Osaka Human Sciences, **9**, 63-81
March 2023

Online Disinhibition: Reconsideration of the Construct and Proposal of a New Model

Ruohan WEN[1,2], Asako MIURA[2]

Abstract

On the Internet, one comes across behaviors that are not observed in real life. The online disinhibition theory, pioneered by Suler (2004), has frequently been cited in empirical studies to explain this phenomenon. However, scholars have not yet reached a consensus regarding the construct of online disinhibition. This study explored an appropriate construct of online disinhibition for psychological research and proposed a model to explain its functioning. Previous studies have examined online disinhibition from three perspectives. This paper discusses the contributions and limitations of previous studies and postulates that psychological research on online disinhibition should be conducted from the perspective of mental state. Three significant models that explain the working of online disinhibition were reviewed: the "benign/toxic disinhibition model," "online disinhibition/behaviors model," and "online disinhibition and deindividuation model." Finally, the "motivation-based online disinhibition model" is proposed as an improved model that solves the limitations of the aforementioned models.

Keywords: Online disinhibition, disinhibitive behavior, benign disinhibition, toxic disinhibition, deindividuation, motivation-based online disinhibition model

---

This article is the English translation of "Wen, R., & Miura, A (2022). Online Disinhibition: Reconsideration of the construct and proposal of new model. *Japanese Psychological Review*, **65**(1) (in Japanese)." The publication of this translation has been permitted by the Society of Japanese Psychological Review.
[1]Social Psychology Lab, Graduate School of Human Sciences, Osaka university, 1-2, Yamadaoka, Suita, Osaka 565-0871, Japan
[2]Email: runningwz@gmail.com

1. Social problems in the Internet age and the online disinhibition effect

In the past 20 years, information and communication technology has completely transformed human life. We can now overcome the limitations of time and space and communicate with others anytime and anywhere. Nevertheless, many adverse circumstances that emerge online have become social concerns. For example, “flaming”—releasing a torrent of irrational and abusive language toward a specific person or organization in the short term—causes undue psychological distress to both the target and those around them. Another example is “Internet trolls”—people who intentionally provoke others online to elicit an argument or emotional reaction, and whose highly malicious nuisance behavior has a destructive impact on the online environment(Buckels, Trapnell, & Paulhus, 2014). However, the psychological mechanisms underlying cyber-violence have not been thoroughly examined.

Notably, people do things in the virtual world that they would not do in reality. For example, some may disclose secrets they rarely share in person, or create and act under a persona different from the real world, perhaps falsifying their gender or age. Suler (2004) proposed an online disinhibition effect theory to explain this phenomenon, suggesting the online disinhibition effect is a phenomenon wherein behavioral inhibition that typically exists in in-person settings lessens or disappears when online. The effect is further divided into two types: benign disinhibition, wherein psychological defenses weaken, allowing one to communicate freely to others and resolve interpersonal challenges, and toxic disinhibition, wherein one speaks or acts aggressively because there is no fear of incurring real punishment.

Suler’s (2004) theory describes the features of computer-mediated communication (CMC) via personal computers or smartphones to explain why people become more open or aggressive online. Since its publication, this theory has significantly influenced human, social science, and informatics research in the online context. For example, behaviors unique to the Internet age, such as online aggression and self-disclosure, are associated with online disinhibition (e.g., Hollenbaugh & Everett, 2013; Lowry, Zhang, Wang, & Siponen, 2016; Udris, 2014; Wu, Lin, & Shih, 2017). Beyond theoretical research, this theory also contributes to resolving social issues by providing system developers with specialized knowledge on features like anonymity to develop countermeasures (Cheung, Wong, & Chan, 2021).

Nonetheless, when Suler (2004) presented his theory, he did not clearly and precisely define terms such as “online disinhibition,” “antecedent factors of online disinhibition,” or “disinhibitive behavior.” This led subsequent studies to explore online disinhibition from an extensive range of perspectives, further complicating the construct. If the trend continues, the confusion surrounding this construct may impede the theory’s cohesive, cross-sectional development. Diverse viewpoints on online disinhibition should be organized and reevaluated based on a consistent framework to promote a more detailed and robust construct of online disinhibition theory.

Accordingly, this study comprehensively reviews the literature on online disinhibition and reconsiders how to appropriately define it based on psychological research, considering the modern online environment. Additionally, we refined the mechanisms that impact human behavior on the Internet through online disinhibition.

2. Perspectives on the online disinhibition effect

Suler's (2004) theory does not accurately define online disinhibition. Therefore, subsequent research has assumed diverse approaches for defining it, ranging from describing attributes of the online environment that facilitate the onset of disinhibition to explaining how people behave differently than they would in actuality due to online disinhibition. A summary of the three approaches is shown in Figure 1.

Figure 1 Various online disinhibition approaches

2.1. Features of each perspective

2.1.1. Attributes related to the onset of online disinhibition

Suler (2004) identified six factors of the online environment and the resulting CMC leading to online disinhibition: dissociative anonymity, invisibility, asynchronicity, dissociative imagination, solipsistic introjection, and minimization of status and authority. He argues that interactions between these factors lead to online disinhibition. Based on this perspective, Udris (2014) and Cheung et al. (2021) developed scales to measure individuals' online disinhibition. Udris (2014) developed an 11-item scale against the background of cyberbullying in Japanese schools. Subsequently, Cheung et al. (2021) performed a more detailed review of Suler's six factors to create a 22-item scale based on Udris' (2014) scale.

2.1.2. Behaviors resulting from online disinhibition

This perspective focuses on behaviors resulting from online disinhibition. Online disinhibition is regarded as freely doing something online that may be restrained in reality (e.g. Joinson, 2007; Lapidot-Lefler & Barak, 2012; Lapidot-Lefler & Barak, 2015). Cyberbullying

(Lai & Tsai, 2016; Wright, Harper, & Wachs, 2019), flaming (Lapidot-Lefler & Barak, 2012), self-disclosure in a blog (Hollenbaugh & Everett, 2013), online aggression (Wu et al., 2017), self-disclosure and prosocial behavior in online chat (Lapidot-Lefler & Barak, 2015) are typical examples of online disinhibition. Additionally, though no empirical psychological research has been conducted yet, phenomena such as the "flame wars" and "shitposting" (mainly referring to replies that make a person feel uncomfortable) that frequently occur in Japan and China are also typical examples of online disinhibition under this definition.

2.1.3. Online disinhibitive mental state

This perspective interprets the phrase “online disinhibition” literally and focuses on the mental state when inhibitions are weakened online. This perspective regards online disinhibition as a mental state in which behavior is not consciously controlled. Kurek, Jose and Stuart (2019) validated the association between “dark” personality traits (narcissism, psychopathy, and sadism) and a disinhibitive mental state. A disinhibitive mental state predicts cyber aggression. Schouten, Valkenburg and Peter (2007) found that awareness of two CMC traits, reduced nonverbal cues and controllability, correlated with this mental state, and the disinhibitive mental state further influenced online self-disclosure. In both cases, this perspective of online disinhibition is clearly distinct from the onset and behavioral perspectives, and may be considered a bridge between the two.

2.2. Contributions and limitations of each perspective

Online disinhibition has been defined from various perspectives, and the developing investigations from each perspective represent the complexity of the concept. This study explores the respective contributions and limitations of the onset and behavioral perspectives based on existing empirical research. Subsequently, it argues that it is necessary to adopt the mental state perspective in psychological research.

2.2.1. Online disinhibition onset perspective

This perspective views online disinhibition by considering attributes of the online environment related to its onset (e.g., anonymity and invisibility). It is a relatively straightforward approach that considers the phenomenon systematically, and hence, has created a basis for subsequent applications of online disinhibition theory across various fields. Udris’s (2014) Online Disinhibition Scale, based on Suler’s six-factor theory, is considered to have made significant progress in developing a tool to measure online disinhibition. As cyberbullying remains a major social concern, the significant impact of Udris’s (2014) findings has extended beyond social psychology (e.g., Lai & Tsai, 2016; Wang et al., 2020; Wright et al., 2019; Wright & Wachs, 2021; Yang, Wang, Gao, & Wang, 2021) to media studies (Saunders, 2016), the sociological context of cyberbullying (e.g., Heirman et al., 2016; Sobba,

Paez, & Bensel, 2017, Udris, 2015), school psychology-based research on cyberbullying for young children (DePaolis & Williford, 2015), and forming education policies to prevent cyberbullying (Cox, Marczak, Teoh, & Hassard, 2017).

However, this has led to criticism that the onset perspective does not distinguish between online disinhibition and its determinants (Stuart & Scott, 2021). For example, the item "The Internet is anonymous, so it is easier for me to express my true feelings or thoughts" from Udris's (2014) scale conflates the cause (The Internet is anonymous) and the effect (it is easier for me to express my true feelings or thoughts). Indeed, relying solely on Suler's (2004) theory or the findings on anonymity accumulated through social psychology research (e.g., Joinson, 2001) suggests that online disinhibition intensifies in anonymous and invisible environments because it is challenging to identify the act's perpetrator. However, recent studies suggest that the relationship between the anonymity and invisibility of the online environment and online disinhibition is complex. Research has observed low reproducibility of the impact of anonymity on online disinhibition or disinhibitive behavior. For example, research on cyberbullying has frequently shown that it occurs even in non-anonymous environments (e.g., Bryce & Fraser, 2013; Huang, Zhang, & Yang 2020; Wright et al., 2019). Studies of blogs have yielded similar conclusions. Hollenbaugh & Everett (2013) analyzed self-disclosure trends in personal blogs and found that those that shared a picture of themselves—those that were more visually identified—disclosed more information. These results are contrary to the predictions based on the onset perspective. A meta-analysis of the relationship between anonymity and online self-disclosure by Clark-Gordon, Bowman, Goodboy, & Wright (2019) found a typically weak, positive correlation between anonymity and self-disclosure ($r = .18$). However, the correlation was negative in some cases. Hence, the relationship between the attributes of the online environment, such as anonymity and invisibility, and online disinhibition should be explored further. For example, the abovementioned item, "The Internet is anonymous, so it is easier for me to express my true feelings or thoughts," from Udris' (2014) scale is too simple to explore the relationship in depth.

Furthermore, it is necessary to reconsider the validity of the other four factors Suler (2004) discussed (asynchronicity, dissociative imagination, solipsistic introjection, and minimization of status and authority). When online disinhibition effect theory was developed, information and communications technology was beginning to spread. Consequently, people's perceptions of and attitudes toward the features of the online environment were also in their early stages, and may have changed significantly with the growth of technology and the popularization of CMC. Suler (2004) believes that asynchronous communication promotes more open communication because the social norms of traditional communication (e.g., the need to reply immediately) do not apply. However, as asynchronous communication becomes more common, new social norms may arise. For example, in social media applications such as "LINE" that display the "read or unread" status of personal messages, not responding to a message for a long time may cause

the conversation partner to experience negative feelings, such as anxiety from feeling ignored (the anxiety of being “left on read”). Conversely, some people are extremely conscientious in preventing their conversation partners from feeling such anxiety, and feel a sense of urgency to respond immediately after reading a message (Tokioka et al. 2017). In other words, people’s perceptions of the asynchronicity of CMC are becoming more complex and varied than when Suler’s theory was proposed. The onset perspective does not consider the possibility of such changes, nor can it concretely examine how the perception of features of the online environment and disinhibition are related, because they are viewed as the same.

2.2.2. Online disinhibitive behavior perspective

This perspective views online disinhibition as a specific behavior. For example, Lapidot and Barak (2012) operationally defined online disinhibition as the use of hostile expressions toward others in online communication. Therefore, they conducted a detailed investigation into the factors causing the onset of this behavior. The researchers asked participants to discuss a dilemma, observe the use of hostile expressions, and examine the effects of anonymity, invisibility, and eye contact. They found that eye contact, or lack thereof, affected the number of hostile expressions more than anonymity. Owing to taking online disinhibition into a particular view, this perspective has been of great value in analyzing individual cases, such as cyberbullying (e.g., Bryce & Fraser 2013), which is of grave concern.

Internet trolling and flame wars frequently occur on the Japanese and Chinese Internet and have become a severe social concern. Applying the behavior perspective of online disinhibition theory can help explore how these behaviors arise from a social psychology perspective. However, according to survey research in Japan, engaging in flame wars is not universal. In an online survey, Yamaguchi (2015) found that only 303 (1.5%) of 19,992 respondents were involved in flaming. According to Koyama, Asatani, Sakaki, and Sakata (2019), who analyzed data using the official application programming interface provided by Twitter, 135,580 users were involved in the six flame wars that occurred on Twitter over two months, which is a mere 0.3% of Twitter’s 45 million active monthly users.

Other studies on aggressive online behavior have yielded similar conclusions. Masui, Tamura, & March (2019) developed a Japanese version of the Global Assessment of Internet Trolling, seeking answers to eight questions (e.g., I enjoy annoying people I don’t know online) on a five-point scale ranging from 1 (Extremely inappropriate) to 5 (Extremely appropriate). They found that the average score for almost every question ranged from 1 to 2 points. Udris (2015) conducted a social survey of Japanese high school students to determine the conditions of cyberbullying. The results demonstrated that in the previous six months, only 1.1% of respondents had spread messages containing insults or negative rumors among their classmates or acquaintances, and a meager 0.2% had sent insulting or abusive messages or e-mails.

Thus, while aggressive online behavior is believed to be universal, in reality, the proportion of people who engage in it is insignificant. When online disinhibition is conceptualized as a specific behavior, research focuses on a small minority of society, making it challenging to measure online disinhibition among Internet users.

Originally, Suler (2004) and Barak, Boniel-Nissim and Suler (2008) described online disinhibition as deriving from the unique objective features of CMC (primarily nonverbal) that differ from traditional (primarily verbal) communication. In other words, people are impacted by online disinhibition to some extent as long as they are in CMC situations. Suler (2004) indicated that online disinhibition is a pervasive phenomenon. The rate of Internet use in Japan has grown to 83.4% (as of 2020), accompanied by the recent popularization of smartphones, tablet terminals, and other digital devices (cf. Ministry of Internal Affairs and Communications, 2021). Moreover, owing to the coronavirus disease 2019 pandemic (COVID-19), remote work and online classes have become necessary, and CMC comprises a growing portion of daily communication. Experiences related to online disinhibition may become ubiquitous in this social context. Therefore, in psychological research, it is prudent to consider the construct of online disinhibition from a perspective that recognizes its universality.

2.2.3. Mental state perspective: A breakthrough construct of online disinhibition

The discussion above clarifies that the onset and behavioral perspectives have significantly contributed to the interdisciplinary applications of online disinhibition theory. However, a more appropriate construct is needed if the theory is to be further developed and utilized in psychological research. We believe that it is most appropriate to define online disinhibition from the perspective of mental state. Thus, this study defines online disinhibition as "a mental state in which the cognition to inhibit behavior lessens or disappears in an online environment."

In this approach, the six factors indicated by Suler (2004), including dissociative anonymity, invisibility, asynchronicity, dissociative imagination, solipsistic introjection, and minimization of status and authority, as well as those investigated by later researchers, such as a lack of eye contact (Lapidot-Lefler & Barak, 2012; Lapidot-Lefler & Barak, 2015), reduced nonverbal cues (Schouten et al., 2007), and controllability (Schouten et al., 2007), are considered potential determinants of online disinhibition. For example, being comfortable with perpetrating poor behavior because it is difficult to identify the perpetrator of an act in an online environment is considered typical of the online disinhibition effect. From the mental state perspective, "difficulty identifying the perpetrator" is considered an objective feature of the online environment (cause) and is distinguished from the online disinhibition of "accepting poor behavior" (effect). It is undoubtedly easy for "difficulty in identifying the perpetrator" to lead to "accepting poor behavior," but this does not mean that people will necessarily "accept

poor behavior" when they enter an environment where it is difficult to identify the perpetrator. Making a clear distinction between cause and effect enables researchers to examine the process whereby "difficulty in identifying the perpetrator" leads to "accepting poor behavior" and identify the impact of other factors, such as personality traits. In other words, it can resolve the limitations observed of the onset perspective.

Furthermore, engaging in behaviors online that are inhibited in the real world should be regarded as "disinhibitive behavior" occurring from online disinhibition. However, this does not suggest that a person in a mental state of online disinhibition will necessarily manifest specific disinhibitive behaviors. In the example given above, even if someone has entered a mental state of online disinhibition wherein they are "accepting poor behavior," this does not suggest they would invariably behave poorly. Therefore, online disinhibition can be viewed as an essentially universal mental state that underlies disinhibitive behavior even if the probability of engaging in a specific behavior is extremely low. In fact, it can be considered a more generalized concept that can help resolve the issues observed regarding the behavioral perspective.

The mental state perspective can help resolve conceptual confusion regarding online disinhibition. One pioneering study from this perspective is that of Schouten et al. (2007), who measured online disinhibition using a simple three-item scale. Stuart and Scott (2021) defined online disinhibition 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." They developed a more specific one-factor, 11-item scale known as the Measure of Online Disinhibition. An investigation of the tool's validity showed that the stronger one's online disinhibition, the greater the likelihood of online trolling behavior or self-disclosure.

Stuart and Scott's (2021) evaluation tool is based on a definition of online disinhibition similar to that argued by the present study, and therefore lends a certain level of validity to the mental state perspective. However, as noted by Stuart and Scott (2021) and Cheung et al. (2021), the construct of online disinhibition includes multiple components (e.g., human public self-consciousness and consciousness of social norms). This simple one-factor evaluation tool can provide a primary outline of online disinhibition, but requires further investigation. Using this tool as a starting point for a more detailed investigation of the complex components involved in online disinhibition is a future direction for psychological research on online disinhibition.

3. Influencing mechanism of online disinhibition

Various attempts have been made to construct a theoretical model of the influencing mechanism of online disinhibition on human behavior. The contributions and limitations of a few major models are as follows:

3.1. Benign or toxic disinhibition model

Suler (2004) explained the characteristic behaviors arising in online settings from the perspective of "inhibition removal" and proposed a dichotomous model that categorizes online disinhibition as benign or toxic based on the positive or negative result of disinhibition removal. Benign disinhibition is believed to lead to behaviors such as casting off psychological defenses, expressing oneself freely to others, and engaging in prosocial interactions, whereas toxic disinhibition leads to rude language, harsh criticism, and threatening behavior to achieve blind catharsis (Figure 2).

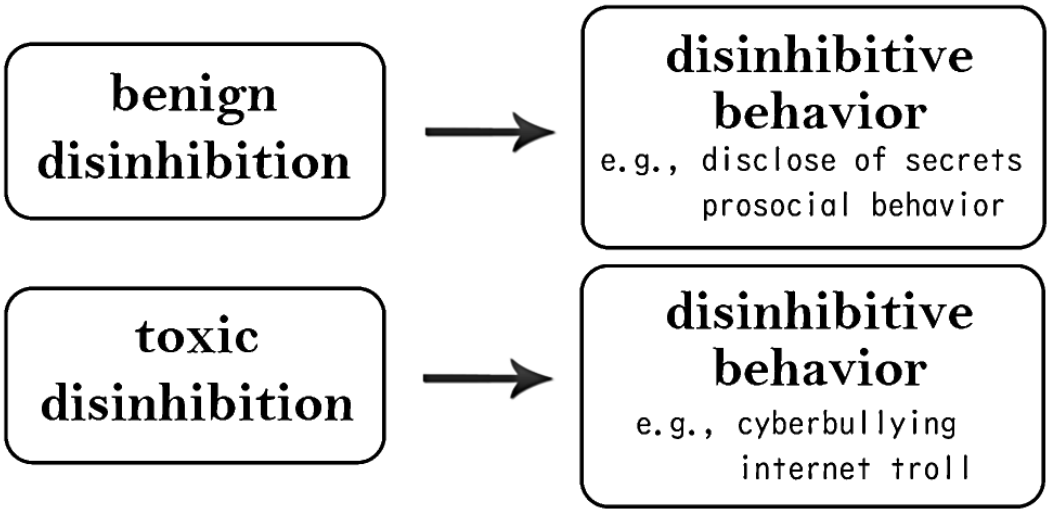

Figure 2 Benign or toxic disinhibition model

This model suggests that online disinhibition can drive behavior toward two distinct extremes; however, this dichotomy is ambiguous. Suler (2004) mentioned the ambiguity, which was also observed in subsequent studies. For example, hostile words in chat encounters could be a therapeutic breakthrough for some people (Suler, 2004). In such cases, toxic disinhibition can positively affect communication. Conversely, Udris (2014) found that both toxic and benign disinhibition can exacerbate cyberbullying. In other words, the benign or toxic disinhibition model can dichotomize online disinhibition as a construct but can be difficult to discriminate in actuality.

Moreover, Suler's (2004) theory suggests that whether online disinhibition is considered toxic depends on whether it leads to aggressive behavior, making both benign and toxic disinhibitions resultant concepts. This indicates that online disinhibition cannot be classified as benign or toxic unless it progresses to observable and concrete behavior. Hence, the benign or toxic disinhibition model faces recursive definition concerns, wherein results are predicted using the resultant concepts. Considering this issue, recent studies have argued that online disinhibition should not be classified as good or bad (e.g., Stuart & Scott, 2021). Classifying

the outcomes that result from the mental state of disinhibition as good or bad can help intuitively understand its impact; however, this dichotomy should be reconsidered as a subject of psychological study.

3.2. Online disinhibition—behavior model

The online disinhibition-behavior model improves upon the dichotomous model and does not classify online disinhibition or its outcomes as good or bad. In this model, online disinhibition determines the disinhibitive behavior (Figure 3).

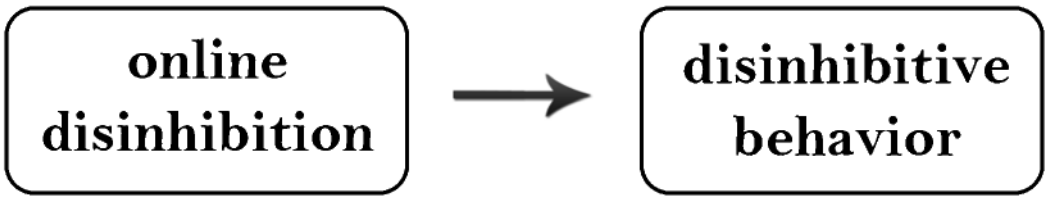

Figure 3 Online disinhibition—behavior model

Rather than engaging with the complex or ambiguous determination of whether online disinhibition is good or bad, the online disinhibition-behavior model postulates a straightforward relationship between them based on the mental state perspective described earlier. Based on this viewpoint, the model demonstrates the relationships between online disinhibition and disinhibitive behaviors such as Internet trolling (e.g., Kurek et al., 2019; Stuart & Scott, 2021), cyberbullying experiences (e.g., Stuart & Scott, 2021), online self-disclosure (Schouten et al., 2007), sending sexual information, and experiences of online sexual harassment (Hernández, Schoeps, Maganto, & Montoya-Castilla, 2021).

The online disinhibition-behavior model applies to an extensive range of subjects because it does not address complex relationships. However, the straightforward approach also has disadvantages because it cannot reflect autonomy in human behavior. Schouten et al. (2007) argue that CMC is often used to fulfill one's need for self-disclosure. Some researchers point out that when experiencing confusion about their identity, many adolescents voluntarily use virtual spaces to practice dissociative self-presentation to bridge the gap between their actual and ideal selves (e.g., Kurek et al., 2019; Michikyan, Dennis, & Subrahmanyam, 2015). In such scenarios, it is believed that they are not unwittingly self-disclosing due to the influence of the online environment but intentionally select the online environment to fulfill their need for self-disclosure and identity exploration. However, the online disinhibition-behavior model cannot reflect this autonomy. It is essential to refine the model further to achieve higher construct validity.

3.3. Relationship between online disinhibition theory and deindividuation theory

A critical collateral theory that considers the influencing mechanism of online disinhibition is deindividuation theory, as discussed below.

3.3.1. Online disinhibition and deindividuation theory

Attempts have been made to theorize how the potential for antisocial behavior increases when an individual joins a group and has increased anonymity before the rise of the Internet(e. g. Diener, 1980; Le Bon, 1895; Zimbardo, 1969).

According to Zimbardo's (1969) deindividuation theory, when the self becomes immersed in a group, antisocial behavior is tolerated to a large degree, leading to aggressive behaviors that are usually inhibited under normal circumstances. In this context, anonymity is considered a determinant of deindividuation. However, subsequent replication studies have yielded inconsistent results regarding anonymity's impact. For example, in an artificially manipulated deindividuation scenario, participants who donated a nurse's uniform were significantly more prosocial (Johnson & Downing, 1979).

An improved theory, known as the social identity model of deindividuation effects (SIDE model), was proposed (Spears & Lea, 1994) to explain such results. The SIDE model emphasizes the importance of situational norms in groups and defines deindividuation as a mental situation wherein situational group norms overlap general social norms (cf. Vilanova, Beria, Costa, & Koller, 2017). In group settings, people begin to identify more with situational group norms because situational social identity is heightened. In other words, situational group norms determine whether one's behavior will become prosocial or antisocial. The SIDE model has become a prominent theory for explaining deviant behavior in CMC settings (Vilanova et al., 2017).

Deindividuation and online disinhibition are both frequently cited when explaining the cause of the many instances of deviant online behavior because they increase the likelihood of an individual behaving antisocially. Lowry et al. (2016) presented a hypothesized model suggesting that people enter a state of disinhibition and de-individuation in anonymous online environments, leading to cyberbullying via social learning. Rösner & Krämer (2016) deemed online disinhibition a type of deindividuation in their analysis of the theoretical context surrounding the phenomenon of emotional venting using aggressive language online. Spears (2017), a proposer of the SIDE model, also argues that when people with similar interests or perspectives join an online environment, the characteristics of the group become salient relative to other characteristics, making individuals more likely to deviate from social norms to adhere to the norms of that group. Spears (2017) did not approach this phenomenon using the term online disinhibition; however, the content and context bear a strong resemblance.

Attempts to cross-sectionally integrate online disinhibition and deindividuation theories have considerable significance for the unified development of both theories and for providing a more multifaceted explanation of widespread deviant Internet behavior.

3.3.2. Distinguishing online disinhibition and deindividuation

There are similarities between online disinhibition and deindividuation, although the two constructs are not identical. However, studies such as those of Lowry et al. (2016) used

expressions like “disinhibition and deindividuation,” and thus do not clearly demonstrate their distinguishing features. This ambiguity in terminology further lowers the validity of both constructs. Therefore, this study attempts to differentiate between the constructs of online disinhibition and deindividuation.

While online disinhibition and deindividuation may sometimes result in highly similar outcomes, they should be considered distinct concepts. As discussed earlier, online disinhibition is derived from CMC, which is inherently different from traditional communication, and may arise in all CMC situations to some extent. Conversely, deindividuation theory, from both Zimbardo’s (1969) perspective and the subsequent SIDE model, focuses on people behaving in more deviant ways in group settings. In other words, an individual’s anonymity through becoming a group member is considered a basic premise of deindividuation. This relationship is illustrated in Figure 4. Consider the following two typical examples of online disinhibition: When people with similar interests come together in an online environment, or a large group of people attacks a specific person or organization in online flaming, the impact of online disinhibition and deindividuation may be strikingly similar. This is because group identity is salient in both situations. However, this does not imply that the same phenomenon occurs in all online settings. For example, social media may provide attractive opportunities for self-presentation. Social media users are faced with a continuous demand for endless self-focused activities such as taking pictures, adding friends, and following people they like (Wallace, 2015). In such an environment, individuality and self-expression are emphasized, and the aim is to set oneself apart from others, reinforcing individual identity. In this case, deindividuation is unlikely to occur even in the presence of online disinhibition because the prerequisite group setting is not present.

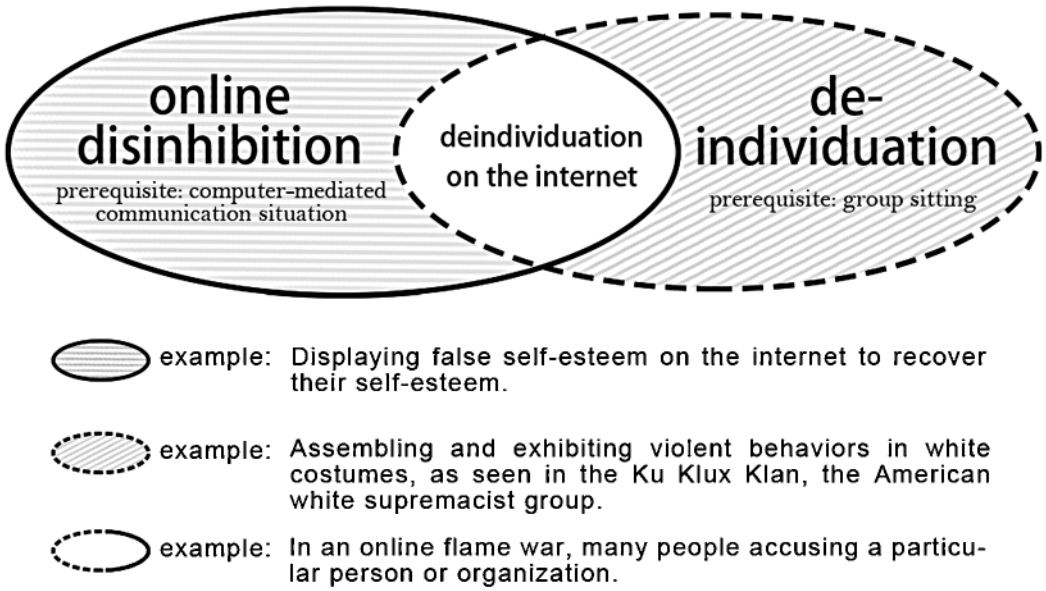

Figure 4 The relationship between online disinhibition and deindividuation

Wu et al. (2017) conducted an empirical study that partially supports the relationship between online disinhibition and deindividuation. They conclusively demonstrated that a state of deindividuation, a decrease in private self-awareness online, is a determinant of antisocial disinhibitive behavior, similar to the asynchronicity and dissociative imagination indicated by Suler (2004). In other words, in contexts where deindividuation plays a significant role in removing inhibition, deindividuation, and online disinhibition yielded highly similar outcomes. Based solely on the results, this suggests that no major problems would arise even if the two were not differentiated. Nevertheless, deindividuation may not occur in individual settings such as online self-presentation or self-disclosure; therefore, it is essential to differentiate the two.

4. Motivation-based online disinhibition model

This review demonstrates that, although online disinhibition theory has achieved a certain level of growth supported by various proposed models and evidence, it should be refined further for psychological research. This study proposes a motivation-based online disinhibition (MOD) model based on the mental state perspective of online disinhibition discussed in Section 2 to compensate for these deficits.

As shown in Figure 5, the MOD model considers the impact of online disinhibition and the crucial role that motivation for a specific action plays in the process of disinhibitive behavior. Moreover, online disinhibition is considered a moderator, not a determinant of disinhibitive behavior . The model assumes that people do not passively act in disinhibitive ways because they are impacted by the attributes of the online environment but that they intentionally utilize various online services. The desire for self-disclosure and exploration of personal identity impact behavior as intrinsic motivators. Conversely, experiencing unexpected abuse from strangers, encountering difficulties in the real world, and reading outrageous news online are

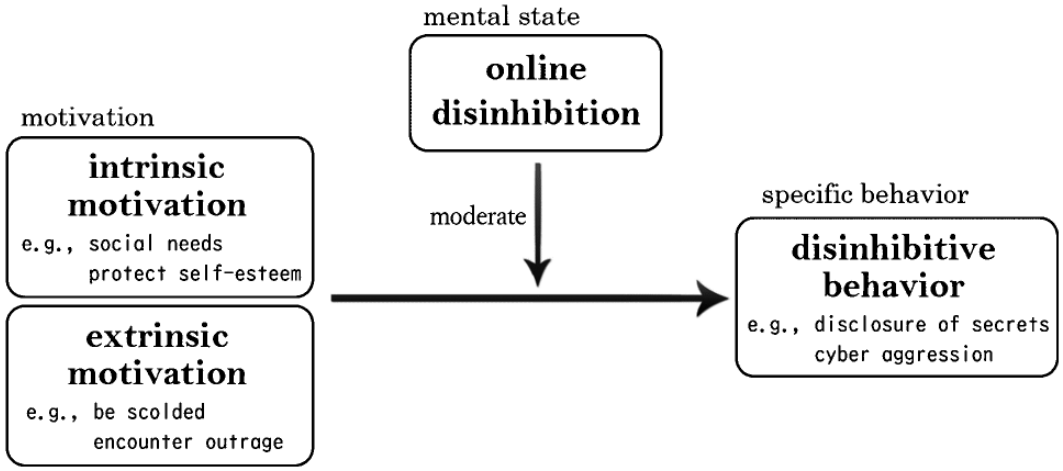

Figure 5 Motivation-based online disinhibition model

common occurrences. Such experiences become extrinsic motivators, leading to attempts to distract oneself through prosocial or antisocial behaviors. In other words, in the process of disinhibitive behavior, intrinsic or extrinsic motivators or a combination of the two determines what specific disinhibitive behavior a person will perform to a significant extent, rather than online disinhibition.

People desire to act in specific ways, either intrinsically or in response to external stimuli; in the real world, this desire is usually inhibited by the cognition of various social phenomena. For example, concern about one's own appearance, the act of speaking, or negative responses from others, such as frowns or sighs, may inhibit the desire to express oneself (Suler, 2004). This inhibition works as the "floodgate" through which desire is allowed to transform into action. However, this inhibition floodgate effect lessens or disappears in online spaces (in the example above, the invisibility of the online environment eliminates these concerns), making it easier to move from desire to action. This mechanism suggests that online disinhibition plays a moderating role, whereas motivation plays a determinant role.

Several studies offer empirical evidence for this stance. Chan (2021) hypothesized that social anxiety would promote online self-disclosure and that the process would be moderated by online disinhibition, and verified this using online survey data of social media users. The results confirmed that the interaction between online disinhibition and social anxiety positively affects online self-disclosure. When social anxiety is high, self-disclosure may occur regardless of the online disinhibition level; when social anxiety is low, self-disclosure may occur with high online disinhibition. Yang et al. (2021) hypothesized that associating with peers who engage in deviant behavior gradually increases tolerance for poor behavior through social learning and may lead to cyberbullying. They found that associating with people who engage in deviant behavior promotes cyberbullying, and that higher online disinhibition strengthens this effect. In this study, socializing with people who engage in deviant behavior was an extrinsic motivator of cyberbullying, and online disinhibition moderated the influence of this motivator, resulting in cyberbullying. These two studies used the scales described by Schouten et al. (2007) and Udris (2014). The construct validity of these scales should be reevaluated and verified, but the findings offer partial support for the MOD model's assertion that online disinhibition is not a determinant, but a moderator.

The discussion above explains how the MOD model solves the problem of the online disinhibition-behavior model overlooking human autonomy by incorporating the influence of motivation. However, the tools used to measure online disinhibition in these studies were not based on adequately valid constructs. Future studies should empirically verify the relationships between the variables proposed in the MOD model using a scale of online disinhibition with higher construct validity.

5. Conclusion

This study discusses the construct of online disinhibition and the mechanism influencing behaviors, emphasizing Suler's (2004) online disinhibition theory considering the current online environment. First, it demonstrated that the construct of online disinhibition has thus far been investigated from three perspectives, and argued that online disinhibition should be considered a mental state in psychological research. Subsequently, the study reviewed critical theoretical models concerning how online disinhibition impacts human behavior and outlined their respective contributions and limitations. Finally, the MOD model was proposed, addressing the gaps in existing theoretical models and providing significant explanatory power in psychological research.

This study shows that diverse perspectives on the construct of online disinhibition have been intermixed in the literature. Therefore, the construct validity of the scales developed thus far requires further examination. The Measure of Online Disinhibition (Stuart & Scott, 2021) evaluates online disinhibition as a mental state; however, its exploration in online disinhibition needs to be further improved, and the development of a precise tool for evaluating online disinhibition is still limited. In the future, it will be necessary to test the validity of our proposed MOD model using a more accurate and specific scale.

Acknowledgment

The authors would like to thank Editage (www.editage.com) for English-language editing. This study was supported by a grant from the Telecommunication Advancement Foundation.

References

Barak, A., Boniel-Nissim, M., & Suler, J. (2008). Fostering Empowerment in Online Support Groups. *Computers in Human Behavior, 24*(5), 1867-1883. https://doi.org/10.1016/j.chb.2008.02.004

Bryce, J., & Fraser, J. (2013). "It's Common Sense That It's Wrong": Young people's perceptions and experiences of cyberbullying. *Cyberpsychology, Behavior, and Social Networking, 16*(11), 783-787. https://doi.org/10.1089/cyber.2012.0275

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls Just Want to Have Fun. *Personality and Individual Differences, 67,* 97-102. https://doi.org/10.1016/j.paid.2014.01.016

Chan, T. K .H. (2021). Does Self-disclosure on Social Networking Sites Enhance Well-Being? The role of social anxiety, online disinhibition, and psychological stress. In Lee, Z. W. Y., Chan, T. K. H., & Cheung, C. M. K. (Eds.) , *Information Technology in Organisations and Societies: Multidisciplinary Perspectives from AI to Technostress* (pp. 175-202). Bingley: Emerald

Publishing Limited. https://doi.org/10.1108/978-1-83909-812-320211007

Cheung, C. M. K., Wong, R.Y.M., & Chan, T. K. H. (2021). Online Disinhibition: Conceptualization, measurement, and implications for online deviant behavior. *Industrial Management & Data Systems*, *121*(1), 48-64. https://doi.org/10.1108/IMDS-08-2020-0509

Cox, T., Marczak, M., Teoh, K., & Hassard, J. (2017). New Directions in Intervention: Cyber-bullying, schools and teachers. In Teresa Mendonça McIntyre, Scott E. McIntyre, David J. Francis, *Educator Stress* (pp. 411-435). Cham: Springer. https://doi.org/10.1007/978-3-319-53053-6_17

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

DePaolis, K., Williford, A. (2015). The Nature and Prevalence of Cyber Victimization Among Elementary School Children. *Child Youth Care Forum 44*(3), 377–393. https://doi.org/10.1007/s10566-014-9292-8

Diener, E. (1980) Deindividuation: The absence of self-awareness and self-regulation in group members. In: Paulus, P.B. (Eds.), *Psychology of Group Influence*(pp. 209-242). London: Psychology Press.

Heirman, W., Walrave, M., Vandebosch, H., Wegge, D., Eggermont, S., & Pabian, S. (2016). Cyberbullying Research in Belgium: An overview of generated insights and a critical assessment of the mediation of technology in a web 2.0 world. In Raúl Navarro, Santiago Yubero, Elisa Larrañaga (Eds.), *Cyberbullying Across the Globe*. Cham: Springer. https://doi.org/10.1007/978-3-319-25552-1_9

Hernández, M. P., Schoeps, K., Maganto, C., & Montoya-Castilla, I. (2021). The Risk of Sexual-erotic Online Behavior in Adolescents: Which personality factors predict sexting and grooming victimization?. *Computers in Human Behavior, 114*, 106569. https://doi.org/10.1016/j.chb.2020.106569

Hollenbaugh, E. E., & Everett, M. K. (2013). The Effects of Anonymity on Self-disclosure in Blogs: An application of the online disinhibition effect. *Journal of Computer-Mediated Communication, 18*(3), 283-302. https://doi.org/10.1111/jcc4.12008

Huang, C. L., Zhang, S., & Yang, S. C. (2020). How Students React to Different Cyberbullying Events: Past experience, judgment, perceived seriousness, helping behavior and the effect of online disinhibition. *Computers in Human Behavior*, *110*. https://doi.org/10.1016/j.chb.2020.106338.

Johnson, R. D., & Downing, L. L. (1979). Deindividuation and Valence of Cues: Effects on prosocial and antisocial behavior. *Journal of Personality and Social Psychology*, *37*(9), 1532-1538. https://doi.org/10.1037/0022-3514.37.9.1532

Joinson, A. N. (2001). Self-disclosure in Computer-mediated Communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, *31*(2), 177-192.

https://doi.org/10.1002/ejsp.36

Joinson, A. N. (2007). Disinhibition and the Internet. In Jayne Gackenbach (Ed.), *Psychology and the Internet* (pp. 75-92). Cambridge, MA: Academic Press. https://doi.org/10.1016/B978-012369425-6/50023-0

Koyama, K., Asatani, K., Sakaki, T., Sakata, I. (2019). Resonance Structure of Online Flaming: The epicenter of information sharing on online flaming. *The 33rd Annual Conference of the Japanese Society for Artificial Intelligence.* https://doi.org/10.11517/pjsai.JSAI2019.0_2E5J602

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*, 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. In *Proceedings of the 3rd Multidisciplinary International Social Networks Conference on Social Informatics 2016, Data Science 2016*. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2955129.2955138

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). https://doi.org/10.5817/CP2015-2-3

Le Bon, G. (1897). *The crowd: A study of the popular mind*. London: T. Fisher Unwin.

Lowry, P. B., Zhang, J., Wang, C., & Siponen, M. (2016). Why Do Adults Engage in Cyberbullying on Social Media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model. *Information Systems Research, 27*(4), 962-986. https://doi.org/10.1287/isre.2016.0671

Masui, K., Tamura, A., March, E. (2018). Development of a Japanese Version of the Global Assessment of Internet Trolling-Revised. *Japanese Psychological Research, 89*(6), 602-610. https://doi.org/10.4992/jjpsy.89.17229

Michikyan, M., Dennis, J., & Subrahmanyam, K. (2015). Can You Guess Who I Am? Real, ideal, and false self-presentation on Facebook among emerging adults. *Emerging Adulthood, 3*(1), 55-64. https://doi.org/10.1177/2167696814532442

Ministry of Internal Affairs and Communications (2021). WHITE PAPER Information and Communication in Japan. Retrieved from https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r03/pdf/01point.pdf

Rösner, L., & Krämer, N. C. (2016). Verbal Venting in the Social Web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media + Society, 2*(3). https://doi.org/10.1177/2056305116664220.

Saunders, K. C. (2016). A Double-edged Sword: Social media as a tool of online disinhibition regarding American Sign Language and Deaf cultural experience marginalization, and as a tool of cultural and linguistic exposure. *Social Media + Society*, *2*(1). https://doi.org/10.1177/2056305115624529

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

Sobba, K. N., Paez, R. A., & Ten Bensel, T. (2017). Perceptions of Cyberbullying: An assessment of perceived severity among college students. *TechTrends, 61*(6), 570-579. https://doi.org/10.1007/s11528-017-0186-0

Spears, R., & Lea, M. (1994). Panacea or Panopticon? The hidden power in computer-mediated communication. *Communication Research*, *21*(4), 427–459. https://doi.org/10.1177/009365094021004001

Spears, R. (2017). Social identity model of deindividuation effects. In Rössler, P (Eds.), *The international encyclopedia of media effects*, Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781118783764.wbieme0091

Stuart, J., & Scott, R. (2021). The Measure of Online Disinhibition (MOD): Assessing perceptions of reductions in restraint in the online environment. *Computers in Human Behavior*, *114*, 106534. https://doi.org/10.1016/j.chb.2020.106534

Suler, J. (2004). The Online Disinhibition Effect. *Cyberpsychology & Behavior, 7*(3), 321-326. https://doi.org/10.1089/1094931041291295

Tokioka, R., Satou, U., Kodama, N., Tazuke, K., Takenaka, Y., Matsunami, M.... & Kuwabara, T. (2017) A Study of High School Students' Cognition about Communication through LINE and the Effect of Friendships among Modern Adolescents on It. *The Japanese Journal of Personality, 26*(1), 76-88. http://doi.org/10.2132/personality.26.1.7

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

Udris, R. (2015). Cyberbullying in Japan: An exploratory study. *International Journal of Cyber Society and Education, 8*(2), 59-80. http://dx.doi.org/10.7903/ijcse.1382

Vilanova, F., Beria, F. M., Costa, Â. B., & Koller, S. H. (2017). Deindividuation: From Le Bon to the social identity model of deindividuation effects. *Cogent Psychology, 4(1)*, 1308104. https://doi.org/10.1080/23311908.2017.1308104

Wallace, P. (2015). *The psychology of the Internet*. Cambridge University Press.

Wang, X., Wang, W., Qiao, Y., Gao, L., Yang, J., & Wang, P. (2020). Parental Phubbing and Adolescents' Cyberbullying Perpetration: A moderated mediation model of moral disengagement and online disinhibition. *Journal of Interpersonal Violence*, *37*(7-8). https://doi.org/10.1177/0886260520961877.

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*(1), 41-45. https://doi.org/10.1016/j.paid.2018.04.001

Wright, M. F., & Wachs, S. (2021). Does Empathy and Toxic Online Disinhibition Moderate The Longitudinal Association Between Witnessing And Perpetrating Homophobic Cyberbullying?. *International Journal of Bullying Prevention, 3*(1), 66-74. https://doi.org/10.1007/s42380-019-00042-6

Wu, S., Lin, T. C., & Shih, J. F. (2017). Examining the Antecedents of Online Disinhibition. *Information Technology & People, 30*(1), 189-209. https://doi.org/10.1108/ITP-07-2015-0167

Yamaguchi, S. (2015). An Empirical Analysis of Actual Examples of "Flaming" and Participants' Characteristics. *Journal of Information and Communication Research, 33*(2), 53-65. https://doi.org/10.11430/jsicr.33.2_53

Yang, J., Wang, N., Gao, L., & Wang, X. (2021). Deviant Peer Affiliation and Adolescents' Cyberbullying Perpetration: Online disinhibition and perceived social support as moderators. *Children and Youth Services Review, 127*, 106066. https://doi.org/10.1016/j.childyouth.2021.106066

Zimbardo, P. G. (1969). The Human Choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In *Nebraska Symposium on Motivation*. Lincoln, NE: University of Nebraska Press.