asked to rate their level of agreement with each statement on a 5-point scale ranging from 1 = strongly disagree to 5 = strongly agree. Negatively worded items were reverse coded.

2.3.7. Trolling

The four item Global Assessment of Internet Trolling (Buckels, Trapnell, & Paulhus, 2014), was used in this study. This scale assesses individuals' experiences and beliefs about trolling others online, with items including "I enjoy harassing other people in online settings". Respondents were asked to rate their level of agreement with each statement on a 5-point scale ranging from 1 = strongly disagree to 5 = strongly agree.

2.4. Data analytic plan

The data was analyzed in three phases. Firstly, the data were subject to a random split into two sub-samples. In the first sub-sample Exploratory factor analysis (EFA) was conducted to examine the structure of the MOD, identify the number of factors to be retained, and reduce the number of items. For the EFA, maximum likelihood estimation and oblique (oblim) rotation were applied. Next, utilizing the second sub-sample Confirmatory factor analyses (CFA) were performed to test the construct validity of the MOD and to further reduce redundancy in the measure. After the structural analyses were complete, the sub-samples were then recombined and as a final step of analysis, we examined whether the model was equivalent by gender. Following methods for testing invariance, models were conducted testing configural, metric, scalar, and residual invariance. We then compared the fit of each nested model to the previous solution by examining change in fit indices, with

Table 2
Study1. Model fit indices and $\Delta\chi^2$ between confirmatory factor analysis models (N = 190).

CFA Model	χ^2	*df*	RMSEA	GFI	CFI	SRMR	$\Delta\chi^2$	Δdf	*p*
1	271.84	90	.10	.82	.91	.05	–	–	–
2	189.18	77	.09	.86	.94	.04	82.66	13	.001
3	132.25	65	.07	.90	.96	.04	56.93	12	.001
4 (Final)	99.01	54	.07	.92	.97	.04	33.24	11	.001

Note: RMSEA, Root Mean Square Error of Approximation; GFI, Goodness of Fit Index; CFI, Comparative Fit Index; SRMR, Standardized Root Mean Residual.

nonsignificant differences in fit between the models indicating that the null hypothesis of invariance can be retained. Following invariance testing, the scale was computed, and the reliability and descriptive statistics of the measure were assessed. The construct validity was then examined through bivariate correlations with the measures of convergent and criterion validity. Finally, exploratory analyses on the relationship between potential predictors (time spent online and on social media) and the potential outcomes (trolling and online self-disclosure) of online disinhibition were examined.

3. Results

3.1. Exploratory factor analyses

Using the exploratory sub-sample (n = 212) the 28 items of the MOD were subjected to statistical analyses to ensure adequate item variance, to check the factor structure, and to establish acceptable item-total correlations. First, the negatively worded items were reverse scored, and then the items were subjected to exploratory factor analysis (EFA) in SPSS version 25 with oblim rotation. The results of the EFA found 2 factors with eigenvalues greater than 1.0, the first of which explained 59.90% of the variance, and the second which explained 7.64% of the variance (eigenvalues = 16.77 and 2.14). One item was found to weakly load onto both factors, and thus was removed and the model re-run. Inspection of the solution indicated that the second factor comprised of only four items loading above 0.30. Two items were found to cross-load substantially across both factors and thus were removed. Re-running of the model indicated that the final two items loading onto the second factor were reverse scored items (I am more private online than I am offline, and I am more likely to keep my opinions to myself online than in person). These items were removed for the sake of parsimony and the model was again re-run.

The remaining 23 items conformed to a single factor which explained 67.52% of the variance. While this was considered a strong solution, in the final step of model building, inter-item correlations were inspected and in order to reduce redundancy, and where items correlated above .75, these were removed one by one. Eight items were removed in this final step resulting in a final model comprising 15 items which all loaded at above 0.75 on a single factor (see Table 1). Reliability analyses of the scale indicated that the items had high levels of internal consistency Cronbach's $\alpha = 0.97$.

3.2. Confirmatory factor analyses

In the second stage of the analysis a Confirmatory Factor Analysis was conducted on the confirmatory sub-sample (n = 191) in Amos version 25. Table 2 outlines all of the model fit statistics. The first model tested all 15 of the observed scale items loading onto a single construct. This model was found to not fit the data adequately (χ^2 (90) = 271.84, $p < .001$; RMSEA = 0.10, GFI = 0.82, CFI = 0.91, SRMR = 0.05). Upon inspection of factor loadings, one of the items was found to load below 0.50 on the latent construct "The way I talk to people online is different than how I talk offline". This item was removed and the model re-run, with model fit indices showing a substantial improvement ($\Delta\chi^2$ (13) = 82.66, $p < .001$). However, a number of fit indices were still sub-par, therefore, modification indices were examined to assess misfit. Modification indices suggested that covariances between the residuals of two items "I find it easier to express myself on the internet than in a face-to-face conversation" and "I am more expressive online than I am offline" with a number of other item residuals were contributing to the misfit. These items were removed one at a time and the model re-run. Each step significantly improved the model fit ($\Delta\chi^2$ (12) = 56.93, $p < .001$ and $\Delta\chi^2$ (11) = 33.24, $p < .001$), effectively reducing item redundancy. The final model comprised 12 items with good model fit (χ^2 (54) = 99.01, $p < .001$; RMSEA = 0.07, GFI = 0.92, CFI = 0.97, SRMR = 0.04). See Table 3 for the Confirmatory factor analyses results and final Measure of Online Disinhibition.

Building upon the CFA, a multi-group model was developed to assess the equivalence of the factor structure by gender. This model was conducted with the overall sample, although 3 individuals were omitted due to missing information on gender. Thus, the male model included 202 participants and the female model 198 participants. Multi-group Model 1 tested the configural equivalence of the structure across the gender, allowing all parameters to be freely estimated. This model yielded a good fit to the data (χ^2 (108) = 270.99, RMSEA = 0.07, GFI = 0.89, CFI = 0.95) illustrating invariance of the overall model structure across the samples. In Model 2 metric equivalence was tested by constraining the factor loadings to be equal across the samples. Results indicate that the model fit did not significantly change following these constraints; χ^2 (119) = 287.15, RMSEA = 0.06, GFI = 0.89, CFI = 0.95, $\Delta\chi^2$ (10) = 16.16, $p > .001$, ΔCFI = 0.001, ΔRMSEA =.002, illustrating metric equivalence across the samples. In Model 3 scalar invariance was tested by constraining the factor intercepts across the groups. Again, results indicate the model fit did not significantly change χ^2 (120) = 287.21, RMSEA = 0.06, GFI = 0.89, CFI = 0.95, $\Delta\chi^2$ (1) = 0.05, $p > .001$, ΔCFI = 0.001, ΔRMSEA = 0.001, illustrating metric equivalence across the samples. Finally, Model 4 tested invariance of residuals. This model, however, was found to differ significantly from the previous model; χ^2 (132) = 321.89, RMSEA = 0.06, GFI = 0.89, CFI = 0.95) $\Delta\chi^2$ (12) = 34.68, $p < .001$, ΔCFI = 0.001, ΔRMSEA =.002, illustrating residual equivalence was not met . These results point to strong evidence for structural invariance of the latent measure among males and females.

Table 3
Study 1. Confirmatory factor analyses results and final measure of online disinhibition (N = 190).

	Factor Loading
1. I act differently online than I do offline	.732
2. I act tougher on the internet than I do face-to-face	.746
3. I am less cautious about what I say online than about what I say in person	.704
4. I am more assertive online than I am offline	.808
5. I am more competitive online than I am offline	.683
6. I am more confident online than I am offline	.774
7. I am more outgoing online than I am offline	.783
8. I am more able to discuss controversial issues online than I am in person	.769
9. I find communicating with others easier on the internet than in person	.772
10. I make friends more easily online than I do offline	.709
11. I say things on the internet that I would not say in person	.805
12. My behaviours online are less restricted than in person	.825

Rating Scale: *1* = Not at all like me, to *5* = Very like me.

Thus, the 12- items were averaged into a composite score and reliability analyses were conducted. The final MOD scale was found to have a Cronbach's alpha of .95, with analyses indicating that removal of any items would not reduce the internal consistency below .94.

3.3. Validity assessments and mediation models

To test the hypotheses that MOD would be positively associated with the measures or convergent and criterion validity, the sub-samples were recombined (N = 403) and bivariate correlations were conducted with the 12-item scale. The pattern of correlations emerged as expected (see Table 4 for correlations, alphas and descriptive statistics). Specifically, the MOD was found to correlate positively and strongly with the benign and toxic subscales of the Online Disinhibition Scale ($rs = 0.79$ and 0.64, $p < .001$ respectively) as well as with the online false self ($r = 0.74$, $p < .001$). The MOD was also positively correlated with average daily internet use and social media engagement ($r = 0.40$ and 0.31, $p < .001$ respectively). Finally, the MOD was also found to correlate significantly and positively with online self-disclosure ($r = 0.29$, $p < .001$) and trolling ($r = 0.61$, $p < .001$).

In order to explore the relationships among the potential predictors and outcomes of MOD, path analyses were conducted in AMOS version 25. This model concurrently tested the associations between MOD and online self-disclosure and trolling as well as the indirect effects of internet and social media use on self-disclosure and trolling via MOD. The significance of the indirect pathways was determined using 5000 bootstrapped samples producing 95% confidence intervals of the indirect effect (Hayes, 2009). The model was fully saturated, so model fit statistics were not available, and there were no missing data. Results of the path model (see Fig. 1) found that, as expected, both higher frequencies of internet use and social media use predicted greater MOD (ßs = 0.27 and 0.24, $p < .001$), and in turn MOD predicted higher levels of trolling ($\beta = 0.56$, $p < .001$) and higher levels of online self-disclosure (ß = 0.27, $p < .001$). Examination of effects found that frequency of using social media had a positive direct effect on trolling (ß = 0.20, $p < .001$), although neither internet nor social media use had direct effects on online self disclosure. However, the indirect effects of both internet use and social media use on self-disclosure and trolling via MOD were all significant. Specifically, frequency of internet and social media use were both indirectly associated with increased trolling via MOD (ß*Internet Use* = 0.14, $CI_{95\%}$ = 0.09 - 0.19, ß*Social Media Use* = 0.13, $CI_{95\%}$ = 0.08 - 0.18) and increased online self-disclosure via MOD (ß*Internet Use* = 0.07, $CI_{95\%}$ [0.04 - 0.12], ß*Social Media Use* = 0.07, $CI_{95\%}$ [0.04 - 0.11]). These results indicate that MOD fully mediates the impact of internet use on both trolling and online self-disclosure as well as fully mediating the effect of social media engagement on online self-disclosure and partially mediating the impact of social media engagement on trolling.

4. Study 2: scale validation and extension of nomological network

As previously mentioned, the influence of online disinhibition on harmful online behaviors has been a key focus of initial research into this largely misunderstood phenomenon (e.g. Varjas et al., 2010; Wright et al., 2019). It has been suggested that the absence of non-verbal cues, physical distance from others, and asynchronous feedback in online interactions diminish self-censorship, which can encourage aggressive behavior toward others, and reduce empathetic responding (Low & Espelage, 2013; Nesi et al., 2018a,b; Udris, 2014; Voggeser et al., 2018). Supporting these associations, previous research has found online disinhibition to be related to greater cyberbullying, cyberaggression, cyberhate, and cybervictimization (Kurek et al., 2019; Udris, 2014; Wachs & Wright, 2019; Wachs et al., 2019). Furthermore, in a latent profile analysis examining patterns of cyberbullying and victimization it was found that cyber bullies, victims, and bully/victims (those who engage in high levels of cyberbullying perpetration and experience high levels of cybervictimization) all had elevated levels of online disinhibition in comparison to those not uninvolved (Antoniadou et al., 2019). For this reason, the first aim of the second study is to further validate the MOD by testing its predictive associations with cyberbullying and cybervictimization. Following the findings of previous research, it was hypothesized that the MOD would be significantly positively associated with both cyberbullying perpetration and victimization.

As past research has primarily focused on the negative effects of online disinhibition, this study also aims to expand the nomological network of the construct in assessing its relationship to positive social and psychological outcomes. A recent systematic review found that there are a variety of benefits of digital technologies for well-being such as increased self-esteem, social support, social capital, and opportunities for self-disclosure (Best, Manktelo, & Tayor, 2014). Some longitudinal studies suggest that the key reasons for the relationship between well-being and digital engagement are that the online environment provides a sense of social comfort that, in turn, results in increases in positive outcomes (Szwedo, Mikami, & Allen, 2012; Valkenburg & Peter, 2009). In fact, the first study in this manuscript, alongside preliminary work on the effects of online disinhibition, found that online disinhibition is related to increased online self-disclosure (Clark-Gordon et al., 2019; Schouten et al., 2007), which itself has been found to be associated with higher levels of well-being in previous research (Ko & Kuo, 2009; Varnali & Toker, 2015). Furthermore, there is evidence to suggest that feeling less restraint online may have social benefits due to the increased ease felt by individuals in being able to express themselves and connect with others online as compared to during in-person interactions (Antoniadou et al., 2019; Scott, Stuart, O'Donnell, & Jose, under review). Thus, in this study we aimed to explore whether the MOD is associated with both greater well-being and greater social connectedness.

The final aim of this research is to further our initial investigations of the mediating effect of online disinhibition on the relationship between time spent online and the positive and negative outcomes outlined in study 1. Indeed, recent meta-analyses indicate that time spent online is positively associated with both cyberbullying and cybervictimization and negatively associated with indicators of well-being (Huang, 2017; Kowalski, Giumetti, Schroeder, & Lattanner, 2014; Çikrıkci, 2016). In this research, we seek to examine the influences of internet and social media use on digital behavior and well-being outcomes, and whether

Table 4
Study 1. Descriptive statistics and correlations between the measure of online disinhibition (MOD) and measures for convergent and criterion validity (N = 403).

	1	2	3	4	5	6	7	8	α	Mean	*SD*
1. MOD									.95	2.90	1.16
2. Benign OD	.79***								.85	3.41	0.89
3. Toxic OD	.64***	.55***							.83	2.47	1.14
4. Internet use	.40***	.35***	.31***						–	5.08	2.81
5. Social media use	.31***	.30***	.28***	.27***					–	3.24	1.15
6. Online False Self	.74***	.60***	.72***	.36***	.32***				.92	2.40	1.26
7. Online Self-Disclosure	.29***	.25***	.11*	.18**	.13*	.16**			.64	2.45	0.84
8. Trolling	.61***	.48***	.78***	.33***	.37***	.77***	.08		.92	2.05	1.29

Note: *$p < .05$, **$p < .01$, ***$p < .001$.

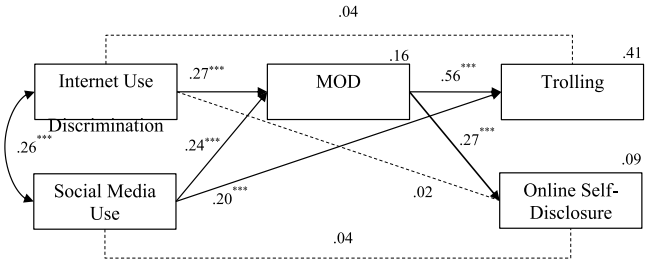

Fig. 1. Study 1 Mediation of Internet Use and Social Media Use on Trolling and Online Self-Disclosure via MOD (N = 403).Note: ***$p < .001$, **$p < .001$, *$p < .05$ Solid lines indicate significant coefficient and dashed lines indicate non-significant coefficient.

these are mediated by online disinhibition in a similar way to our previous results which found positive indirect associations on both negative (trolling) and positive (online self-disclosure) outcomes.

5. Method

5.1. Procedure and participants

A sample of undergraduate students were invited to participate in a confidential online survey through Qualtrics for course credit in a first year introductory Psychology course. No restrictions were put on inclusion with the exception that participants must be self-defined regular internet users. The research was approved under the delegated authority of the Griffith University's Human Ethics Committee.

In total 242 students aged between 17 and 66 ($M = 22.25$ years, SD = 8.74) completed the survey. The majority of the sample were female (74.1%, n = 180) and were full time students (73.6%). Participants identified predominantly as White (77.5%), followed by Asian (6.3%), with small numbers identifying with other ethnicities (e.g., Pacific Islander, Middle Eastern, African, Aboriginal and Torres Strait Islander).

5.2. Materials

5.2.1. Measure of online disinhibition (MOD)

The 12-item MOD that was developed and validated in study 1 was included in the survey. See Table 3 for scale details. Participants were asked to rate how much each item was representative of themselves on a rating scale from 1 = not at all like me, to 5 = very like me. Higher scores indicate greater levels of online disinhibition.

5.2.2. Internet use

As for Study 1, the amount of leisure time spent using the internet was assessed by computing the mean of two items concerning how much time was spent online for leisure during weekdays and weekends. See Study 1 materials for more details.

5.2.3. Social media use

One item asked respondents how often they used the internet to use social networking sites (SNSs) including Facebook, Instagram, Snapchat and Twitter. See Study 1 materials for more details.

5.2.4. Cyberbullying and Cybervictimization

Cyberbullying and cybervictimization were each assessed 10 items adapted from a scale developed by Kurek et al. (2019) to measure aggressive online behavior. The scales asked participants to rate how often in the last month (30 days) they had engaged in a series of online behaviors as either the aggressor or as a victim. Example items include "Have you made comments or posts to make someone upset or uncomfortable", "sent rude or upsetting images to someone", "threatened to physically hurt someone". Items were scored on a 5-point Likert-type scale (ranging from 1 = never to 5 = 7 or more times).

5.2.5. Social connectedness

Social connectedness was assessed using the 20-item Social Connectedness Scale-Revised (Lee, Draper, & Lee, 2001). The Social Connectedness Scale-Revised is measured on a 6-point agreement scale, from 1 = strongly disagree to 6 = strongly agree. Example items include "I am able to connect with other people", and "I see myself as a loner" (reverse scored). Higher scores indicate stronger feelings of social connectedness.

5.2.6. Flourishing

The Flourishing Scale (Diener et al., 2010) was utilized as an indicator of well-being. This is an 8-item assessment of social-psychological success, measuring a range of psychological needs including relatedness and competence. Example items include "I actively contribute to the happiness and well-being of others", and "I am optimistic about my future". The items are measured along a 7-point scale, from 1 = strongly disagree to 7 = strongly agree, with higher scores representative of higher psychosocial flourishing.

6. Results

6.1. Validity assessments

To test the hypotheses that MOD would be positively associated with the measures of validity, bivariate correlations were computed. The pattern of correlations emerged predominantly as expected (see Table 5 for correlations, alphas and descriptive statistics). Specifically, similar to the results of study 1, the MOD was found to be moderately positively correlated with daily internet use and social media engagement ($r = 0.27$ and 0.31, $p < .001$ respectively). Regarding criterion measures, the MOD was also found to correlate significantly and positively with cyberbullying ($r = 0.17$, $p < .01$), but not with cybervictimization ($r = 0.12$, $p = .07$). In exploring the extended network of associations between MOD and the positive indicators, disinhibition was also found to have significant associations with both social connectedness and flourishing, although these were in the opposite direction to what was expected ($r = -.34$. and $-.36$., $p < .001$ respectively).

Table 5
Study 2. Descriptive statistics and correlations between the measure of online disinhibition (MOD) and validity measures (N = 242).

	1	2	3	4	5	6	α	Mean	*SD*
1. MOD	–						.92	2.31	0.89
2. Internet Use	.27***	–					–	6.29	4.23
3. Social Media Use	.31***	.15*	–				–	3.92	1.16
4. Cyberbullying	.17**	.08	.04	–			.92	1.12	0.28
5. Cybervictimization	.12	.18**	-.03	.35***	–		.83	1.26	0.43
6. Flourishing	-.36***	-.27***	-.01	-.15*	-.11	–	.86	4.40	0.81
7. Social Connectedness	-.34***	-.17**	.09	-.03	-.12	.84***	.95	4.24	0.88

Note: *$p < .05$, **$p < .01$, ***$p < .001$.

6.2. Path analysis

A single path model examining indirect associations of internet use and social media use on cyber behaviors (cyberbullying and cybervictimization) and well-being (social connectedness and flourishing) via MOD was conducted. Similar to the model conducted in study 1, this path analysis was estimated in AMOS version 25 (see Fig. 2), and the significance of direct, indirect, and total were determined using 5000 bootstrapped samples producing 95% confidence intervals of (Hayes, 2009). The effect of covariates (age and gender), the covariance between internet and social media use as well as the covariances among residuals for cyber behavior and well-being respectively were controlled for in the full model. The covariance between the residuals of cyberbullying and flourishing was subsequently added by examining the modification indices.

The model was found to fit the data very well (χ^2 (5) = 2.16, p = .827, CFI = 1.00, GFI = .99, RMSEA = .01). Similar to the results of study 1, it was found the both frequency of internet use (β= .18, p = .003) and social media use (β= .22, $p < .001$) were associated with increased MOD. Regarding cyber behaviors, MOD was associated with increased cyberbullying (β= .15, $p < .05$), but not cybervictimization. In contrast, the frequency of internet and social media were not found to be associated with cyberbullying, but internet use was directly associated with greater cybervictimization (β= .14, p = .02). There was also evidence for significant indirect effects of internet use and social media use via MOD on cyberbullying. Specifically, frequency of internet and social media were both indirectly, albeit weakly, associated with increased cyberbullying via MOD (ß*Internet Use* = 0.04 = 0.03, CI95% [0.01 to 0.08], ß*Social Media Use* = 0.04, CI95% [0.01 to 0.08]). Results for cybervictimization found that frequency of internet use, social media use, and MOD were not associated with cybervictimization. However, the total effect of internet use on cybervictimization was significant (ß = 0.17, CI95% [0.01 to 0.04]).

Regarding well-being, the frequency of social media use was found to be positively associated with connectedness (β= .22 $p < .001$), whereas internet use was significantly negatively associated with flourishing (β = -.18, $p < .001$) and MOD was significantly negatively associated with both connectedness and flourishing (β = -.39 and -.33, $p < .001$) respectively. Furthermore, there was evidence of a significant, negative indirect effect such that internet use was associated with lower levels of connectedness (β = -.07, CI95% [-0.11 to -.04]) and flourishing (β = -.09, CI95% [-0.14 to - 0.05]), via MOD. Similarly, social media had a significant, negative indirect effect on connectedness (β = -.09, CI95% [-0.11 to -.04]) and flourishing (β = -.08, CI95% [-0.14 to - 0.05]) via MOD. These indirect effects reduced the total positive effect of social media use on connectedness (Total effect β = 0.16, CI95% [0.02 to 0.29]) and on flourishing (Total effect β = 0.05, CI95% [-0.08 to 0.18]).

7. Discussion

Technology has changed the way that people interact, with an emerging body of research finding that the features of the internet can promote online disinhibition, whereby people act, think, and feel differently online when compared to face-to-face settings (Suler, 2004). However, to date, measures that assess the experience of online disinhibition have not been well operationalized or validated. Specifically, previous measures often blend the theorized features of the internet and antecedents or outcomes of online disinhibition, such that the precursors, experiences and effects of online disinhibition are not treated as distinct factors. In fact, the most commonly used scale currently

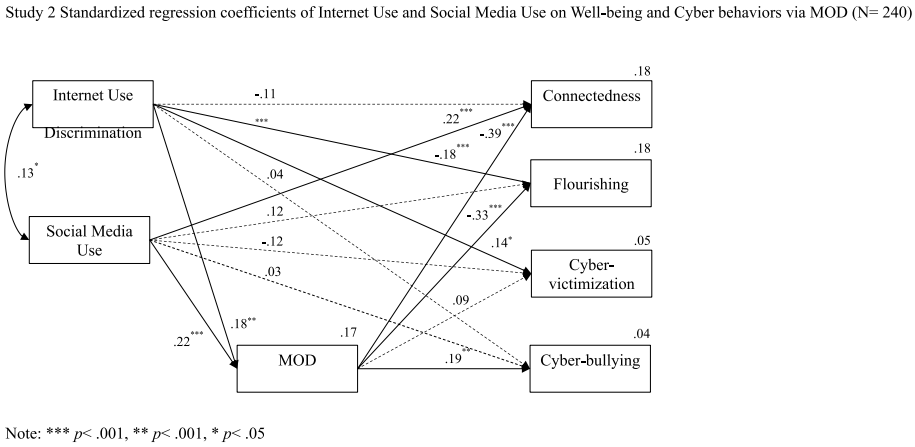

Fig. 2. Study 2 Standardized regression coefficients of Internet Use and Social Media Use on Well-being and Cyber behaviors via MOD (N = 240). Note: ***$p < .001$, **$p < .001$, *$p < .05$ Solid lines indicate significant coefficient and dashed lines indicate non-significant coefficient. Effects of covariates (age and gender) and covariances between residuals of exogenous variables not depicted for ease of interpretation.

available in the literature considers negative or "toxic" aspects of online disinhibition as compared to neutral or positive (i.e. "benign") disinhibition in its measurement (Udris, 2014), rather than examining the positive and negative outcomes of the experience of online disinhibition.

Through consideration of issues with existing measures of online disinhibition, as well as the growing need to investigate how the digital environment influences our behavior, we highlighted the importance of developing a robust, ecologically valid instrument of online disinhibition. As such, the primary aims of the current study were to: (1) to define and operationalize online disinhibition, (2) to develop a reliable instrument measuring online disinhibition, (3) to validate this measure, and (4) to begin to explore the relationship between internet use, online disinhibition, and both social and psychological outcomes. In our research, online disinhibition was defined as; the perception or experience of reductions in restraint in the online environment such that individuals may act, think, and feel differently online when compared to face-to-face interactions. Using two studies with unique samples, the Measure of Online Disinhibition (MOD), was developed as a comprehensive and valence free instrument for assessing self-perceptions of psychological and behavioral change in the online as compared to the offline environment.

In our studies we found a single factor solution that was replicated with confirmatory factor analysis. Having established a psychometrically robust structure for the MOD, evidence for construct validity was exhibited via relationships with toxic and benign disinhibition, false self-presentation online, frequency of internet and social media use, as well as trolling and online self-disclosure. Notably, and in line with previous research, the MOD was significantly positively associated with both positive cyber behaviors (i.e., benign disinhibition scale and online self-disclosure) as well as negative cyber behaviors (i.e., toxic disinhibition, trolling, cyberbullying, and cybervictimization) suggesting that the new construct appropriately captures online disinhibition in a balanced hedonic fashion.

Finally, in order to examine the relationships between MOD and its potential predictors and outcomes, two path models were constructed. The analysis found that greater time spent online and on social media were associated with greater online disinhibition, and in turn with higher levels of online self-disclosure, trolling, and cyber-bullying and lower levels of social connectedness and flourishing. These results indicate that online disinhibition may play an important role in influencing how time spent in the digital context translates into potential positive and negative behaviors online for individuals. This is a core contribution of the current study as previous research regarding individual level predictors of online disinhibition have predominantly focused on relatively uncommon personality characteristics as influences of disinhibition, and in turn, aggressive forms of cyber behavior (e.g., Antoniadou et al., 2019; Kurek et al., 2019). Our results extend this emerging literature by suggesting that frequency of engagement in online contexts both in terms of time spent for general internet use and time spent on social media use can influence the likelihood of experiencing disinhibition, which subsequently impacts on one's behaviors (for both the positive and negative) in the online context.

Additionally, contrary to our expectations and research extolling the benefits of digital technologies for well-being (Best, Manktelow, & Taylor, 2014), we found that online disinhibition was negatively associated with well-being and time spent online exerted a negative effect on well-being via the MOD. One interpretation of these results is that online disinhibition predominantly influences behaviors in the online context, which themselves may or may not have flow-on effects for well-being both on and offline (or indeed a range of other outcomes). For example, research has found that the positive effects of the online environment are often a result of feeling a sense of belonging and perceiving the ability to safely self-disclose online (Ko & Kuo, 2009; Szwedo et al., 2012; Valkenburg & Peter, 2009; Varnali & Toker, 2015). Therefore, it may be the case that online disinhibition increases opportunities for social support online, which then influences well-being. Another interpretation of these results is that online disinhibition has positive effects on well-being for some people, but not for others. There is evidence to suggest that online disinhibition has benefits for those who are socially anxious as the online context enables those who have difficulty in expressing themselves in-person to engage with others online (Antoniadou et al., 2019; Scott et al. under review). Those who do not face such difficulties interacting with others offline may not reap the benefits of a less restrained self online. Specifically, as online disinhibition has been found to be related to false self presentations (Kurek et al., 2019), individuals higher in disinhibition may be aware of and face subsequent decrements in well-being that result from perceptions of inauthenticity on the online environment when interacting with others.

While acknowledging the exploratory nature of the present research, there are some limitations that must be considered. First, the studies included relatively small and non-representative samples meaning that we are unable to generalize the results more broadly. Future research should aim to use more diverse samples, confirming both the structure of the measure and the associated relationships with individuals from different age groups, cultural contexts, and digital literacy levels. Second, self-report measures were used for all constructs in the current study. Such data are consistently employed in social media research and are essential for measuring self-perceptions of psychological and behavioral change in the online environment. However, future research would benefit from employing the measure alongside experimental, observational, and qualitative studies of online disinhibition such as those undertaken by Lapidot-Lefler and Barak (2012) and Voggeser et al. (2018). Additionally, the Cronbach's alpha of the MOD was above $\alpha = 0.90$, which has been suggested could indicate that some items in the measure are redundant (Peterson, 1994). However, the final 12 items were retained because of their conceptual significance in measuring the construct, and in following guidelines that suggest inspection of the relationships between items in these circumstances is an appropriate way to measure such redundancy (Taber, 2018). Lastly, it is notable that both studies were cross-sectional, and as such, the directionality of the relationships between internet and social media use, the MOD, and outcome variables of interest that were tested in the exploratory analyses cannot be considered to be conclusive evidence of the sequence of relationships. We recommend that future research use the MOD in with diverse samples in longitudinal studies as well as continue exploration of the nomological network of online disinhibition.

A key strength of the present research was that the MOD was designed to be distinct from the internet attributes which are thought to be antecedents of online disinhibition (e.g. dissociative anonymity, invisibility, and asynchronicity; Suler, 2004). This differentiates the MOD from past measures of this construct that have conflated several precursors to online disinhibition with experiences of disinhibition. Although these antecedents were not a focus of the present study, it must be noted that they are theoretically a key factor in encouraging online disinhibition. Therefore, it is necessary that future research continue work in defining and operationalizing the antecedents of online disinhibition as perception of the attributes of the internet. Specifically, if individual perceptions of the affordances of the online environment are, indeed, the factors which precede experiences and perceptions of online disinhibition, more work into understanding these is warranted in order to effectively investigate how features of digital social environments influence online behavior.

The development of the MOD has extended the available literature on online disinhibition and provides opportunities for robust and generalizable research into the future. Specifically, the current study has contributed by defining and developing a robust, valid, and valence-free measure of online disinhibition that ostensibly overcomes several limitations within existing measures. We also subsequently explored the relationship between time spent online, online disinhibition, cyber behaviors, and well-being in an extension of past literature. In our increasingly digitally connected world, there is a growing need to understand how online environments change people for the better and for

the worse. This study adds to the growing, and important body of research that investigates the social, psychological, and cognitive impacts of our increasing use of the internet. We believe that this research, and the MOD as an instrument, can help to elucidate the impacts of the internet on everyday life. Specifically, it is expected that the MOD can be utilized in future research in order to better understand why individuals might think, act, and feel differently online as compared to offline and what implications this has for their health and well-being.

CRediT authorship contribution statement

Jaimee Stuart: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing. **Riley Scott:** Methodology, Formal analysis, Writing - original draft, Writing - review & editing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chb.2020.106534.

References

Antoniadou, N., Kokkinos, C. M., & Fanti, K. A. (2019). Traditional and cyber bullying/victimization among adolescents: Examining their psychosocial profile through latent profile Analysis. *International Journal of Bullying Prevention, 1*(2), 85–98. https://doi.org/10.1007/s42380-019-00010-0.

Antoniadou, N., Kokkinos, C. M., & Markos, A. (2019). Psychopathic traits and social anxiety in cyber-space: A context-dependent theoretical framework explaining online disinhibition. *Computers in Human Behavior, 99*, 228–234. https://doi.org/10.1016/j.chb.2019.05.025.

Armstrong, L., Phillips, J. G., & Saling, L. L. (2000). Potential determinants of heavier Internet usage. *International Journal of Human-Computer Studies, 53*(4), 537–550. https://doi.org/10.1006/ijhc.2000.0400.

Barlett, C. P., & Helmstetter, K. M. (2018). Longitudinal relations between early online disinhibition and anonymity perceptions on later cyberbullying perpetration: A theoretical test on youth. *Psychology of Popular Media Culture, 7*(4), 561.

Best, P., Manktelow, R., & Taylor, B. (2014). Online communication, social media and adolescent wellbeing: A systematic narrative review. *Children and Youth Services Review, 41*, 27–36. https://doi.org/10.1016/j.childyouth.2014.03.001.

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences, 67*, 97–102.

Casale, S., Fiovaranti, G., & Caplan, S. (2015). Online disinhibition. *Journal of Media Psychology, 27*, 170–177. https://doi.org/10.1027/1864-1105/a000136.

Cheung, C. M. K., Wong, R. Y. M., & Chan, T. K. H. (2016). Online disinhibition: Conceptualization, measurement, and relation to aggressive behaviors. *Proceedings of International Conference on Information Systems 2016*, 1–10.

Çikrıkci, Ö. (2016). The effect of internet use on well-being: Meta-analysis. *Computers in Human Behavior, 65*, 560–566. https://doi.org/10.1016/j.chb.2016.09.021.

Clark-Gordon, C. V., Bowman, N. D., Goodboy, A. K., & Wright, A. (2019). Anonymity and online self-disclosure: A meta-analysis. *Communication Reports, 32*(2), 98–111. https://doi.org/10.1080/08934215.2019.1607516.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research, 97*(2), 143–156.

Gibbs, J. L., Ellison, N. B., & Heino, R. D. (2006). Self-presentation in online personals: The role of anticipated future interaction, self-disclosure, and perceived success in Internet dating. *Communication Research, 33*(2), 152–177.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408–420.

Huang, C. (2017). Time spent on social network sites and psychological well-being: A meta-analysis. *Cyberpsychology, Behavior, and Social Networking, 20*(6), 346–354. https://doi.org/10.1089/cyber.2016.0758.

Joinson, A. (1998). Causes and implications of disinhibited behavior on the internet. In J. Gackenbach (Ed.), *Psychology and the internet: Intrapersonal, interpersonal, and transpersonal implications* (pp. 43–60). San Diego, CA: Academic Press.

Kim, H., & Chang, Y. (2017). Managing online toxic disinhibition: The impact of identity and social presence. *SIGHCI 2017 Proceedings*, 1–5.

Ko, & Kuo. (2009). Can blogging enhance subjective well-being through self-disclosure? *CyberPsychology and Behavior, 12*(1), 75–79. https://doi.org/10.1089/cpb.2008.016.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin, 140*(4), 1073–1137. https://doi.org/10.1037/a0035618.

Kurek, A., Jose, P. E., & Stuart, J. (2019). 'I did it for the LULZ': How the dark personality predicts online disinhibition and aggressive online behavior in adolescence. *Computers in Human Behavior, 98*(98), 31–40. https://doi.org/10.1016/j.chb.2019.03.027.

Lai, C. Y., & Tsai, C. H. (2016). Cyberbullying in the social networking sites: An online disinhibition effect perspective. *Proceedings of the 3rd multidisciplinary international social networks Conference on social Informatics*. August.

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior, 28*(2), 434–443. https://doi.org/10.1016/j.chb.2011.10.014.

Lapidot-Lefler, N., & Barak, A. (2015). The benign online disinhibition effect: Could situational factors induce self-disclosure and prosocial behaviors? *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 9*(2).

Lee, R. M., Draper, M., & Lee, S. (2001). Social connectedness, dysfunctional interpersonal behaviors, and psychological distress: Testing a mediator model. *Journal of Counseling Psychology, 48*(3), 310.

Litt, D. M., & Stock, M. L. (2011). Adolescent alcohol-related risk cognitions: The roles of social norms and social networking sites. *Psychology of Addictive Behaviors, 25*(4), 708–713. https://doi.org/10.1037/a0024226.

Low, S., & Espelage, D. (2013). Differentiating cyber bullying perpetration from non-physical bullying: Commonalities across race, individual, and family predictors. *Psychology of Violence, 3*(1), 39.

Michikyan, M., Subrahmanyam, K., & Dennis, J. (2014). Can you tell who I am? Neuroticism, extraversion, and online self-presentation among young adults. *Computers in Human Behavior, 33*, 179–183.

Morahan-Martin, J., & Schumacher, P. (2000). Incidence and correlates of pathological Internet use among college students. *Computers in Human Behavior, 16*, 13–29.

Nesi, J., Choukas-Bradley, S., & Prinstein, M. J. (2018a). Transformation of adolescent peer relations in the social media context: Part 1—a theoretical framework and application to dyadic peer relationships. *Clinical Child and Family Psychology Review, 21*(3), 267–294. https://doi.org/10.1007/s10567-018-0261-x.

Nesi, J., Choukas-Bradley, S., & Prinstein, M. J. (2018b). Transformation of adolescent peer relations in the social media context: Part 2—application to peer group processes and future directions for research. *Clinical Child and Family Psychology Review, 21*(3), 295–319. https://doi.org/10.1007/s10567-018-0262-9.

Niemz, K., Griffiths, M., & Banyard, P. (2005). Prevalence of pathological Internet use among university students and correlations with self-esteem, the General Health Questionnaire (GHQ), and disinhibition. *CyberPsychology and Behavior, 8*(6), 562–570.

Peter, J., Valkenburg, P. M., & Schouten, A. P. (2007). Precursors of adolescents' use of visual and audio devices during online communication. *Computers in Human Behavior, 23*(5), 2473–2487.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21*(2), 381–391.

Rimal, R. N., & Real, K. (2005). How behaviors are influenced by perceived norms: A test of the theory of normative social behavior. *Communication Research, 32*(3), 389–414.

Schouten, A. P., Valkenburg, P. M., & Peter, J. (2007). Precursors and underlying processes of adolescents' online self-disclosure: Developing and testing an "Internet-attribute-perception" model. *Media Psychology, 10*(2), 292–315. https://doi.org/10.1080/15213260701375686.

Starkstein, S. E., & Robinson, R. G. (1997). Mechanism of disinhibition after brain lesions. *The Journal of Nervous and Mental Disease, 185*(2), 108–114.

cott, R. A., Stuart, J., O'Donnell, K. J., & Jose, P. E. (under review). Adolescent perceptions of online interactions: Can the Internet reduce the negative impacts of social vulnerability for young people?.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology and Behavior, 7*(3), 321–326. https://doi.org/10.1089/1094931041291295.

Szwedo, D. E., Mikami, A. Y., & Allen, J. P. (2012). Social networking site use predicts changes in young adults' psychological adjustment. *Journal of Research on Adolescence, 22*(3), 453–466. https://doi.org/10.1111/j.1532-7795.2012.00788x.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296.

Thacker, S., & Griffiths, M. D. (2012). An exploratory study of trolling in online video gaming. *International Journal of Cyber Behavior, Psychology and Learning (IJCBPL), 2*(4), 17–33.

Udris, R. (2014). Cyberbullying among high school students in Japan: Development and validation of the Online Disinhibition Scale. *Computers in Human Behavior, 41*, 253–261. https://doi.org/10.1016/j.chb.2014.09.036.

Valkenburg, P. M., & Peter, J. (2009). The effects of instant messaging on the quality of adolescents' existing friendships: A longitudinal study. *Journal of Communication, 59*(1), 79–97. https://doi.org/10.1111/j.1460-2466.2008.01405x.

Varjas, K., Talley, J., Meyers, J., Parris, L., & Cutts, H. (2010). High school students' perceptions of motivations for cyberbullying: An exploratory study. *Western Journal of Emergency Medicine, 11*(3), 269–273.

Varnali, K., & Toker, A. (2015). Self-disclosure on social networking sites. *Social Behavior and Personality: International Journal, 43*(1), 1–13.

Voggeser, B. J., Singh, R. K., & Göritz, A. S. (2018). Self-control in online discussions: Disinhibited online behavior as a failure to recognize social cues. *Frontiers in Psychology, 8*, 2372.

Wachs, S., & Wright, M. F. (2018). Associations between bystanders and perpetrators of online hate: The moderating role of toxic online disinhibition. *International Journal of Environmental Research and Public Health, 15*(9). https://doi.org/10.3390/ijerph15092030.

Wachs, S., & Wright, M. F. (2019). The moderation of online disinhibition and sex on the relationship between online hate victimization and perpetration. *Cyberpsychology, Behavior, and Social Networking, 22*(5), 300–306. https://doi.org/10.1089/cyber.2018.0551.

Wachs, S., Wright, M. F., & Vazsonyi, A. T. (2019). Understanding the overlap between cyberbullying and cyberhate perpetration: Moderating effects of toxic online

disinhibition. *Criminal Behaviour and Mental Health, 29*(3), 179–188. https://doi.org/10.1002/cbm.2116.

Weidman, A. C., Fernandez, K. C., Levinson, C. A., Augustine, A. A., Larsen, R. J., & Rodebaugh, T. L. (2012). Compensatory internet use among individuals higher in social anxiety and its implications for well-being. *Personality and Individual Differences, 53*(3), 191–195.doi. https://doi.org/10.1016/j.paid.2012.03.003.

Wright, M. F., Harper, B. D., & Wachs, S. (2019). The associations between cyberbullying and callous-unemotional traits among adolescents: The moderating effect of online disinhibition. *Personality and Individual Differences, 140*, 41–45.

Wu, S., Lin, T. C., & Shih, J. F. (2017). *Examining the antecedents of online disinhibition.* Information Technology & People.

Zuckerman, M. (1979). *Sensation seeking: Beyond the optimal level of arousal.* Hillsdale, NJ: Erlbaum.