International Conference on Psychology UMBY
ISSN: 0033-3077

UNDERSTANDING BENIGN AND TOXIC ONLINE DISINHIBITION: THE ROLE OF SELF-ESTEEM IN GENERATION Z SOCIAL MEDIA USERS

Lenissia Marie Roessli[1]*, Rahma Widyana[2]
1,2 Universitas Mercu Buana Yogyakarta, Indonesia
*marieleni330@gmail.com

ABSTRACT

The rapid expansion of social media has intensified the manifestation of online disinhibition, where individuals communicate more freely, openly, or impulsively in digital spaces. This phenomenon is especially prevalent among Generation Z, a cohort of digital natives whose online behaviors are strongly intertwined with their psychological characteristics. This study investigates the role of self-esteem in shaping two distinct dimensions of online disinhibition—benign and toxic—among Generation Z social media users. Using a quantitative correlational design and purposive sampling, data were collected from 224 participants. Self-esteem was measured using the Rosenberg Self-Esteem Scale, while online disinhibition was assessed using the Online Disinhibition Scale (ODS). Spearman's rank correlation was employed to analyze the relationships among variables. The findings revealed no significant correlation between self-esteem and overall online disinhibition ($r = –0.068$, $p = 0.318$). However, self-esteem showed a significant positive association with benign disinhibition ($r = 0.132$, $p = 0.049$) and a significant negative association with toxic disinhibition ($r = –0.204$, $p = 0.002$). These results emphasize the multidimensional nature of online disinhibition and indicate that self-esteem may differentially influence individuals' adaptive and maladaptive online behaviors. The implications for digital well-being, online behavior interventions, and future research are discussed.

Keywords: generation z social media users; self-esteem; online disinhibition; benign disinhibition; toxic disinhibition

Introduction

Social media has become a primary platform for digital interaction and the construction of self-identity in modern society. Rapid technological advancements have transformed the ways individuals communicate, share information, and build social relationships. In Indonesia, internet use has increased significantly in recent years (Rahmadani, 2023). According to the APJII (2024) report, the number of internet users has surpassed 221,563,479 individuals, with a penetration rate of 79.5%. Generation Z represents the group with the highest usage rate at 87.02%, reinforcing their position as digital natives who have grown up alongside technology. The intensive engagement of Generation Z with social media makes them particularly vulnerable to psychological dynamics in online interactions. A lack of self-control may lead to inappropriate behavior and difficulties in forming social relationships,

both in offline and digital contexts (Fitriansyah, 2018). Social media provides an interaction space that allows individuals to feel freer in expressing themselves, particularly due to anonymity and physical separation from communication partners. This condition indicates that technological development not only changes the way individuals communicate, but also affects emotional regulation, behavior, and social norms in everyday interactions, especially within online environments (Haqie et al., 2024).

One prominent phenomenon in online behavior is online disinhibition, which refers to the tendency of individuals to express thoughts, emotions, or behaviors more freely in digital environments compared to face-to-face settings (Suler, 2004). Suler distinguishes two forms of online disinhibition: benign disinhibition, which includes positive expressions such as openness and empathy, and toxic disinhibition, encompassing negative behaviors such as insults, verbal aggression, and cyberbullying. These contrasting expressions highlight that online disinhibition is not a unitary construct but a multidimensional phenomenon that shapes the quality of digital interactions. Empirical studies have demonstrated the psychological consequences of online disinhibition among adolescents and young adults. For example, toxic disinhibition has been shown to mediate the relationship between life stress and cyberbullying behavior (Chu et al., 2023), whereas benign disinhibition is associated with emotional expression and the receipt of social support in online contexts. Despite this conceptual distinction, many empirical studies continue to operationalize online disinhibition as a single construct, potentially obscuring meaningful psychological differences between adaptive and maladaptive online behaviors. As a result, findings that rely on aggregated disinhibition scores may lack conceptual clarity and offer limited explanatory value for understanding diverse online behavioral outcomes.

Online disinhibition is influenced by several situational and psychological factors, including dissociative anonymity, invisibility, asynchronicity, solipsistic introjection, dissociative imagination, and the minimization of status and authority (Suler, 2004). These conditions reduce direct social pressure and external regulation, allowing individuals to feel freer in expressing thoughts and emotions in online environments compared to face-to-face interactions. In such contexts, internal psychological factors become increasingly important in shaping online behavior. One such factor is self-esteem, which has been identified as a key psychological variable influencing how individuals use social media as a space for self-expression. Niemz et al. (2005) suggested that self-esteem plays a role in individuals' preference for online versus offline communication, indicating that self-evaluative processes may influence how people respond to reduced social constraints in digital interactions.

Self-esteem refers to an individual's overall evaluation of personal worth and capability (Coopersmith, 1967). Rosenberg (1965) conceptualizes self-esteem as comprising two core components: self-acceptance, which reflects the ability to acknowledge both strengths and weaknesses, and self-respect, which refers to the extent to which individuals value themselves positively. As a central aspect of psychological functioning, self-esteem has been widely associated with patterns of self-expression and interpersonal behavior, including behavior in digital environments.

Previous research has examined the relationship between self-esteem and online disinhibition among Generation Z populations. For instance, Ariesandy and Ariana (2021) found that Generation Z Twitter users with higher self-esteem were more likely to express themselves openly in online interactions. Similarly, Fitria (2021) reported that adolescents with higher self-esteem tended to exhibit online disinhibition behaviors more frequently when using social media. However, these studies largely conceptualized online disinhibition as a single, undifferentiated construct, without distinguishing between its benign and toxic forms.

This unidimensional approach represents an important limitation, as benign and toxic online disinhibition may be driven by different psychological processes and may relate to self-esteem in distinct ways. Treating online disinhibition as a single construct risks obscuring opposing behavioral tendencies and may contribute to inconsistent or ambiguous empirical findings. This limitation is particularly critical for Generation Z, who represent the most active and digitally immersed cohort and are simultaneously more exposed to both positive and problematic online behaviors. Without a clearer understanding of how individual psychological factors, such as self-esteem, relate to specific forms of online disinhibition, interventions aimed at improving digital well-being may remain overly general and less effective.

To address this gap, the present study examines the relationship between self-esteem and online disinhibition by explicitly distinguishing between benign and toxic dimensions among Generation Z social media users. By adopting a multidimensional perspective, this study seeks to provide a more precise understanding of the psychological dynamics underlying online behavior and to inform the development of evidence-based digital literacy programs, preventive interventions, and educational initiatives that promote healthy and ethical online engagement.

Based on the theoretical framework and previous empirical findings, the following hypotheses are proposed:

H1: There is a significant relationship between self-esteem and benign online disinhibition among Generation Z social media users.

H2: There is a significant relationship between self-esteem and toxic online disinhibition among Generation Z social media users.

Methods

This study involved 224 Generation Z respondents (born 1995–2010, aged 15–30) who are active social media users. Participants were selected using purposive sampling based on social media usage intensity and willingness to participate (Sugiyono, 2015). Participants were selected using purposive sampling, with inclusion criteria consisting of (1) belonging to Generation Z (aged 15-30), (2) being active users of social media platforms. This sampling technique was considered appropriate for the present correlational study, as it ensured the inclusion of participants who were directly relevant to the research objectives.

Two instruments were employed: the Online Disinhibition Scale (ODS) and the Rosenberg Self-Esteem Scale. The ODS was adapted from Udris (2014) and translated into Indonesian by Mantara et al. (2023). It consists of 11 items, measuring benign online disinhibition (7 items) and toxic online disinhibition (4 items), rated on a five-point Likert scale. As the ODS is a recently adapted instrument, no separate pilot testing was conducted prior to data collection; however, reliability analysis in the present study indicated good internal consistency (Cronbach's $\alpha = 0.82$). Self-esteem was measured using the Rosenberg Self-Esteem Scale (RSES) developed by Rosenberg (1965) and adapted by Azwar (1979). A pilot test conducted on 87 respondents resulted in the removal of one item due to low item–total correlation, yielding a final 9-item scale with satisfactory reliability (Cronbach's $\alpha = 0.846$).

Data were collected online via Google Forms, and all participants provided informed consent prior to participation. Participation was voluntary and anonymous to ensure confidentiality. A quantitative correlational approach was employed to examine the relationship between self-esteem and online disinhibition. Preliminary assumption testing indicated that the data were not normally distributed; therefore, Spearman's rank-order correlation was used to analyze the relationships between self-esteem and benign online disinhibition as well as toxic online disinhibition, using SPSS version 27.

Results

Descriptive Statistics

This study involved 224 participants who are part of Generation Z (aged 15–30 years). The descriptive statistics for each variable are presented in Table 1.

Table 1. Descriptive Stastistic of Main Variable

Variable	N	Min	Max	Mean	SD
Online Disinhibition	224	22	53	40.55	6.975
Self-Esteem	224	29	45	33.12	5.772

Table 1 provides an overview of distribution of scores for the two main variables. The descriptive analysis shows that the empirical scores of Online Disinhibition range from 22 to 53, with a mean of 40.55 and a standard deviation of 6.975. Meanwhile, Self-Esteem scores range from 29 to 45, yielding a mean of 33.12 and a standard deviation of 5.772. These findings indicate adequate score variability for both variables, allowing further analysis to be conducted.

Table 2. Descriptive Stastistic by Subdimensions

Variable	N	Min	Max	Mean	SD

Benign Disinhibition	224	14	35	28.63	4.017
Toxic Disinhibition	224	4	20	11.91	4.771
Self-Esteem	224	10	45	33.16	5.784

Table 2 provides a more specific understanding of respondents' online behavior. Empirical scores of Benign Disinhibition range from 14 to 35, with a mean of 28.63 and a standard deviation of 4.017. Toxic Disinhibition scores range from 4 to 20, with a mean of 11.91 and a standard deviation of 4.771. Self-Esteem scores show a range of 10 to 45, with a mean of 33.16 and a standard deviation of 5.784. All empirical ranges fall within the theoretical limits, indicating that the scale functions appropriately for further analysis.

Correlation Analysis

Spearman's rank-order correlation was used to examine the relationships between self-esteem and online disinhibition. The analysis revealed no significant relationship between self-esteem and overall online disinhibition ($r = –0.068$, $p = 0.312$), as shown in Table 3.

Table 3. Hypothesis Testing Test (Online Disinhibition & Self-Esteem)

Variable	R (Spearman)	p-value	Conclusion
Self-Esteem – Online Disinhibition	-0.068	0.312	Not significant

When analyzed by dimension (Table 4), self-esteem showed a significant positive correlation with benign online disinhibition ($r = 0.132$, $p = 0.049$) and a significant negative correlation with toxic online disinhibition ($r = –0.204$, $p = 0.003$). These findings indicate that self-esteem is differentially associated with adaptive and maladaptive forms of online disinhibition.

Table 4. Hypothesis Testing Test (Online Disinhibition & Self-Esteem)

Variable	R (Spearman)	p-value	Conclusion
Self-Esteem – Benign Disinhibition	0.132	0.049	Significant
Self-Esteem – Toxic Disinhibition	-0.204	0.003	Significant

Discussions

The findings of this study indicate that the total score of online disinhibition does not show a significant relationship with self-esteem. This suggests that when benign and toxic disinhibition are combined, the opposing directions of their correlations cancel each other out, resulting in a non-significant overall association. This underscores Suler's (2004) assertion that online disinhibition is a multidimensional construct, encompassing both

adaptive and maladaptive components. By distinguishing between benign and toxic disinhibition, the present study advances Suler's framework by demonstrating how these dimensions operate differently in the context of Generation Z, highlighting the nuanced psychological mechanisms that a unidimensional approach may overlook.

A more informative pattern emerges when the dimensions are analyzed separately. Benign disinhibition demonstrates a significant positive relationship with self-esteem, indicating that individuals with higher self-esteem are more likely to express openness, empathy, and supportive behavior in digital interactions. This finding extends Coopersmith's (1967) and Santrock's (2011) perspectives, suggesting that stable self-esteem not only enhances confidence in self-expression but also facilitates prosocial engagement in online settings. Mechanistically, this may be explained by better emotion regulation, a stable self-concept, and lower reliance on external validation, which allow individuals to interact online without fear of judgment. Anonymity and psychological distance in digital environments may further enable prosocial behaviors, as suggested by Lapidot-Lefler and Barak (2015).

In contrast, toxic disinhibition is negatively associated with self-esteem. Individuals with lower self-esteem are more prone to aggressive, impulsive, or hostile behaviors online, such as trolling, offensive comments, or rash postings. This aligns with Suler's (2004) disinhibition model and is supported by Verma et al. (2023), Putri and Pratama (2023), and Pickhardt (2013), who emphasize that poor emotion regulation, self-hate, and low self-esteem contribute to maladaptive online behaviors. These findings highlight that combining benign and toxic disinhibition into a single score, as done in prior studies (e.g., Ariesandy & Ariana, 2021; Fitria, 2021), may obscure opposing behavioral tendencies and lead to misleading conclusions.

Practically, these findings have important implications for digital literacy programs, cyberbullying prevention, and interventions targeting Generation Z. Programs can be tailored to strengthen self-esteem and emotion regulation, fostering benign disinhibition while mitigating toxic tendencies. Educators and practitioners can use these insights to develop platform-specific guidelines and activities that encourage positive online engagement.

This study has several limitations. First, the reliance on self-report measures may be subject to social desirability bias. Second, the cross-sectional correlational design limits causal inference, and cultural specificity may constrain generalizability beyond Indonesian Generation Z users. Third, the quantitative approach provides limited insight into the contextual and psychological processes underlying online disinhibition. Future research could adopt longitudinal, experimental, or mixed-method designs, or focus on platform-specific behaviors, to better understand the mechanisms and causal pathways linking self-esteem and the two dimensions of online disinhibition.

Conclusion

In conclusion, this study demonstrates that self-esteem influences online disinhibition differently across its dimensions: higher self-esteem is associated with benign online

disinhibition, while lower self-esteem is linked to toxic disinhibition. By distinguishing between these two forms, the study contributes both theoretically, by clarifying the psychological mechanisms underlying online behavior-and practically, by informing interventions aimed at promoting healthier digital interactions among Generation Z. Future research could build on these findings by examining specific social media platforms or particular online behaviors to better understand how self-esteem shapes digital interactions.

Acknowledgement

The authors would like to express their gratitude to all participants who generously contributed their time to this research. Appreciation is also extended to the academic advisors and colleagues who provided valuable guidance and feedback throughout the research process. The support from the Faculty of Psychology, Universitas Mercu Buana Yogyakarta, is gratefully acknowledged.

References

Anggraini, D., Nurmayasari, M., & Saripah, S. (2023). Penggunaan media sosial Tik Tok dan pengaruhnya terhadap motivasi berprestasi siswa SMK Al Khairiyah Bahari Jakarta. *Jurnal Pendidikan Tambusai*, *7*(1), 2239-2244.

Ariesandy, S. (2023). *Hubungan antara self-esteem dan online disinhibtion effect pada generasi z pengguna media sosial Twitter* (Doctoral dissertation). Universitas Airlangga, Surabaya, Indonesia.

Asosiasi Penyelenggara Jasa Internet Indonesia. (2024). *Survei penetrasi & perilaku pengguna internet Indonesia tahun 2024*. Retrieved from https://apjii.or.id/

Bernadine, J., & Astuti, N. W. (2024). Hubungan antara school well-being dan self-esteem dalam keberhasilan nilai belajar siswa. *JLEB: Journal of Law, Education and Business*, *2*(1), 648-659.

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

Chu, X., Li, Q., Fan, C., & Jia, Y. (2023). Life stress and cyberbullying: examining the mediating roles of expressive suppression and online disinhibition. *Journal of Youth and Adolescence*, *52*(8), 1647-1661.

Coopersmith, S. (1967). *The antecedents of self-esteem*. San Francisco, CA: W.H. Freeman and Company.

Correa, T., Hinsley, A. W., & de Zuniga, H. G. (2010). Who interacts on the web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, *26*(2), 247-253.

Dalifa, P. A. (2021). Hubungan antara parent attachment dengan self-esteem pada mahasiswa di Sumatera Barat. *Jurnal Pendidikan Tambusai*, *5*(2), 3621-3626.

Dirmayanti, S., & Setiasih, S. (2024). Role of social media uses and social comparison on self-esteem. *Jurnal Psikologi Tabularasa*, *19*(2), 239-243.

Firamadhina, F. I. R., & Krisnani, H. (2020). Perilaku generasi Z terhadap penggunaan media sosial TikTok: TikTok sebagai media edukasi dan aktivisme. *Share Social Work Journal*, *10*(2), 199-208.

Fitriah, A., & Hariyono, D. S. (2019). Hubungan self-esteem terhadap kecenderungan depresi pada mahasiswa. *Psycho Holistic*, *1*(1), 8-17.

Gil de Zúñiga, H., Diehl, T., Huber, B., & Liu, J. (2017). Personality traits and social media use in 20 countries: How personality relates to frequency of social media use, social media news use, and social media use for social interaction. *Cyberpsychology, Behavior, and Social Networking*, *20*(9), 540-552.

Haqie, D. A., Hapsari, W., & Karsiyati, K. (2024). Peran anonimitas dan konsep diri terhadap online disinhibition effect pada remaja. *Jurnal Ilmiah Wahana Pendidikan*, *10*(16), 238–252.

Hasanati, U., & Aviani, Y. I. (2020). Hubungan social comparison dengan self-esteem pada pengguna Instagram. *Jurnal Pendidikan Tambusai*, *4*(3), 2391-2399.

Joinson, A. N. (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). Academic Press.

Kernis, M. H. (2005). Measuring self-esteem in context: The importance of stability of self-esteem in psychological functioning. *Journal of Personality, 73*(6), 1569–1605.

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).

Mantara, A. Y., Sa'id, M., Zahra, G. A., Rizkina, A. T., Febriyanti, L., & Prastika, S. B. (2023). Adaptation of the online disinhibition effect scale. *KnE Social Sciences*, 317–323.

Matondang, A. F. (2023). The relationship between hoax behavior and toxic disinhibition among Indonesian high school students. *Journal of Educational, Health & Community Psychology (JEHCP)*, *12*(4).

Mayliyan, A. K., Marizza, H. M., Azizah, N., & Budiarto, E. (2023, January). Gambaran self-esteem warga binaan dengan kasus penyalahgunaan NAPZA di Rutan. In *Prosiding University Research Colloquium* (pp. 33–35).

Mueller-Coyne, J., Voss, C., & Turner, K. (2022). The impact of loneliness on the six dimensions of online disinhibition. *Computers in Human Behavior Reports*, *5*, 100169.

Naibaho, F., Agustina, V. F., & Wijayani, M. R. (2022). Pengaruh kontrol diri terhadap perilaku disinhibition online effect di komunitas remaja Gereja Santo Nikodemus ciputat. *Afeksi: Jurnal Psikologi*, *1*(2), 263-273.

Putri, M. A., & Pratama, M. D. (2023). Pada remaja, kebencian terhadap diri sendiri, dapat memicu perilaku toxic disinhibition online. *Jurnal Ilmiah Psikologi Mind Set*, *2*(01), 112-118.

Ramadhani, A. Z., & Merida, S. C. (2024). Self-control and the phenomenon of toxic online disinhibition in teenagers who have Twitter. *Nusantara Journal of Behavioral and Social Sciences*, *3*(2), 45-52.

Santrock, J. W. (2008). *A topical approach to life-span development: Motor, sensory, and perceptual development* (pp. 172–205). McGraw-Hill Higher Education.

Satriawan, N. (2016). *Hubungan antara konsep diri dengan toxic disinhibition online effect pada siswa SMK N 8 Surakarta.* (Undergraduate Thesis) Universitas Sebelas Maret Surakarta, Surakarta, Indonesia.

Sherman, L. E., Payton, A. A., Hernandez, L. M., Greenfield, P. M., & Dapretto, M. (2016). The power of the like in adolescence: Effects of peer influence on neural and behavioral responses to social media. *Psychological Science*, *27*(7), 1027-1035.

Sholikin, R. A. P. (2019). Hubungan antara gambaran diri dengan disinhibition effect pada remaja. Program DIII Keperawatan.

Steinberg, L. (2002). *Adolescence* (6th ed.). McGraw-Hill.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, *7*(3), 321-326.

Suriani, N., & Jailani, M. S. (2023). Konsep populasi dan sampling serta pemilihan partisipan ditinjau dari penelitian ilmiah pendidikan. *IHSAN: Jurnal Pendidikan Islam, 1*(2), 24-36.

Triananda, S. F., Dewi, D. A., & Furnamasari, Y. F. (2021). Peranan media sosial terhadap gaya hidup remaja. *Jurnal Pendidikan Tambusai*, *5*(3), 9106-9110.

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

Valkenburg, P. M., & Peter, J. (2011). Online communication among adolescents: an integrated model of its attraction, opportunities, and risks. *Journal of Adolesc Health.* 48(2),121-7. doi:10.1016/j.jadohealth.2010.08.020.

Van den Bos, K., Van Lange, P. A., Lind, E. A., Venhoeven, L. A., Beudeker, D. A., Cramwinckel, F. M., ... van der Laan, J. (2011). On the benign qualities of behavioral disinhibition: Because of the prosocial nature of people, behavioral disinhibition can weaken pleasure with getting more than you deserve. *Journal of Personality and Social Psychology*, *101*(4), 791.

Verma, A., Islam, S., Moghaddam, V., & Anwar, A. (2024). Digital emotion regulation on social media. *Computer*, *57*(6), 82-89.

Wibowo, Y., & Silaen, S. M. J. (2018). Hubungan self-esteem dan penggunaan media sosial instagram dengan perilaku narsisme di kalangan siswa kelas VIII SMPK Penabur Bintaro Jaya. IKRA-ITH Humaniora: *Jurnal Sosial Dan Humaniora*, *2*(2), 109-115.

Winch, R. F. (1965). Rosenberg: Society and the adolescent self-image (book review). *Social forces*, *44*(2), 255.

Zeigler-Hill, V. (2013). *Self-esteem*. Psychology Press.