Profiling the Offline and Online Risk Experiences of Youth to Develop Targeted Interventions for Online Safety

ASHWAQ ALSOUBAI, Vanderbilt University, USA and King AbdulAziz University, KSA
AFSANEH RAZI, Drexel University, USA
ZAINAB AGHA, Vanderbilt University, USA
SHIZA ALI, Boston University, U.S.A
GIANLUCA STRINGHINI, Boston University, U.S.A
MUNMUN DE CHODHURY, Georgia Institute of Technology, U.S.A
PAMELA J. WISNIEWSKI, Vanderbilt University, USA

We conducted a study with 173 adolescents (ages 13-21), who self-reported their offline and online risk experiences and uploaded their Instagram data to our study website to flag private conversations as unsafe. Risk profiles were first created based on the survey data and then compared with the risk-flagged social media data. Five risk profiles emerged: Low Risks (51% of the participants), Medium Risks (29%), Increased Sexting (8%), Increased Self-Harm (8%), and High Risk Perpetration (4%). Overall, the profiles correlated well with the social media data with the highest level of risk occurring in the three smallest profiles. Youth who experienced increased sexting and self-harm frequently reported engaging in unsafe sexual conversations. Meanwhile, high risk perpetration was characterized by increased violence, threats, and sales/promotion of illegal activities. A key insight from our study was that offline risk behavior sometimes manifested differently in online contexts (i.e., offline self-harm as risky online sexual interactions). Our findings highlight the need for targeted risk prevention strategies for youth online safety.

CCS Concepts: • **Human Computer Interaction**; • **Youth Online Safety**; • **Youth Risk Profiles** → Mixture Factor Analysis;

Additional Key Words and Phrases: Empirical analysis, profiles

ACM Reference Format:
Ashwaq Alsoubai, Afsaneh Razi, Zainab Agha, Shiza Ali, Gianluca Stringhini, Munmun De Chodhury, and Pamela J. Wisniewski. 2024. Profiling the Offline and Online Risk Experiences of Youth to Develop Targeted Interventions for Online Safety. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 114 (April 2024), 37 pages. https://doi.org/10.1145/3637391

1 INTRODUCTION

Modern-day youth are found to experience a myriad of online and offline risks such as substance misuse [4], cyberbullying [13], or unwanted online sexual experiences, including exposure to

Authors' addresses: Ashwaq Alsoubai, ashwaq.alsoubai@vanderbilt.edu, atalsoubai@kau.edu.sa, Vanderbilt University, Nashville, Tennessee, USA and King AbdulAziz University, Jeddah, Saudi Arabia, KSA; Afsaneh Razi, afsaneh.razi@drexel.edu, Drexel University, Philadelphia, Pennsylvania, USA; Zainab Agha, zainab.agha@vanderbilt.edu, Vanderbilt University, Nashville, Tennessee, USA; Shiza Ali, shiza@bu.edu, Boston University, Boston, Massachusetts, U.S.A; Gianluca Stringhini, gian@bu.edu, Boston University, Boston, Massachusetts, U.S.A; Munmun De Chodhury, munmund@gatech.edu, Georgia Institute of Technology, 30318, Atlanta, Georgia, U.S.A; Pamela J. Wisniewski, pamela.wisniewski@vanderbilt.edu, Vanderbilt University, Nashville, Tennessee, USA.

ACM 2573-0142/2024/4-ART114
https://doi.org/10.1145/3637391

sexually explicit materials [74]. The research community has shown a great interest in understanding the boundaries between the offline and online behaviors of youth, mainly in the context of these risk experiences [39]. In fact, offline and online risk experiences of youth were found to be driven by the same underlying factors related to the general propensity to experience risks [51]. There is well-established research that has focused on examining the correlation between youth online and offline risk experiences such as between online sexual risks and self-harm [86, 124] or substance misuse [17], or between online harassment and self-harm [113] or substance misuse [96, 137]. This line of research has examined such correlations by primarily conducting survey-based and longitudinal studies that used self-reported data from youth, which could be subject to recall and social desirability biases [120]. To overcome these biases, Pinter et al. [97] called for using social media trace data to have an in-depth understanding of youths' lived experiences online. Several researchers have used digital trace data, like social media data, to examine the online risk experiences of youth (e.g., [10, 102, 112, 139]). By studying youth's public or semi-public online disclosures, these studies uncovered nuances in youth's online risk experiences such as how consented sexual conversations between youth and their friends or significant others were laden by indicators of mental health problems (e.g., threats of self-harm) [10], which could not be precisely captured through self-reports only. However, these studies have not gone back to understand how these insights correlate with the social science literature that is based primarily on self-reports. In addition, relying on public or semi-public social media data has limited the ecological validity of research studies, given the interactions youth have online vary dramatically depending on various factors such as whether they are conducted in public versus private spaces [78, 130]. Therefore, this study fills these gaps by correlating the self-reports of youths' offline and online risk experiences with their social media "private" conversations.

Additionally, while the current efforts for online safety education and prevention initiatives [1, 2] are valuable in raising youth awareness about online safety, such efforts are often designed for general populations of youth, rather than being tailored to the unique differences and needs of subgroups of youth who share similar offline and online experiences. Yet, targeted education strategies and intervention-based programs for youth have consistently been found to positively affect their personal and social competence, self-awareness, confidence, and relationships [79]. Therefore, there is a need to personalize online safety prevention and education strategies for youth to improve these outcomes. Profiling youth based on their similar risk experiences is one way to achieve this level of personalization. In this paper, we addressed the following research questions:

- **RQ1:** *What are the differing profiles of youth based on their self-reported offline and online risk behaviors (i.e., online harassment, sexting, self-harm, and offline risk behavior)?*
- **RQ2:** *How do these risk profiles correlate and/or differ when compared to youths' risk-flagged private Instagram trace data?*
- **RQ3:** *How do unsafe Instagram conversations differ linguistically between the different risk profiles?*

To address these questions, we collected survey and social media data from youth (N = 173). First, we administered a web-based survey using pre-validated psychological constructs to measure the frequency in which youth reported Risky Behavior Questionnaire [14], Inventory of Statements About Self-harm [64], Cyber-Aggression Victimization [116], Cyber-Aggression Perpetration [116], Unwanted Sexual Solicitations and Approaches [84], and Youth Produced Sexual Images (Sexting) [84]. Then, we asked the participants to upload their social media (Instagram) data into our study web system and to self-assess their private conversations and flag messages that made them feel uncomfortable or unsafe. The participants were between the ages of 13 and 21 (avg.= 17 yrs, Std.=2.15). To answer RQ1, we conducted a Mixture Factor Analysis (MFA) to create youth profiles

based on the self-reported risk experiences. Profiling is one of the common approaches to identify groups of a human or nonhuman subject based on analyzing the correlations between data [58]. For RQ2, we conducted a between-group analysis using a (χ^2) test to uncover key differences between the youth profiles based on their risk-flagging behaviors. We then used machine learning generative models of text to identify the unique linguistic differences between the profiles' unsafe conversations to answer RQ3.

Through these analyses, we identified five unique youth risk profiles: 1) Low Risks (51% of the participants), 2) Medium Risks (29%), 3) Increased Sexting (8%), 4) Increased Self-Harm (8%), and 5) High Risk Perpetration (4%). Low Risks profile who self-reported the least number of online and offline risk experiences, was found to be exposed to and flagged a wider variety of spam and scam messages and the self-harm disclosures of others. Medium Risks profile self-reported medium levels of risk experiences, yet, this profile mostly encountered harassment in the form of sexual flirtations and harassing comments. The rest of the profiles faced higher levels of risk experiences. Youth in the Increased Sexting profile self-reported the most sexting and often flagged in-person meeting requests they received within their private sexual conversations. Youth in the Increased Self-Harm profile reported the highest levels of offline self-harm, but their unsafe conversations did not contain digital self-harm content; instead, they engaged in more unsafe sexual conversations, including sugar daddy relationships with strangers. Youth that belonged to the High Risk Perpetration profile reported the highest average scores of online harassment perpetration and offline risk behaviors, including conflicts they initiated offline that escalated online, as well as promotions of illegal products.

The ACM Special Interest Group on Computer-Human Interaction (SIGCHI) and CSCW communities have been studying adolescents' risk experiences in networked spaces. Our work makes several important contributions for these communities by providing insights and improving our understanding of youth online safety. We work toward building a comprehensive understanding of how youth experience offline and online risks by triangulating their self-report survey data with their private social media data. In summary, this work makes the following novel research contributions:

- It identifies correlations between youths' self-reported responses and their Instagram digital trace data to address the methodological gaps within the literature, which contributes to the body of works related to youth online safety.
- It illustrates the importance of acknowledging the multi-dimensional nature of youth online and offline risk behaviors to identify unique clusters of youth.
- It improves our understanding of youth risk assessments and perceptions by aligning their explicit risk flagging to their self-reports, the evidence of which can help provide youth more agency for their online/offline safety.
- It demonstrates how some offline behaviors of youth were manifested differently online. We provide implications for designing safer online interactions for youth based on their different profiles.

2 BACKGROUND

In this section, we synthesize literature on youths' salient offline and online risk behaviors and the connection between the two. Then, we highlight the knowledge gaps that motivate the importance of delineating multidimensional risk profiles of youth that contextualize their self-reports of risks with their own social media trace data.

2.1 Identifying the Salient Offline and Online Risk Behaviors and Experiences of Youth

A plethora of prior works have examined the offline risk behaviors of youth, including alcohol consumption [81], substance misuse [46, 89], and self-harm [28, 121]. One line of research has focused on investigating the prevalence, or consequences of substance misuse on youths' future, the likelihood of committing risky behaviors [21, 89], mental well-being [81, 91, 104], and school performance [27, 71]. While another line of research has extensively focused on youths' mental well-being [73], mainly by examining the antecedent contextual factors that motivate youths' non-suicidal self-harm, defined as "the deliberate destruction of one's body tissue without suicidal intent and for purposes not socially sanctioned" [64]. These research efforts provided insights into the significant role of the youths' community environment such as family [90, 107, 138], peer influence [46, 57], traumatic experiences such as child sexual abuse [3], and social isolation or bullying experiences [77, 114]. During the COVID-19 pandemic, several reviews have reiterated the importance of monitoring youth offline risk behaviors to identify important trends over time [67, 73]. For instance, while studies on offline risk behaviors have highlighted a decline in some problematic behaviors, such as alcohol misuse, among youth [67], others have indicated an increase in youth self-harming behaviors [73]. Given the prevalence of research on the offline risk behaviors of youth, we deemed it important to also study these behaviors along with youths' online risk experiences.

Meanwhile, the youth online safety literature has identified the most prevalent online risks that youth encounter, including online harassment or cyberbullying, sexual risks, such as sexual solicitations, or exposure to explicit content [31, 74, 133]. Along with establishing the prevalence of such risks amongst youth, prior work has also examined the contextual factors when studying cyberbullying and online sexual risks to better understand youths' online risks. For cyberbullying, prior work has investigated the impact that youth roles have on their online harassment as victims [38, 99], perpetrators [125, 136], or as victim-perpetrators (i.e., dual experiences) [23, 69, 100]. For instance, prior work has shown how the victims and victim-perpetrator of cyberbullying had less social-emotional intelligence [60] as well as less social preference [53] compared to perpetrators, which illustrates the importance of studying youth online harassment experiences by disentangling such roles and understanding their unique differences. While for online sexual risks, scholars have emphasized investigating other factors such as distinguishing between online sexual risks and interpersonal sexting [10], which are often grouped together and studied under the same lens [48, 86]. Interpersonal sexting is defined as sending or receiving sexually explicit messages, especially images, online [37]. Although it is well known that youth sexting is important for their sexual development [98], it is often coupled with adverse outcomes such as sexual risk behaviors, coercion, and substance misuse that could negatively impact youth in their adulthood [17, 55, 86]. These negative effects amplify the importance of examining youth sexting experiences yet as a separate measure from the online sexual risks. These prior works studying youths' online risk experiences show that it is important to have a nuanced understanding of these behaviors and experiences to examine the interplay of the contextual factors that contribute to these risks. This study builds upon the online risk literature by adopting a more nuanced lens in measuring online risk behaviors of youth by disentangling online sexual risks from interpersonal sexting, as well as capturing the experiences of cyberbullying from both the perspectives of victimization and perpetration. Most importantly, our work examines both the offline and online risk behaviors of youth together, rather than focusing on one or the other. We describe how we operationalized and measured these constructs in our methods (Section 3.2). Next, we establish how past studies often examine the relationship between offline and online risk behaviors of youth.

2.2 Studying the Relationship between the Offline and Online Risk Behaviors and Experiences of Youth

Prior research has studied correlations or causal relationships between analogous online and offline risk behaviors of youth [12, 36, 140]. Most of these studies focused on the negative impact of social media use and/or exposure to online risk content on offline risk behaviors. For instance, Davis et al. [36] conducted a multi-wave longitudinal study over ten years to collect self-reported data for measuring the impact of online exposure to substance-related content on the substance misuse behavior of 4,840 youths' (aged 12 to 22). They found that exposure to substance misuse content in social media was associated with an increase in alcohol drinking amongst youth. Arendt et al. [12] also conducted a two-wave panel survey with 729 young adults and found that youth who were exposed to self-harm content online were at higher risk of experiencing self-harm or suicide. Overall, this body of literature has emphasized concerns related to youth use of social media, mainly related to normalizing their perception of risks or reinforcing such behaviors in their offline environment. In contrast, other researchers have focused on how analogous online behaviors manifested offline or vice versa [16, 26]. Several studies have investigated how bullying and cyberbullying behaviors of youth impact their well-being [26] or to what extent online and offline sexual behaviors of youth were prevalent [16]. Examining the associations or correlations between distinct online and offline behaviors is also well-established within the online safety literature [106, 124, 137]. For instance, Wachs et al. [124] conducted a survey with 2506 adolescents (ages 13–16 years) and found that online sexual risks (i.e., pressured sexting) were positively and significantly correlated with offline non-suicidal self-harm. While Yoon et al. [137] have also surveyed 10th-grade students (N = 2,768) and after the 12-month follow-up, they found that youth who experienced cyberbullying, regardless of their roles as witnesses, perpetrators, or victims, had higher odds of substance misuse.

This body of literature demonstrates that it is worthwhile to study online and offline risk behaviors in unison, especially when analogous behaviors cross the boundary between the virtual and physical worlds. Yet, offline and online risk behaviors are likely bidirectional in nature, meaning that engaging in one type of behavior may influence the other, with a mutually causal relationship [132]. Further, engaging in one risk behavior may influence or increase the likelihood of engaging in other types of risk behaviors (e.g., cyberbullying with increased sexual risks [47]), which is not directly related. Various individual and contextual factors, such as personality traits [24], sociocultural influences [44], and personal circumstances [122] can contribute to the likelihood and directionality of these relationships. Therefore, taking a multidimensional approach to understanding youths' offline and online risk behaviors is important to look beyond analogous online to offline risk behaviors to examine possible *unanticipated* correlations between these different risk behaviors that have not been uncovered in the past literature.

2.3 Creating Multidimensional Risk Profiles of Youth and Triangulating Youth Self-Reported Risk Behaviors with Their Risk-Flagged Social Media Data

In examining different approaches that account for the multidimensionality of offline and online behaviors, several works have presented promise in using profiling methods [56, 119]. There is a growing body of knowledge within youth safety literature that has attempted to elucidate unique profiles of adolescents to better understand the heterogeneity in this population based on their distinct risks experiences [18, 56, 63, 119]. Using Latent Class Analysis (LCA), Bishop et. al [18] identified four profiles of substance misuse for gang-involved youth: Non-Users (38%), Past Users (15%), Casual Users (27%), and Frequent Multi-Users (21%). These profiles revealed a nuanced understanding of the differences among the gang-involved youth in their substance misuse along with the ecologies that either promoted or inhibited certain patterns of misuse, against the common

perceptions that all youth in this population are users. Recently, emergent works have acknowledged the importance of incorporating the offline and online context of risks when creating youth profiles to provide a more holistic understanding of such risks. For instance, Kim et al. [62] created profiles for adolescents based on their offline and online bullying behaviors. Four profiles emerged: (1) Low Risk (85.3%), who reported the lowest levels of engaging in both offline and cyberbullying, (2) High Risk (2.4%), who showed high levels of engagement in both bullying and victimization online and offline, (3) Offline Risks (5.1%), who had high scores for offline bullying, but low scores for cyberbullying, and (4) Online-Risk Group (7.2%), who reported high scores of engagement in the online domain, but low scores in offline bullying. Through these profiles, they were able to confirm the co-occurrence in the roles of bullying (i.e., victim and perpetration) across the online and offline contexts. These studies highlight the value of examining youths' heterogeneity related to their risk experiences, which would help scholars and practitioners to better understand the dynamics of risks in the youth population and therefore delineate targeted and evidence-based intervention initiatives and education plans. Our work builds upon this prior literature by adopting a similar approach to profiling youth based on a myriad of different risk experiences (as described in Section 2.1 and operationalized in Section 3.2). Therefore, we contribute to the literature by moving beyond a narrow view of a subset of related risks to studying a wider array of risks, including offline risk behaviors, offline self-harm, unwanted online sexual risks, online harassment roles (e.g., perpetrator vs. victim), and online sexting.

Additionally, we build upon and expand on existing literature by addressing the gap identified in the youth online safety literature by Pinter et al. [97], regarding the heavy reliance on youth self-report data. Since this gap was identified, several scholars have leveraged digital trace data as a way to examine the online lived experiences of youth [54, 102, 139]. Yet, a limitation of this approach was that online behaviors may not always be indicative of offline reality, and digital trace data by itself could easily be taken out of context. Online behaviors can be influenced by various factors, such as anonymity, disinhibition, and the absence of immediate consequences, which may lead youth or individuals, in general, to engage in behaviors that differ from their offline behaviors [72, 125]. Yet, to our knowledge, research has not yet addressed the alignment or misalignment between youth self-reports regarding their risk behaviors and digital trace data indicative of these experiences. Therefore, we address this gap by triangulating youth self-report data with their digital trace social media data (RQ2), which is a novel contribution to the literature. Such triangulation could be used to cross-validate findings from different data sources and establish whether or not there are statistically significant correlations between youth self-report and the actual experiences that occur online. These findings could help identify consistencies and discrepancies between self-reported data and digital trace data, providing a more accurate and nuanced picture of individuals' risk behaviors and enhancing the validity and reliability of future research.

Beyond triangulating self-report data with digital trace data when analyzing youth risk behaviors, this approach yields another significant contribution. Contextualizing the self-reported risk profiles (RQ1) with more nuanced digital trace data (RQ3) provides a more holistic understanding of the profiles' risk experiences beyond the measured behaviors. Prior research mainly presented profiles of youth without going beyond the statistical interpretations of these profiles to thicker qualitative descriptions. As such, relying solely on self-reported measures to interpret the profiles might fail to detect nuances of the lived experiences of youth and to provide insights into the specific contexts of these experiences [7, 108]. In summary, we take a data-driven empirical approach by triangulating self-reports of youth about their online and offline risk experiences with their social media (i.e., Instagram) data to inform an in-depth understanding of such experiences.

3 METHODS

Below, we give an overview of our study, followed by a detailed account of our research methods.

3.1 Study Overview

We developed a secure web-based system, where participants first completed a web-based survey; then, they were asked to login into their Instagram accounts to download their Instagram data. These files included profile information, direct message conversations, including posted photos and videos, archived stories, comments and likes, followers and following accounts, recent searches, and history of shopping and advertisement. For this study, we only processed and analyzed the private messages conversations, as the riskiest interactions most likely occur in private online spaces [5]. We selected Instagram because it is the most popular social media platform after YouTube and TikTok among youth [13]. As Pew Research recently found that six in ten teens engaged as active users of the platform [13]. Instagram also enables users to easily download their data based on the General Data Protection Regulation (GDPR) [49], which mandates social media companies to allow users to download their own personal data. After uploading their Instagram data to our secured system, participants were asked to review their private message conversations, flag messages that made them or someone else feel uncomfortable or unsafe, and provide contextual information (e.g., what happened, with whom) about the interaction. In the subsections below, we provide more details about the survey constructs and Instagram data donation procedure.

3.2 Survey Design

To measure youth risk experiences, we drew from prior literature (Section 2.1) and utilized pre-validated survey measures, including Risky Behavior Questionnaire [14], Inventory of Statements About Self-harm [64], Cyber-Aggression Victimization [116], Cyber-Aggression Perpetration [116], Unwanted Sexual Solicitations and Approaches [84], and Youth Produced Sexual Images (Sexting) [84] (Appendix A). The Likert scale of these measures was from 1-5 (1- Never, 2- Rarely, 3-Sometimes, 4- Often, 5- All the time). We slightly updated the wording of the question for the online risk experience constructs to ask specifically about the participants' experiences when using Instagram, rather than in general. Below, we explain the measures in more detail.

Offline Risk Behaviors. In this study, we leveraged the Risky Behavior Questionnaire (RBQ) scale by Auerbach and Gardiner [14] to measure the frequency of a myriad of offline risk behaviors during adolescence, including substance misuse, unsafe sex, cheating, and gambling. Using this construct gave us a more holistic view of youth risk experiences rather than only focusing on their online behaviors. Non-suicidal self-harm behavior has become more prevalent recently, especially among vulnerable populations like youth [75]; therefore, we included the Inventory of Statements About Self-harm (ISAS) scale by Klonsky and Glenn [64], which was designed to understand the non-suicidal self-harm behaviors. This scale quantifies the frequency of intentional youths' self-harming behaviors, including cutting, scratching, and hitting away from suicidal reasons.

Online Risks behaviors. We included The Cyber-Aggression Victimization (CAV) scale developed by Shapka and Maghsoudi [116], which measured youth experiences of online harassment in different forms such as receiving hurtful comments, gossip about them, or having an embarrassing post, photo, or video on Instagram. To better understand youth risk experiences, we not only used the online harassment construct of youth as victims but also included the online harassment perpetration, where the youth were the perpetrators. To do this, we adopted the Shapka and Maghsoudi [116] scale for Cyber-Aggression Perpetration (CAP) to measure the prevalence of online harassment experiences that youth committed online. The questions were similar to the

Cyber-Aggression Victimization scale but were rephrased to be about the participants committing harassment instead of being victims. We utilized the Unwanted Sexual Solicitations and Approaches scale from the Youth Internet Safety Survey (YISS) developed by Mitchel et al. [84] to measure the unwanted online sexual risks, including sexual messages, requests to engage in sexual activities and/or sexual conversations, and unexpected exposures to nude pictures or people having sex. This scale combines two scales, which were the Unwanted sexual solicitations and approaches and Unwanted exposure to pornography, into one scale called Unwanted Online Sexual Risks. We did not include the harassment questions to avoid repeating the cyberbullying questions from the CAV and CAP scales, which were more comprehensive of the online harassment experiences. We also used Mitchel et al.'s [84]Youth Produced Sexual Images (Sexting) construct from the same YISS. This scale consists of five questions three of them about the possession and distribution of digital imagery depicting nudity of a minor (under the age of 18). Since the possession and distribution of such materials is considered a federal crime, we did not ask our participants these questions in the survey. Therefore, this measure will mainly quantify the production and distribution of youth sexual imagery or videos, particularly whether they send or receive any personal nude/semi-nude media (pictures or videos).

Type	Measures	No. Items	Cronbach's alpha	Mean	SD	Skewness	Kurtosis
Risk Experiences	**Offline Risk Behaviors**	20	0.82	1.54	0.38	1.20	2.71
	Offline Self-harm	12	0.85	1.55	0.62	1.58	2.14
	Unwanted Online Harassment	12	0.90	1.99	0.72	1.01	0.43
	Online Harassment Perpetration	12	0.90	1.28	0.44	2.62	7.21
	Unwanted Sexual Experiences	5	0.84	2.18	0.88	0.36	-0.69
	Interpersonal Online Sexting	2	0.76	1.4	0.73	2.24	0.43

Table 1. Measures descriptive statistics.

Table 1 summarizes the descriptive statistics of the pre-validated constructs used in this study. Cronbach's α's, which measures the internal consistency of survey constructs [33], was higher than the acceptable 0.7 threshold [29]. More importantly, the Interpersonal Online and the Unwanted Online Sexual scales remained reliable after the changes we did with Cronbach's α of 0.76 and 0.84 respectively. Table 2 demonstrates the significant positive correlations between the risk experience measures, which suggest that youth who encountered one type of risk experience were also more likely to encounter another.

Type	Measure	RBQ	ISAS	CAV	CAP	YISS	SEXT
Risk Experiences	**Offline Risk Behavior (RBQ)**	1					
	Offline Self-Harm (ISAS)	.466**	1				
	Unwanted Online Harassment (CAV)	.378**	.224*	1			
	Online Harassment Perpetration (CAP)	.283**	.183*	.367**	1		
	Unwanted Sexual Experiences (YISS)	.476**	.218*	.554**	.275**	1	
	Interpersonal Online Sexting (SEXT)	.399**	.204*	.393**	.234**	.506**	1

Table 2. The correlation between the risk experiences measures (Unwanted Online Sexual Risks, Interpersonal Online Sexting, Unwanted Online Harassment, Online Harassment Perpetration, Offline Self-harm, and Offline Risk Behaviors). All risk experiences measures were significantly positively correlated with each other. * p-value < 0.05, ** < 0.01, and *** < 0.001

Lastly, the participants also answered demographic questions about their sex, age, race, and sexual orientation.

3.3 Instagram Data Collection

Once participants successfully uploaded their Instagram files, their direct conversations were displayed back to them in reverse chronological order to review their past conversations. For the participants who are 18 years and older, their conversations when they were under 18 were shown to them first to make this process faster for participants. For each unsafe conversation, they were then asked to review and flag the unsafe messages and evaluate them based on the risk severity levels (i.e, low, medium, high), which were adopted based on prior literature [134], and risk types. The risk types included harassment, sexual messages or solicitations/nudity, hate speech/threat of violence, sale or promotion of illegal activities, digital self-injury, or spam, which were derived from Instagram reporting feature risk categories[1]. For the unsafe conversations, the participants were also asked to specify their relationship with the other person involved as a stranger, acquaintance, friend, family, or significant other. This will be mainly used to supplement the qualitative reading that will be presented in section (4.3) to contextualize the conversations better. Please be aware that throughout the study, we used the term "risky" to refer to uncomfortable or unsafe conversations.

The data collection efforts above required significant development efforts to create a secure web-based system. On the front-end, we utilized PHP, HTML, and CSS to facilitate the upload of participants' Instagram files in the form of zipped JSON files. On the back-end, we leveraged several Amazon Web Services (AWS), including Lambda functions to process the files, a Relational Database Service (RDS) to store private conversations and risk annotations, and S3 buckets to securely store images and raw data files. All the data transitions were encrypted using RDS at-rest and in-transit encryption, and lambda environment variables encryption (further details about this system can be found in our published case study that will be cited upon acceptance to avoid de-anonymizing ourselves during blind review). Due to institutional data compliance requirements for securing highly sensitive data, we performed all development and data storage on an institutional AWS account that was frequently audited by professionals employed by the university's IT research and security audit teams.

3.4 Participant Recruitment and Demographics

For this study, participants were recruited based on the following selection criteria: 1) between 13 and 21 years old, 2) English speakers, 3) located in the United States, 4) had an active Instagram account for at least 3 months during the time they were a teen (ages 13-17), had direct conversations with at least 15 people, and 5) had at least two conversations made them feel unsafe or uncomfortable. Per the requirements of our Institutional Review Board (IRB), participants who were under the age of 18 were required to provide their parents' consent and their own assent prior to their enrollment to the study. Participants who were over 18 years old completed the adult consent form. During the study, we disclosed our status as mandated child abuse reporters and warned participants that any instances of child pornography would have to be reported to the proper authorities. Therefore, we clearly requested the participants to not upload any content that includes the nudity of minors and described to them the required steps to delete such content from their Instagram files prior to uploading to our study system. Additionally, we obtained a Certificate of Confidentiality from the National Institute of Health, which protects the participants' privacy and prevents the subpoenaing of the data during legal discovery. All personally identifiable information from any textual or image data was removed and all quotations were paraphrased in our results to further protect the youths' privacy. The participants were compensated with a $50 Amazon gift card for their data and time after verifying the quality of their data.

[1]https://www.facebook.com/help/instagram/192435014247952

For this study, ($N = 173$) youth participants were able to successfully complete both parts of the study. Most of the participants were females (67%), (23%) males, and (10%) non-binary. Half of the participants identified themselves as heterosexual or straight 50%, while the rest were bisexual (28%), homosexual (9%), and 13% preferred to self-identify. In order to examine the impact of these demographics on the generated youth profiles, we conducted between-group analysis (χ^2) between the profiles based on sex, age, and sexual orientation. From the 173 participants, we collected ($N = 33,469$) conversations and out of these conversations ($N = 32,256$) were flagged as safe conversations and ($N = 1,213$) as unsafe conversations. Out of these unsafe conversations, ($N = 3,066$) messages were flagged for risk levels and types. The following section presents the results of this study.

3.5 Data Analysis Approach

We combined the analysis of self-reported risk experiences and social media trace data from participants. We applied the Mixture Factor Analysis (MFA) to create the youth profiles based on online and offline risk experiences. Next, between-group analysis (χ^2) was performed to examine any significant differences between the profiles based on their risk flagging. Then, an unsupervised language modeling approach was performed to extract key linguistic differences in the profiles' unsafe private conversations. The following sections describe these approaches in more detail.

3.5.1 Mixture Factor Analysis (MFA) to Profile Participants' Self-Reported Risk Experiences (RQ1). To create the youth risk profiles, we used the self-reported measures of youths' risk experiences that were explained in Section 3.2. We conducted a series of Mixture Factor Analyses (MFAs) with a robust maximum likelihood estimator to group like-minded youth. Mixture Factor Analysis is one of the variations of Factor Mixture Models (FMM) that allows the means and variances to vary between the classes when creating the profiles based on a "mixture" of factor mean scores of the self-reported measures [30, 88]. This approach is useful because it demonstrates the relationship between each factor (risk experience) for different groups of youth. Studying youth risk experiences based on the factors improves the interpretability and generalizability of the findings as we study key risk experience patterns (factors) rather than many discrete risk behaviors (items) [65].

The MFA provides only indicators for the optimal number of profiles rather than explicit information to compare the relative quality of the resulting solutions with differing numbers of profiles. An optimal profile solution might exist with a maximum value for the Shannon entropy [115], a minimum value of Bayesian Information Criterion (BIC), which assesses the profile solution parsimony [70], or when the log-likelihood starts to level off, especially that the higher number of profiles may increase the overall model fit, which might not be significant [65]. These indicators may not agree on the optimal profile solution, which usually leads to also use the substantive grounds beside the fit measure indicators [92]: the optimal cluster solution could be decided based on the interpretability and reasonability of the cluster distributions (e.g., avoid solutions with very small clusters and/or have very large cluster means). For this study, we thus leveraged both this substantive ground and the fit measures (in our case a maximum level of entropy and the log-likelihood levels off) were used to decide on the optimal number of the youth risk profiles.

To this end, table 3 lists the resulted profile solutions of the MFA based on different numbers of profiles. For the youth risk profiles, we did not observe any substantial improvements beyond a 5-cluster solution. The BIC reached a minimum value for the 3 and 5 cluster solutions. For the 5-cluster solution, the entropy level reached its maximum value and the log-likelihood started to taper off after the 5-cluster solution. Thus, for this study, a 5-cluster solution was the optimal solution for youth risk profiles. To confirm the existence of significant differences between the profiles based on our participants' risk experiences, we conducted a series of ANOVA tests [34]

Classes	BIC	Entropy	LL
Risk Profiles			
2	1024.35	0.81	-473.68
3	915.43	0.85	-435.715
4	959.59	0.88	-412.424
5	**929.04**	**0.90**	**-382.716**
6	945.15	0.90	-376.33

Table 3. Risk MFA model fit statistics. The bold values indicate the optimal solution (5 profiles) based on the maximum levels of entropy and log-likelihood.

with the risk experiences: offline risk behaviors, offline self-harm, unwanted online harassment, online harassment perpetration, unwanted sexual experiences, interpersonal sexting, as dependent variables and the generated profiles as the independent variable. We also applied a series of post-hoc tests, to compare individual profiles with one another [80]. The identified differences provide a comprehensive overview of the distinct risk experiences of the youth profiles.

3.5.2 Between Group Analysis based on the Risk Flagging of Youth Profiles (RQ2). To correlate profiles' self-reported risk experiences and actual risk flagging behaviors and answer RQ2, between-group analysis (χ^2) was performed between the youth risk profiles based on the risk levels and risk types of the participants' flagged messages ($N = 3,066$). χ^2 tests of independence are between-group tests which are used for two or more variables with normal distribution [117]. The standardized residuals, which are calculated by "dividing the product of subtracting expected from observed values by the square root of the expected value" [117], were used in this study to show the significant associations between the youth risk profiles and their risk flagging. Through these χ^2 tests, we mapped youth risk profiles' self-reported responses to their actual risk flagging (risk levels and types) of their unsafe private messages.

3.5.3 Linguistic Differences in Youth Risk Profiles' Social Media Conversations (RQ3). To answer RQ3, we used an unsupervised language modeling technique, Sparse Additive Generative Model (SAGE) [42] to examine the linguistic differences in the youth risk profiles' unsafe conversations. SAGE extracts distinguishing keywords in given texts by comparing the parameters of logistically parameterized multinomial models using self-tuned regularization to control the trade-off between frequent and rare terms [42]. We applied SAGE to identify *n*-grams (*n*=1,2,3) that differentiate the unsafe Instagram private conversations flagged by the five youth risk profiles (after merging the conversations for each profile). Specifically, based on a generative model framework, SAGE uses the metric of a log-odds ratio to compare the word distribution between a target and reference corpus [42]. Therefore, we conducted one versus the rest by comparing the private unsafe conversation of one profile with the other profiles' conversations to identify the keywords that have an impact on each profile, similar to other works such as [32]. The SAGE value of an *n*-gram indicates the level of its "uniqueness", meaning that the positive SAGE values (above 0) indicate that the n-gram was more distinctive in a given profile's private conversations than the rest. SAGE results were packed up with content analysis [59] to contextualize the extracted keywords and better understand their context within the profiles' unsafe conversations. Using SAGE and the qualitative reading enabled us to capture the distinctive and salient characteristics of the profiles' unsafe conversations content to compare them, contextualize their risk flags, and holistically understand their self-reported responses to the online/offline risk experiences survey. The percentages provided for the messages based on the risk types in Section 4.3 were calculated based on the total messages flagged by each

profile. In addition, the quotations presented within our results represent exemplars reflective of themes observed in the larger dataset.

4 RESULTS

In this section, we presented the created profiles based on the self-reported risk experiences (RQ1), the types and levels of risks these profiles flagged (RQ2), and the linguistic differences in their unsafe private Instagram conversations (RQ3).

4.1 Youth Self-Reported Risk Profiles (RQ1)

The resulting five MFA clusters represented youth risk profiles that described a set of distinct risk experiences that the members of each cluster encountered online and offline. Figure 1 shows the profiles along with the percentages breakdown of the number of participants who were members of each profile. The profiles' average scores of the risk experience constructs mainly ranged from 1 to 3.5, where the highest reported scores were used to semantically label the profiles. In the web graph, the constructs are shown clockwise in descending order by frequency from experienced the most to the least by the aggregated average scores of all of the groups. Mean scores and standard deviations of the profiles' responses are listed in Table 4. Table 5 and Table 6 show that ANOVA yielded significant differences between the youth profiles based on the risk experience constructs. The following section describes these profiles along with the significant differences in detail.

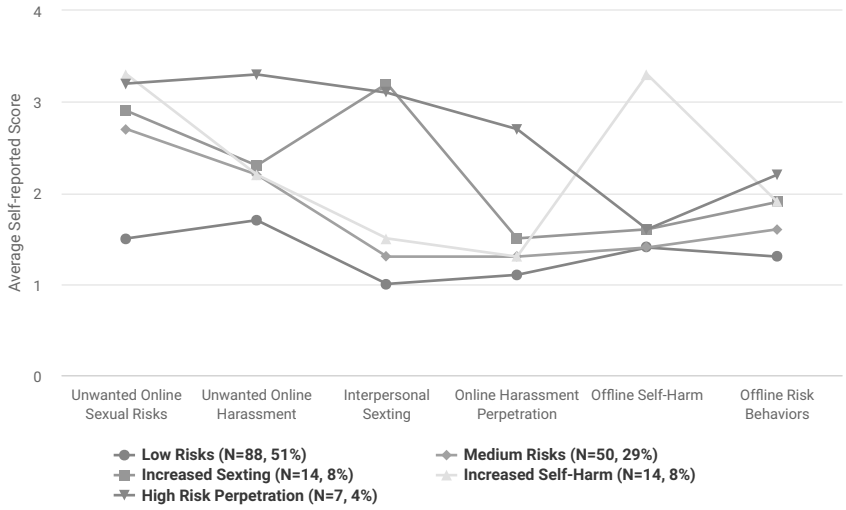

Fig. 1. Risk profiles of (N=173) youth based on the average self-reported scores of their offline and online risk behaviors. This figure compares the distinct risk behavior patterns for the youth profiles related to their online and offline behaviors.

Low Risks (51%): This profile represented the largest group forming (51%) of our participants. Based on the ANOVA results listed in Table 5 and Table 6, participants in the Low Risks profile self-reported significantly lower scores for all the risk experience measures online and offline, compared with other profiles. Overall, Low Risks profile encountered fewer risk experiences online and offline as this profile's responses mostly fell between "Rarely" and "Never" on the Likert scale.

Medium Risks (29%): As the second largest group, youth in the Medium Risks profile reported middle-level scores for the online and offline risk experience measures, compared to other profiles.

	UOSR		UOH		IS		OHP		OffSH		ORB	
Youth Profiles	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Low Risks	1.52	0.43	1.60	0.49	1.06	0.19	1.17	0.24	1.43	0.43	1.38	0.28
Medium Risks	2.77	045	2.33	0.64	1.35	0.40	1.30	0.34	1.48	0.42	1.68	0.36
Increased Sexting	2.94	0.57	2.33	0.54	3.22	0.60	1.54	0.43	1.65	0.48	1.90	0.48
Increased Self-harm	3.35	0.82	2.91	0.80	1.59	0.58	1.39	0.35	3.32	0.75	1.99	0.52
High Risk Perpetration	3.21	0.42	3.21	0.76	3.10	0.68	2.69	0.46	1.67	0.64	2.23	0.77

Table 4. Mean and standard deviation of the Unwanted Online Sexual Risks (UOSR), Unwanted Online Harassment (UOH), Interpersonal Sexting (IS), Online Harassment Perpetration (OHP), Offline Self-harm (OS), and Offline Risk Behaviors (ORB) by the youth risk profiles.

Constructs	df	F	p-value
Unwanted Online Sexual Risks	4	92.11	$p < 0.001$
Unwanted Online Harassment	4	27.58	$p < 0.001$
Interpersonal Sexting	4	128.80	$p < 0.001$
Online Harassment Perpetration	4	50.11	$p < 0.001$
Offline Self-harm	4	48.64	$p < 0.001$
Offline Risk Behaviors	4	16.38	$p < 0.001$

Table 5. ANOVA results. There were significant differences between the youth profiles based on the listed constructs (p-values less than 0.05).

Looking at the ANOVA results Table 5 and Table 6, we found that the Medium Risks profile reported significantly higher average scores than the Low Risks profile for unwanted online sexual risks, unwanted online harassment, interpersonal sexting, and offline risk behaviors. We also found that the Medium Risks profile experienced significantly fewer unwanted online sexual risks than increased self-harm and less interpersonal sexting than the Increased Sexting profile. On the scale, this profile's responses mostly fell between "Sometimes" and "Rarely" for all the measures.

Increased Sexting (8%): Compared to other profiles, youth in the Increased Sexting profile self-reported the highest levels of interpersonal sexting (mostly "Sometimes" or "Often" on the scale). An ANOVA (Table 5) yielded that the youth in this profile experienced significantly more frequent interpersonal sexting than the Low and Medium Risks profile. It was clear that this profile also experienced incidents of unwanted sexual risks and unwanted online harassment, which were reported on the scale as "Sometimes" or "Rarely." This profile rarely experienced online risk perpetration, offline self-harm, and offline risk behaviors.

Increased Self-Harm (8%): An ANOVA (Table 5 and Table 6) showed that the Increased Self-Harm profile self-reported significantly more frequent offline self-Harm experiences than all profiles. An ANOVA also revealed that this profile had significantly more unwanted online sexual risks than Low and Medium Risks profiles. Youth in this profile reported between "Often" and "Sometimes" on the scale for offline self-harm and "Sometimes" for unwanted online sexual risks.

High Risk Perpetration (4%): High Risk Perpetration represented the smallest profile. In comparison to other profiles, High Risk Perpetration profile reported the highest average scores for online harassment perpetration as their responses fall mostly in "Sometimes" on the scale. Looking at the ANOVA results (Table 5 and Table 6), the High Risk Perpetration profile had significantly higher

Constructs	Significant Pairwise Differences (Mean)
Unwanted Online Sexual Risks	Medium Risks (m=2.77), Increased Sexting (m=2.94), Increased Self-harm (m=3.35), and High Risk Perpetration (m=3.21) > Low Risks (m=1.52) Increased Self-harm (m=3.35) > Medium Risks (m=2.77)
Unwanted Online Harassment	Medium Risks (m=2.25), Increased Sexting (m=2.33), Increased Self-harm (m=2.91), and High Risk Perpetration (m=3.21) > Low Risks (1.60)
Interpersonal Sexting	Increased Sexting (m=3.22), Medium Risks (m=1.35), Increased Self-harm (m=1.59), and High Risk Perpetration (m=3.1) > Low Risks (m=1.06) Increased Sexting (m=3.22) > Medium Risks (m=1.35)
Online Harassment Perpetration	High Risk Perpetration (m=2.69) > Low Risks (m=1.17), Medium Risks (m=1.30), Increased Sexting (m=1.54), and Increased Self-Harm (m=1.39)
Offline Self-harm	Increased Self-Harm profile (m=3.32) > Low Risks (m=1.43), Medium Risks (m=1.48), Increased Sexting (m=1.65), and High Risk Perpetration (m=1.67)
Offline Risk Behaviors	Medium Risks (m=1.68), Increased Sexting (m=1.90), Increased Self-Harm (m=1.99), and High Risk Perpetration (m=2.23) > Low Risks (m=1.38)

Table 6. Summary of significant pairwise differences.

online harassment perpetration experiences than all profiles. Overall, this profile experienced significantly more online and offline risk experiences than Low Risks except that this profile reported significantly less levels of offline self-harm experiences than the Increased Self-Harm profile.

Demographics	Parameter	Low Risks	Medium Risks	Increased Sexting	Increased Self-Harm	High Risk Perpetration	χ^2
Sex	Female	32%	23%	5%	5%	2%	p-value = 0.24
	Male	14%	5%	2%	1%	2%	
	Non-Binary	4%	2%	1%	2%	0%	
	Prefer to self-identify	2%	0%	0%	0%	0%	
Age	13-15	15%	7%	1%	1%	1%	p-value = 0.12
	16-18	23%	13%	5%	6%	2%	
	19-21	14%	9%	2%	0%	0%	
Sexual Orientation	Heterosexual or straight	29%	14%	4%	1%	2%	p-value = 0.08
	LGBTQ+	22%	15%	5%	6%	2%	

Table 7. Distribution of demographics by the risk profiles and χ^2 results. % Out of the total number of participants (N=173).

Overall, the youth profiles highlight the multidimensionality of youth risk experiences. When analyzing these profiles based on the reported demographics, we found that the profiles were not impacted by the youths' demographics (there were no significant differences yielded between the youth profiles based on sex, age, and sexual orientation using χ^2 as listed in table 7). This result is noteworthy as it suggests that profiling youth based on their risk experiences, rather than their demographic characteristics, yields additional insight that may be missed if we focused on demographic information alone.

4.2 Youth Risk Profiles Significantly Differed Based on the Flagged Risk Levels and Types (RQ2)

Next, we examined whether the self-reported risk experiences had any relationship with the youths' risk-flagged social media data. The χ^2 test uncovered key differences between the youth profiles and their flagged risk messages based on risk severity levels and types, which will be presented in the following sections.

4.2.1 Youths' Flagged 'Risk Levels' Aligned with their Self-Reports. The χ^2 test indicated a significant association between the youth profiles and their flagged risk levels ($\chi^2(df = 8, N = 3,066) = 94.38, p < 0.001$). As illustrated in Figure 2, a strong positive association was found between the Low Risk profile and the number of conversations flagged as low risk levels in the social media data. Further, a significant negative association was found with medium and high risk levels. This indicated that youth in this profile were most likely to flag their unsafe messages with a low risk level, which clearly aligned with their low average scores for the self-reported risk experiences. For the Medium Risks profile, by looking at the standardized residuals in Figure 2, we found a significant positive association between the Medium Risks profile and medium risk level and a significant negative association with high risk level. This suggested that the unsafe messages of this profile were most likely flagged as medium risk. This also showed an alignment between their medium scores for the self-reported risk experiences and their risk level flagging.

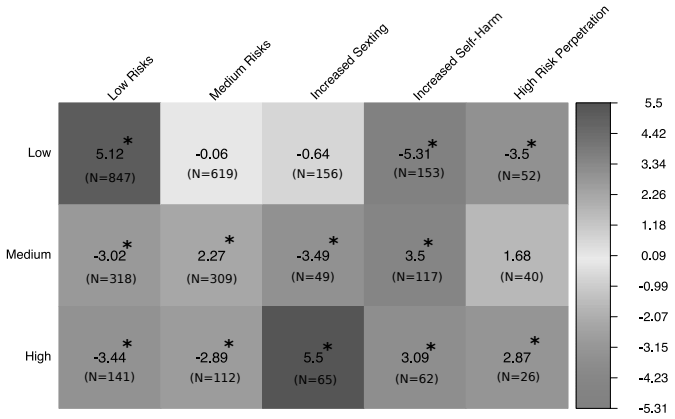

Fig. 2. Results (standardized residuals) of the between-group analysis for risk levels of the risk profiles ($N = 3,066$). (*) indicates significant association. Note that green denotes a positive association, while red denotes a negative one.

For the rest of the profiles (Increased Self-Harm, Increased Sexting, and High Risk Perpetration), the standardized residuals showed significant positive associations between these profiles and the high risk level as demonstrated in Figure 2. This finding suggested that the unsafe messages flagged by these profiles were more likely to be a high risk level. Statistically significant negative associations were found between the Increased Sexting profile and medium risk level, suggesting that adolescents in this profile were less likely to flag their unsafe messages as medium level. On the contrary, the Increased Self-Harm profile showed a significant positive association with medium risk level, which indicated that the unsafe messages of this profile were more likely to be high and/or medium risk levels. Significant negative associations were found between the Increased Self-Harm and High Risk Perpetration profiles and low risk level, which suggested that these profiles were less likely to have unsafe messages with low risk levels. These findings show a clear

alignment between these profiles and their self-reports as each profile reported the highest average scores for certain risk experiences (displayed in Figure 1).

4.2.2 Youth's Flagged 'Risk Types' Mostly Aligned with their Self-Reports. The youth risk profiles were significantly different based on their flagged risk types using χ^2 test ($\chi^2(df = 20, N = 3,066) = 167.43, p < 0.001$). Looking into the standardized residuals in Figure 3, a strong positive association was found between the Low Risks profile and risk types including digital self-injury and spam/others, along with a significant negative association with sexual messages/solicitation/nudity risk. This suggested that when adolescents in the Low Risks profile flagged their messages, they most likely flagged them as digital self-injury or spam/others and were less likely to flag for sexual messages/ solicitation/nudity. Generally, it was not surprising to see this profile mostly flagged spam/others, which matched with their self-reports and low risk level flag; however, finding the digital self-injury as part of their flagging warranted further qualitative unpacking, which will be done in section 4.3. For the Medium Risks profile, a significant positive association was found between this profile and harassment and a significant negative association with hate speech/threat of violence as shown in Figure 3. This finding suggested that the risk messages of this profile were more likely to be flagged as harassment and less likely to be flagged as hate speech/threat of violence. Medium Risk profile self-reported medium levels scores for unwanted online harassment, which suggested a fair alignment between their self-reports and actual risk flagging.

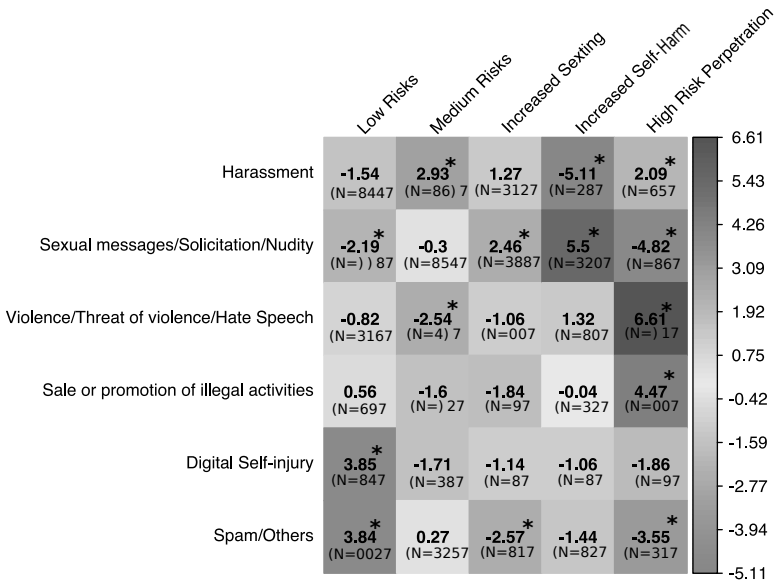

Fig. 3. Results (standardized residuals) of the between-group analysis for risk types of the youth risk profiles ($N = 3,066$). (*) indicates significant association. Note that green denotes a positive association, while red denotes a negative one.

For the Increased Self-Harm and Sexting profiles, the χ^2 test yielded strong positive associations between these profiles and the sexual messages/solicitations/nudity risk type as shown in Figure 3. This result indicated that the risk messages flagged by these profiles were mostly flagged as sexual messages/solicitations/nudity. For the Increased Sexting profile, a significant negative association was found between this profile and spam/others risk type, which suggested that this profile was less likely to flag their messages for spam/others. While for the Increased Self-Harm profile, there was a negative significant association between this profile and harassment risk type, which suggested that

this profile was less likely to flag their messages for harassment. These findings aligned with their self-reports since both of these profiles self-reported higher levels of unwanted online sexual risks. For the High Risk Perpetration profile, Figure 3 shows a significant positive association between this profile and risk types including harassment, hate speech/threat of violence, and sales or promotion of illegal activities, and a significant negative association with sexual messages/solicitations/nudity and spam/others risk types. This indicated that youth in this profile mostly flagged their messages for harassment, hate speech/threat of violence, and sales or promotion or illegal activities and were less likely to flag them for sexual messages/solicitations/nudity and spam/others risk. Overall, this profile's risk flagging suggested an alignment with their self-reported risk experiences for online harassment perpetration and unwanted online harassment. Understanding the role of the participants in these risks motivated us to uncover the nuances of their direct unsafe conversations, which will be presented in the following section.

4.3 The Unsafe Conversations of the Youth Risk Profiles Differed Linguistically (RQ3)

To qualitatively examine the differences that we uncovered statistically in the profiles above, we conducted additional linguistic analyses. Table 8 lists the SAGE results as top keywords that were salient in the profiles' unsafe conversations. The following subsections unpack these conversations for each profile.

Low Risks		**Medium Risks**		**Increased Sexting**		**Increased Self-Harm**		**High Risk Perpetration**	
***n*-gram**	**SAGE**	***n*-gram**	**SAGE**	***n*-gram**	**SAGE**	***n*-gram**	**SAGE**	***n*-gram**	**SAGE**
dysphoria	3.86	smelly america	2.60	u wanna come	3.68	cum	3.68	murdered	4.47
brand ambassadors	3.36	respect	2.55	boob pic	2.46	submissive	3.41	die	4.47
kills myself	2.64	hypocritical	2.55	invite you	2.47	nice dick	3.41	burn in hell	4.25
cut it off	2.47	gay	2.51	naked	2.47	mylol	3.41	kick your ass	4.25
blood	2.39	all lives matter	2.49	attached	2.47	eat u out	3.20	stay away	4.25
get rid of	2.32	beautiful	2.48	sleep with you	2.44	jerking off	3.19	weed	4.25
giftcard	2.50	even hot	2.47	airbnb	2.41	vagina	3.19	do not come	4.25
trust	1.81	hoe	2.46	spend the night	2.38	rubbing your pussy	3.19	leave me alone	3.98
your pictures	1.77	are single	2.45	cant be today	2.38	princess	3.13	bitcoin	3.79
hurry	1.76	absolutely gorgeous	2.45	chill	2.38	thigh pics	3.10	where you stay	3.59
someone posted	1.68	thanks friend	2.45	meet you	2.35	i am horny	3.07	vape	3.59
internship	1.60	dumbass	2.44	hang out	2.35	fishnets	3.00	shut	3.51
weekly allowance	1.59	alone	2.43	online school	2.33	lubricant	3.00	little bitch	3.48
wall	1.57	nig	2.42	youre free	2.33	throat	2.98	spyderco	3.32
product reviewer	1.57	ugly	2.42	work	2.29	me so wet	2.97	fortnite	3.15
list name	1.56	stupid	2.41	call or play a game	2.29	leggings	2.97	report you	3.11
sugar	1.56	add snapchat	2.40	today	2.28	your room	2.97	serious danger	3.09
http	1.55	weak ass bitch	2.37	my boobs	2.27	see your pussy	2.96	uncomfortable	3.05
believe	1.55	sexy	2.35	lyk	2.26	ur breast	2.94	kick	2.72
instafans	1.55	cute	2.34	u want	2.26	hurry up bitch	2.94	hurting	2.69

Table 8. Top 20 salient *n*-grams (*n*=1,2,3) in the flagged unsafe conversations across the youth risk profiles (SAGE [42]). The *n*-grams with higher SAGE scores are more distinguishing of the profile when compared with the rest of the profiles.

4.3.1 Low Risks Profiles: Ignored Spam messages, but Engaged in Scam Conversations and Self-Harm Disclosures of Others. Most of the unsafe conversations of youth that belonged to the Low Risk profile were flagged as low risk, which predominantly flagged by youth as sexual messages, solicitation and nudity (N = 443, 48%), harassment (N = 377, 40%), while this profile overall has most of the messages proportionally across all profiles for the spam and scam (N = 225, 24%). Based on SAGE results, this profile's unsafe conversations were distinguished by a variety of spam and scam messages (*brand ambassadors, giftcard, your pictures, someone posted, sugar, weekly allowance, giftcard, instafans etc*), and mental health indicators (*dysphoria, kills myself, cut it off, blood*). The SAGE scores for the self-harm keywords were higher than the spam and scam content, suggesting that other profiles did not have as much self-harm content as the Low Risks profile.

Regarding spam, unsafe conversations contained posted pictures on other accounts or walls (*your pictures*), offers for increasing the number of followers or likes (*instafans*), messages offering gift cards or winning prizes (*giftcard*), or advertisements from business Instagram accounts for their products, services, or their websites. Spam messages were mostly easy for youth to recognize and ignore, but they flagged these messages as risky, as these spam messages made them feel uncomfortable. Further reading at the conversations that contained sexual and harassment messages, we found that many of these conversations could be recategorized as spam or scam messages, such as the following quotes where P26 (a 15-years-old male) participant flagged the first one as harassment and the second quote was flagged by P2 (an 18-years-old female) as sexual messages.

> **Other user:** *"Completely free! Up to 30'000 followers and 5000 likes to all your posts! For details go to our main page"*

> **Other user:** *"Hello sweety!* 🖤🔥 *http://[site link]"*

The most noticeable type of the scam conversations youth in this profile received was "sugar daddy" requests, where the youth received offers from individuals with promises for high weekly or monthly allowances. During the conversations, although the other individuals usually showed proof of payment, participants refused the offers once they were asked for critical personal information such as credit card or Instagram account credentials. For example, P90 (a 14-year-old female) engaged with someone who promised to send money in exchange for access to her Instagram account, but she soon realized the risk posed by this request and refused.

> **Other user:** *"I'm single 40 years old I have a kid I don't have a wife .. Am looking for an honest baby that will keep my company ... And am ready to spoil her with my money.. can you do this?"*
> **P90:** *"of course! could you just send money via cash app? i mean gift card"*
> **Other user:** *"Yes baby, but before I send it should I have access on your instagram account for 2 days to gain your trust baby"*
> **P90:** *"No don't i trust u"*
> **Other user:** *"You don't. If you can let me have access I just wanna buy you gift card"*
> **P90:** *"Actually I changed my mind I'm not desperate enough for money to give out my info, Sorry"*

As indicated in the example above, the youth appeared to not feel threatened by such advances (i.e., considering them low risk), even trying to take advantage of them, until they assessed the possible risk involved with disclosing their personal information.

Youth in the Low Risk profile also flagged exposure to self-harm as low risk, which was initially surprising to us. After inspecting the unsafe conversations flagged, we found that others (mostly strangers) sought support from the youth by sending private messages, disclosing about their self-harm and/or mental health experiences in their daily lives. The quote below illustrated the self-harm disclosures that the youth in the Low Risk flagged, which was about a stranger sharing their thoughts of self-harm with P158 (19-year-old male) seeking support, which was flagged as low risk self-harm.

> **Another user:** *" i have been having thoughts of wanting to cut i just want to scream and cry for hours"*
> **P158:** *"i'm here if it'd help to talk about things further"*
> **P158:** *"I'm sorry to hear it, self harm urges are awful."*
> **Another user:** *"i can't even do the dishes without picking up a knife to wash it and having the feeling to want to cut right then and there and every night i have been falling*

asleep crying"
P158: *"ugh i'm sorry"*

In these cases, youth mostly provided emotional support to others by responding with sympathy. Participants encouraged others to confide in them. Yet, even though these self-harm disclosures were flagged as low risk because participants themselves were not directly in harm, the fact that they were flagged by youth participants indicates that these conversations still made them (or someone else) feel uncomfortable or unsafe.

4.3.2 Medium Risk Profiles: "Cool" Responses to Flirtations and Blowing Off Harassing Comments. Compared with other profiles, youth who fell within the Medium Risks profile flagged conversations that had a set of distinguished keywords related to sexual flirtations (*beautiful, even hot, absolutely gorgeous, sexy, cute*), personal questions or requests (*are u single, add snapchat*), and harassing comments (*ugly, smelly america, weak ass bitch, stupid, gay*). Participants in the Medium Risks profile flagged most of their messages for sexual messages (N = 387, 37%) and harassment (N = 364, 35%), which were mostly found in the form of sexual flirtations they received in Instagram direct messages as harassment, but they usually either ignored these advances or responded positively. For example, when the flirtations came from random strangers (as indicated by participants), like in the following quote from P77, youth rarely responded.

> *"Hi you do not know me but I just wanted to stop by your page to tell you that you're very beautiful ", flagged by P77 15 years old female*

However, when the flirtations came within conversations with non-strangers (i.e., acquaintance, friend, family, or significant other), youth often accepted the compliment with responses, such as *"that's true i'm beautiful, i agree, thank you!"*, emojis that showed they liked these flirtations, or encouragement to send more personal pictures (though not nude). Interestingly, participants still flagged these conversations as harassment, even though their responses appeared as if they welcomed the advances.

Youth in the Medium Risk profile mainly left the harassing comments they received without response, either in one-to-one or in group conversations, yet they flagged these conversations as harassment. The youth who identified themselves as LGBTQ+ in this profile often considered the threats to inform their family or friends about their sexual identity as harassment, violating their rights to practice their sexuality in private. While in other cases heterosexual youth in this profile were often name-called as "gay" regardless of their sexual identity, which made them feel uncomfortable or unsafe. For instance, if youth shared their sexual identity publicly, they were harassed privately as exemplified in the following conversation between P51 (14-year-old, non-binary youth) and another individual.

Other user: *"If u have a set of balls you a dude .. Is it that hard to understand?"*

In other cases, youth were not the target of harassment within group conversations, but they often flagged these conversations as medium risk, reflecting that they felt uncomfortable being involved in a group chat that mainly involved harassment.

4.3.3 Increased Sexting Profiles: Flagged Offline Meeting Requests within their Sexual Conversations. Although Increased Sexting and Self-Harm profiles mostly flagged conversations for sexual messages, solicitation, and/or nudity, (N = 133, 44%) and (N = 152, 52%), respectively (displayed in Figure 3), SAGE resulted in keywords from the profiles' unsafe conversations that indicated that youth in the Increased Sexting flagged in-person meeting requests while the youth in the Increased Self-Harm flagged sexual content. Youth within the Increased Sexting profile mostly flagged their

conversations for in-person meeting requests resulting in SAGE keywords related to meetings in real life (*u wanna come, invite you, sleep with you, bnb (i.e., Airbnb)*).

While the Increased Sexting profile engaged in sexual conversations, we observed a level of hesitation when the other individuals were willing to transition the relationship into physical meetings sometimes for sex as stated. For instance, in the following conversation, P121 (17-year-old, female) expressed hesitation towards an in-person meeting request from someone within their sexual conversation "in the least sexual way."

Other user: *"My house is free if u wanna come this weekend"*
Other user: *"I want to sleep with you (I mean that in the least sexual way possible)"*
P121: *"i would but i'm always working on the weekends, but I'll keep it in mind"*

Hesitation from other participants also came in the form of not explicitly refusing these requests; instead, when the other person usually insisted to meet, they sent various excuses to refuse meeting the other person such as being sick, having work, school, or having their family over as well as promising the other person of future meetings.

In addition to the in-person meetings, youth in the Increased Sexting profile flagged other sexual conversations for the requests they received to send their nudes or have sexual video calls. The youth in this profile seemed to only want to be the recipient of these sexual conversations. For instance, while they willingly engaged in such conversations and explicitly agreed to receive nudes from others, many refused to send them and flagged the requests as unsafe, as in the following conversation from P3 (16-year-old, female).

Other user: *"is it ok if i can send nudes"*
PP3: *"yeah"*
Other user: *"To you?"*
P3: *"sure"*
Other user: *"ur boob pic?"*
P3: *"no"*

4.3.4 Increased Self-Harm Profiles: No Self-Harm Content was Found but Engaged in High-Risk Sexual Conversations with Strangers. Compared with other profiles, Increased Self-Harm, even in comparison with the Increased Sexting profile, had the most sexual SAGE keywords that indicated sexual interactions within their unsafe conversations (*cum, nice dick, jerking off, rubbing my pussy, i am horny, lubricant, me so wet*). Meanwhile, no self-harm content was found in the keywords; however, when inspecting their conversations more closely, we did find indications of mental health struggles, as they flagged the support they received from others as harassment.

Youth in the Increased Self-Harm profile often engaged in sexual exchanges with others who they indicated were strangers. Similar to the below conversation that P82 (17-year-old, female) has, where she explicitly exchanged nudes with another male.

Other user: *"I wanna see your pussi"*
Other user: *"Do you wanna see mine??"*
P82: *" sure yes"*
Other user: *"And can I see yours?"*
P82: *"yes"*
P82: [user sent attachment] *"looks nice"*

Although youth in this profile stated their enjoyment of sexually pleasing others as in the following example from P94 (16-year-old, non-binary youth) who answered that they enjoyed the experience, they regretted engaging in these conversations based on their reflections (e.g.,"*This conversation was grooming. I regret it.*" from P94), mostly because they discovered who tried to

sexually solicit them. In fact, the Increased Self-Harm profile was the only profile that did not only flag the received messages but also flagged their own messages as unsafe.

> **P94:** *You make me so wet.. yes, daddy Fuck me* [Flagged as high-risk sexual messages/solicitation]
> **Other user:** *"oh girl, u are a good submissive. I fucking love u"*
> **P94:** *"I really want to make you happy daddy"*
> **Other user:** *"awww, daddy is happy"*
> **Other user:** *"Tell me are u enjoying this ? Don't lie"*
> **P94:** *"I am enjoying this daddy"*

Regarding the self-harm content, SAGE did not result in any keyword related to sharing any self-harm content within their unsafe conversations. Instead, after we inspected their unsafe conversations, we noticed that the youth in this profile flagged the support they received from people (mostly known) as unsafe. For example, P67 (15-year-old, female) flagged her friend's (based on her reflection on the conversation) advice as unsafe.

> *"I know Ur Mental Is not as Strong as Mine So…I've told u before Not to trust anyone from MyLOL and Not to trust ANYONE like Literally ANYONE Except Me. "*

This profile seemed to be annoyed by others' protective support in the form of providing unsolicited advice and their stated hesitation on this profile's ability to make correct choices. This profile ignored these messages along with the constant requests of taking care of them and flagged them as unsafe.

4.3.5 High Risk Perpetration Profiles: Flagged Threats and Illegal Products. The High Risk Perpetration profile majorly flagged their conversations for risks including harassment ($N = 68$, 39%), violence and threats of violence ($N = 40$, 23%), and sales or promotion of illegal activities ($N = 22$, 13%). This profile had the highest SAGE scores for the keywords yielded from their unsafe conversations, indicating a completely different risk content from other profiles. This profile's unsafe conversations had serious indications of physical threats and harm (*murdered, die, burn in hill, kik your ass, stay away, etc*) and illegal products (*weed, vape, spyderco*), which were keywords that were not observed in other profiles' conversations. There were serious indications of threats and bodily harm within the unsafe conversations of the youth in the High Risk Perpetration profile. Most of the conversations appeared to occur after an in-person (mostly at schools or neighborhood) conflict already happened or started online in community-based game platforms, like Fortnite or communication servers like Discord as stated in the context of the messages then escalated in Instagram direct conversations, like in the following threats P117 (17-year-old, female) received because of what she did at school.

> **Other user:** *"You was talking all that shit at school ho"*
> **Other user:** *"When you come back I'm going to kick your ass"*
> **Other user:** *[participant name] "ass bitch"*
> **Other user:** *"Stay away from my nigga"*
> **P117:** *"Your white self saying the n word I'm gone report you to the principal Monday"*
> **Another user:** *"Bitch shut you ho ass up. You a weak ass bitch"*

The threats ranged from hacking social media accounts to more serious aggressive physical threats of killing or beating. Social exclusion was also observed among the threats received as others threatened this profile of removing them from an online gaming community or staying away from them or their friends in school. In most cases, the youth in this profile tried mostly to avoid further escalations by explicitly mentioning that they will report the threats to either authorities or their friends.

The youth in the High Risk Perpetration profile flagged another set of messages for illegal products. We found that the youth in this profile engaged with Instagram accounts that sell vapes, weed, knives, or products used for marijuana and sent promotions like *"I have good stuff for sale, I got weed, pills, research chemicals, carts, vapes, wax, LSD and more."* Unfortunately, the youth in this profile often did not only buy these products easily from such accounts, they also voluntarily reviewed them with others within their private conversations, especially in group conversations, such as in the following conversation that represented how illegal products were discussed between youth privately, from P134 (18-year-old, male) who discussed substance misuse.

Other user: *"Big blunt today?"*
P134: *"smoked a joint.. Im good"*

Overall, we demonstrated how each youth risk profile had distinct characteristics in private conversations that were flagged risky based on their perspective. These findings will be further discussed in the following section.

5 DISCUSSION

This study presents an approach for better understanding the multidimensional nature of youth offline and online risks, as well as demonstrating how their self-reported risk profiles' align with their social media data. We first identified five profiles of youth risk behavior (RQ1): 1) Low Risks (51% of the participants), 2) Medium Risks (29%), 3) Increased Sexting (8%), 4) Increased Self-Harm (8%), and 5) High Risk Perpetration (4%). These self-reported risk profiles were fairly well-aligned with youths' risk-flagged social media trace data (RQ2). For RQ3, we uncovered key linguistic differences in unsafe conversations across the profiles that described each profile's unique risk experiences. The sections below unpack the implications of these results in comparison to other youth safety literature and provide recommendations towards designing targeted interventions to protect at-risk youth online.

5.1 Unpacking the Nuanced Risk Behaviors, Experiences, and Perceptions of Youth

Overall, we found that youth in our study experienced a wide array of risks within their Instagram private conversations. Contradictory to Pabiana's et al. [93] study that suggested that most of youth's offline conflicts or threats remain offline, we found that the youth who belonged to the High Risk Perpetration profile received and flagged threats that stemmed from escalated conflicts as violence/threat of violence/ hate speech. Through unpacking their unsafe conversations on Instagram, we uncovered the possible reason behind preserving violence/threat of violence/ hate speech as a higher risk than general harassment, which may rely on the fact that these threats indicated purposeful physical harm to the youth. This implication presents an important point for future research to consider separating these risks, which is in line with prior research that highlights the importance of not confiding harassment with hate speech and violence [50]. Violence or threats are more targeted to make the person fearful, hate speech is an extreme bias expressed, while harassment is more of emotional torment [123]. In addition, while prior research on youths' violence in the offline context, such as school or within family violence, is well established [129, 141], online violence has been underinvestigated [15]. In fact, online violence has been found to lack standard definitions and methods [15]. Therefore, our findings motivate future research to investigate the youths' experiences of online violence, especially since these experiences could present imminent risks to their lives. This could be accomplished in several study designs, including conducting in-depth interviews with youth or professionals who work closely with them to provide deeper and broader insights regarding online violence, leveraging digital trace data (either public or private)

of youth to deeply understand the dynamics of this phenomena using a more naturalistic or in-situ form of observation (e.g., diary studies), or examining the impacts of these online violence experiences on youths' lives through longitudinal investigations.

In this study, an interesting yet unanticipated finding was that youth who self-reported the highest scores of offline self-harm did not exhibit evidence of digital self-harm. This finding contradicts previous research that found youth who self-harmed often shared self-harm content on social media [25]. A possible explanation is that the youth who self-harmed may not manifest these self-harm behaviors within private online spaces. Another explanation is that non-suicidal self-harming behaviors of youth may be a cry for help or a way to get attention [43]; therefore, posting about these self-harm experiences publicly may achieve this goal over sharing in private conversations. Instead, their social media data revealed that they encountered high risks of online sexual interactions. This implication points again to the importance of calibrating self-reported responses with social media data to uncover such nuance. Thus, instead of only relying on social media content to identify youth at risk for self-harm, which is the case of most of the current literature [45, 82], future research is warranted to consider high-risk sexual behaviors and/or self-reported mental health concerns as a proxy for identifying self-harm risks. Using these as proxies would help scholars, and possibly even professional clinicians, to more precisely identify the youth who are at risk of self-harm; and therefore, intervene before physical harm occurs.

Importantly, our findings emphasize how youth often flagged conversations as unsafe, even when they appeared to engage and enjoy these interactions. For instance, youth in the Medium Risks profile engaged with flirty comments, and youth in the Increased Sexting and Self-Harm profiles engaged in sexual conversations, but then all these profiles flagged these conversations as unsafe. In addition, youth in the Low Risks profile emotionally supported others who disclosed their self-harm or mental illness experiences, which could reduce the urges of self-harm to others [135], but, could also trigger the imitative self-harm behaviors of the supporters [12, 20]. These findings suggest that the youth in these profiles may have a hard time setting healthy boundaries and coping strategies in such situations. Therefore, a line of research should be established to identify the ethical considerations of youths' support provisions for others by investigating the potential harms and benefits of disclosure and confidentiality within their online private supportive environments. More importantly, our finding about youth flagging their unsafe engagement online could be tied back to resilience-based literature that found youth to benefit from their low to medium risk behaviors by learning how to cope before they encounter higher-level risk behaviors [61, 131, 134]. Therefore, future research should focus more on identifying the individual and contextual factors that would foster youth resilience and coping skills when they experience low to medium-risk behaviors while providing risk interventions only for those who are at high risk.

5.2 Profiling Youth Risk Behaviors to Develop Targeted Risk Interventions for At-Risk Youth

Profiling youths' risk experiences (RQ1) uncovered a nuanced picture of the risks youth encounter both offline and online. Importantly, we found only around 20% of youth experienced the most concerning risks (i.e., Increased Sexting, Self-Harm, and High Risks Perpetration) that intertwined across online and offline contexts and levels of involvement (i.e., victim vs. perpetrator). This finding is noteworthy as it shows that while most of our youth participants were active social media users, the majority did not report high-risk experiences across offline and online contexts. Meanwhile, the literature and news media regarding online risks has framed these negative experiences as an "epidemic" among youth [95], adopting a "moral panic" stance that associates youths' social media usage with criminality and mental health crises [126]. In contrast, our youth profiles showed that this might not be the case for the majority of youth. Our study recruited youth specifically who had

previously encountered at least two online risk experiences in their private message conversations, which coincidentally resulted in a larger than the census-level proportion of LGBTQ+ youth in our sample. Still, the risk profiles that emerged uncovered a silverlining – about 80% of the youth in our study encountered only low to medium levels of risk both offline and online. Therefore, we caution future researchers to not over-problematize the risk behaviors of youth in a way that promotes deficit-based narratives that are counterproductive.

Instead, our work helps researchers and practitioners hone in on the approximately 20% of youth (reminiscence of the Pareto principle) who encountered the most problematic risk behaviors and experiences offline and online so that we can design effective interventions targeted to protect them from harm. We argue that future risk intervention efforts should focus more time and resources on these youth who are more vulnerable because they are at disproportionate risk; thus, have higher and more imminent need. "Triage," or the ability to prioritize and administer aid to those who need it most [87], is a well-established concept within the medical community and has gained traction as a formidable goal of AI-based medical systems [111]. However, this concept is new to the automated risk detection and youth online safety literature. This may be because, as a society, we are altruistic and want to keep *all* youth out of harms way. Yet, once high-risk sub-populations of youth have been identified, risk prevention programs can administer targeted prevention strategies that cater to the specific needs of these youth in a way that would be much more effective than broad-spectrum interventions.

For example, we found that youth in the increased self-harm profile reported the highest scores of online sexual risks and offline self-harm, while the youth in the High Risk Perpetration profile reported the highest scores for unwanted online harassment and harassment perpetration (refer to Figure 1). In contrast, existing cyberbullying interventions that often either target victims (e.g., the U.S. government stopbullying initiative [1]) or perpetrators (e.g., the Centers for Disease Control and Intervention initiative to Reduce Youth Violence [35]) may fail to adequately protect youth who are victims and perpetrators of online harassment. In our results, we observed that sometimes youth experienced the dual experiences of online harassment, which is important to be noted as they could be at a higher risk of reacting to harassment with aggression, depression, and somatic symptoms [52]. The stressful experiences of unwanted online harassment may trigger youth to be perpetrators and have stronger emotions to harm others as a coping mechanism, especially with the anonymity afforded by online spaces [6, 40, 127]. Therefore, we recommend future research to design evidence-based interventions that take into consideration both behaviors of youth who are victims of online harassment and may harass others to appropriately help them and mitigate harm.

5.3 Acknowledging the Complex Nature of Youth Risk Behaviors and Experiences in Future Research

A key contribution from this research is that it confirms that youth risk behaviors and experiences spanning both offline and online contexts should not be examined within a vacuum, as they are multi-faceted and multi-dimensional, often correlating with one another in unsuspected ways. Our results for RQ2 also indicated that the self-reports from youth showed ecological validity and strong correlations with their social media trace data. At the same time, our research demonstrated the value of triangulating youth self-report data with their social media data to uncover key nuances in their lived online risk experiences that would not have been found if we had only focused on the quantitative analysis of survey data. For example, it was unexpected to find youth self-reported self-harm as low risk, which only made sense after we examined their social media data to find that they were acting as supporters of others who brought up self-harming. Disentangling these

[1] https://www.stopbullying.gov/

incidents was important for better understanding why and how youth characterized risks. Therefore, we urge future research to also leverage self-reports of youth backed by their digital trace data to provide more accurate and nuanced conclusions.

While we emphasize the importance of leveraging the private digital trace data of youth, we also acknowledge that obtaining such a dataset is difficult and comes with its unique ethical and legal challenges [103]. Therefore, collecting, analyzing, and sharing such data is a complicated and effortful endeavor, requiring large collaborative efforts. To facilitate overcoming these challenges, researchers within the HCI community have launched an initiative called MOSafely[2] (i.e., Modus Operandi Safely), which is a task force aimed at building an open source and multidisciplinary community of researchers, clinicians, industry professionals, and civil society dedicated to the online protection of youth [22]. Such collaborative initiatives may provide a way forward for sharing resources, such as datasets and algorithms, to address the online safety of at-risk youth. Thus, future research is recommended to engage with such efforts and benefit from the provided resources, keeping in mind to approach such endeavors with an ethical framework and a focus on protecting the rights and well-being of vulnerable youth populations.

For instance, a challenge that such a community might be able to tackle is the complex nature of youth risk behaviors and how youth risk profiles are likely to change over time. Indeed, the resulting youth profiles in our analyses were static and mutually exclusive, due to the statistical measures that classified youth based on a snapshot of their self-reported data. Yet, factors such as brain and emotional development [11] or family dynamics [118] would have an impact on the lived experiences of youth, which would result in a new classification of the youth profile in the future. For example, youth in the Low Risk profile, who were exposed to others' mental health disclosure, might change their profile to the Increased Self-Harm profile in the future if they do not set healthy boundaries and optimize their support provisions to protect themselves. Therefore, deeply understanding the changes in youths' risk profiles is important to reflect on the required risk interventions. Future research is recommended to conduct longitudinal studies to examine the stability and changes in youth risk profiles over time, which was done in other fields such as investigating the changes in teachers' profiles of success [66] or burnout profiles among workers [19]. In addition, a benefit from the empirically validated relationships between our youth profiles' self-reported data and their digital trace data and given the dynamic nature of social media posted data over time, leveraging the social media data could be a potential way to build a more dynamic system that recreates youth risk profiles in real-time. By doing so, these dynamic profiles could reflect the risk experiences of youth in a timely manner, which could be leveraged as an early detection system for protecting youth from harm both online and offline in an effective way.

5.4 Recommendations for Designing Targeted Risk Interventions for Youth

Our results highlight the importance of moving beyond a unidimensional view of online risks, towards considering the multidimensionality and interplay between different risk types and settings (online vs. offline). Targeted risk detection and prevention strategies need to rely on this multidimensionality of risks for more evidence-based, impactful, and better-equipped strategies to address youths' needs. Therefore, we provide the following recommendations for designing targeted risk detection systems and interventions that are tailored to the unique risk profiles of different youth.

5.4.1 Build Context-Aware and Youth-Centered Online Risk Detection Systems. A key prerequisite of early and personalized assistance based on youth risk profiles is automatic and accurate risk detection [5] In fact, recent works by CSCW scholars on automatic detection of youths' online risks,

[2]https://www.mosafely.org/

such as online sexual risks, have emphasized the importance of considering the human-centered context of the risk experiences when building risk detection algorithms, such as relationship type and age [8–10, 94, 101]. Therefore, we recommend that future machine learning risk detection models could leverage our youth risk profiles' based on their self-reported data along with their digital trace data. Relying on these both sources of data could not be only a reliable way to comprehensively capture the context of youth risk experiences, but also would help the classifiers to target youth who are experiencing multiple high risks or in most need of intervention. A practical example to utilize youth risk profiles in building risk detection classifiers could be by incorporating the profiles' self-reported scores of offline and online risks, salient SAGE keywords, and the online content as features, which would result in a more reliable model to detect and mitigate risks. For instance, a classifier for the Increased Self-Harm profile would jointly try to detect indications of sexual and self-harm risks while a classifier for the Low Risks profile would monitor the patterns of self-harm disclosures of others and detect any indications of developing self-harm or mental health issues of the youth in this profile.

5.4.2 Design Personalized Intelligent Defaults for Risk Mitigation. Risk intervention designers are recommended to move toward "Safety by Design" [83] by providing ways to empower youth to protect themselves online, similar to how the CSCW community has advocated for "Privacy by Design" [76]. Our work supports this new Safety by Design paradigm by providing a practical approach to identifying problematic profiles of youth risk experiences, which could aid in early intervention and support by delivering educational content for coping with such situations. Once a youth is categorized into one of the five profiles, immediate recommendations could be made on how they could best manage and personalize their online content and interactions. For instance, when a youth from the High Risk Perpetration profile experiences conflict or threats sent to them publicly or privately, such a system could provide personalized recommendations providing tailored support resources based on the nature of conflicts or threats identified in these conversations. These resources would include evidence-based coping strategies, reporting mechanisms, or access to helplines specifically designed to address such issues. In recent work, youth preferred this type of personalized online safety intervention, compared to more general warnings [5]. Targeted interventions, rather than generalized ones, have been shown to more effectively reduce risk behavior in other youth contexts, such as risky sexual behavior [68], substance misuse use [41], and depression [128]. Therefore, we argue that these evidence-based approaches should also be applied to online contexts and the design of systems where youth encounter these risks.

5.4.3 Profile Youth Based on Unique Risk Experiences, Not Personal Characteristics. Importantly, the differences in youth profiles did not arise based on their personal characteristics but were rooted in their lived risk experiences, highlighting the importance of profiling youth based on these experiences, rather than their personal traits, which may result in harmful racial or gendered profiling [105, 110]. Current Artificial Intelligent (AI) risk detection solutions were developed and marketed without public evaluation [109], especially from the youth. A recent trend in the CSCW works have shed light on the importance of incorporating youth evaluations when building AI risk detection solutions would not only enhance the accuracy of the AI models but also ensure digital equity, especially for socio-economically disadvantaged youth [103]. Youth in this study provided a level of risk awareness demonstrated in the alignment of their self-reported data and lived experiences; therefore, we recommend that an optimal way to move towards designing youth-centric AI solutions is by leveraging youth assessments for the AI risk predictions to inform about the quality of the risk predictions, using the human-in-the-loop approach [85]. By doing so, we could further enhance the accuracy and reliability of the models. At the same time, being able to accurately flag and reflect on these past experiences may help youth develop a higher level of risk

self-awareness when the risk occurred, which would benefit their future online communication through reflective learning instead of gaining awareness from regrettable past interactions. The real-time self-assessments would equip them with the necessary skills, resilience, and awareness for navigating risks in the future. It is important to note that these design implications should be conducted in an ethical way that does not unintentionally harm vulnerable youth populations.

5.5 Limitations and Future Work

While collecting Instagram private conversations from youth is a key strength of this study, it may also affect the interpretability and generalizability of our results. We chose to take an agnostic approach, which avoided making normative judgments on what we perceived as risky or unsafe; instead, we used a youth-centered lens by leveraging our participants' ground truth for risks in order to provide insights into what they perceived as unsafe. Yet, future research is recommended to triangulate our results using the opinions of other stakeholders, such as clinical experts and parents to provide a more holistic understanding of youths' risk experiences. Additionally, since our results were based on youth experiences on Instagram, they might not be generalizable to other social media platforms characterized by different youth demographics, moderation strategies, and/or affordances. Therefore, we recommend future research to investigate risks that occur on other platforms to validate the alignment between self-reports and digital trace data. Furthermore, this study was conducted with youth (ages 13-21) in the United States; therefore, the results should be generalized to only this youth population. Future research is warranted to conduct the study across different countries, where the GDPR rules might be applied as well as the different cultural norms that may influence youth offline and online behaviors. The participants were self-selected to participate in this study, which investigated the risk behaviors of youth. This sample could be subject to sample bias toward the ones who are victims of online risks or abuse.

6 CONCLUSION

In an increasingly digitized society, the lives of modern-day youth are ever-complex as they encounter new opportunities and risks offline, online, and at the intersection of these two converging realms. Therefore, the risk prevention strategies we develop to keep youth safe both offline and online need to do a better job in reflecting these complicated interactions and multi-faceted risk behaviors and experiences. While most youth typically encounter low to moderate risks that manifest in online spaces, we must be mindful and proactive in addressing the needs of higher risk youth, who struggle, for whatever reason, with increased propensity to engage in sexually risky interactions, self-harm, violent, and illegal activities online that can lead to increased real-world physical and emotional harm. As we learned through our study and analyses, many youth are willing to share their personal struggles with researchers, so that we can better understand the real struggles they are up against; and in return, it is our job to listen and continue to find ways to help them.

ACKNOWLEDGMENTS

This research is supported in part by the U.S. National Science Foundation under grants #IIP-2329976, #IIS-2333207, #CNS-1942610 and by the William T. Grant Foundation grant #187941. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors. We would also like to thank all the participants who donated their data and contributed to the research.

REFERENCES

[1] [n. d.]. Internet safety: Online safety for youth. https://www.ojp.gov/feature/internet-safety/online-safety-youth

[2] 2022. Online safety (for teens) - nemours kidshealth. https://kidshealth.org/en/teens/internet-safety.html
[3] Mona Abdelraheem, John McAloon, and Fiona Shand. 2019. Mediating and moderating variables in the prediction of self-harm in young people: A systematic review of prospective longitudinal studies. *Journal of affective disorders* 246 (2019), 14–28.
[4] Substance Abuse and Mental Health Services Administration (SAMHSA). 2023. Samhsa announces National Survey on Drug Use and Health (NSDUH) results detailing mental illness and substance use levels in 2021. https://www.hhs.gov/about/news/2023/01/04/samhsa-announces-national-survey-drug-use-health-results-detailing-mental-illness-substance-use-levels-2021.html
[5] Zainab Agha, Karla Badillo-Urquiola, and Pamela J Wisniewski. 2023. " Strike at the Root": Co-designing Real-Time Social Media Interventions for Adolescent Online Risk Prevention. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–32.
[6] Robert Agnew. 1992. Foundation for a general strain theory of crime and delinquency. *Criminology* 30, 1 (1992), 47–88.
[7] Maliha Ali, Tiffany R Gray, Diane J Martinez, Laurel E Curry, and Kimberly A Horn. 2016. Risk profiles of youth single, dual, and poly tobacco users. *Nicotine & Tobacco Research* 18, 7 (2016), 1614–1621.
[8] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Chen Ling, Munmun De Choudhury, Pamela J. Wisniewski, and Gianluca Stringhini. 2023. Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 132 (apr 2023), 30 pages. https://doi.org/10.1145/3579608
[9] Ashwaq Alsoubai, Xavier V Caddle, Ryan Doherty, Alexandra Taylor Koehler, Estefania Sanchez, Munmun De Choudhury, and Pamela J Wisniewski. 2022. MOSafely, Is that Sus? A Youth-Centric Online Risk Assessment Dashboard. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*. 197–200.
[10] Ashwaq Alsoubai, Jihye Song, Afsaneh Razi, Nurun Naher, Munmun De Choudhury, and Pamela J Wisniewski. 2022. From'Friends with Benefits' to'Sextortion:'A Nuanced Investigation of Adolescents' Online Sexual Risk Experiences. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.
[11] Jack L Andrews, Saz P Ahmed, and Sarah-Jayne Blakemore. 2021. Navigating the social environment in adolescence: The role of social brain development. *Biological Psychiatry* 89, 2 (2021), 109–118.
[12] Florian Arendt, Sebastian Scherr, and Daniel Romer. 2019. Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults. *New Media & Society* 21, 11-12 (2019), 2422–2442.
[13] Sara Atske. 2022. Teens and cyberbullying 2022. https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/
[14] Randy P Auerbach and Casey K Gardiner. 2012. Moving beyond the trait conceptualization of self-esteem: The prospective effect of impulsiveness, coping, and risky behavior engagement. *Behaviour research and therapy* 50, 10 (2012), 596–603.
[15] Emma Louise Backe, Pamela Lilleston, and Jennifer McCleary-Sills. 2018. Networked individuals, gendered violence: A literature review of cyberviolence. *Violence and gender* 5, 3 (2018), 135–146.
[16] Susanne E Baumgartner, Sindy R Sumter, Jochen Peter, and Patti M Valkenburg. 2012. Identifying teens at risk: Developmental pathways of online and offline sexual risk behavior. *Pediatrics* 130, 6 (2012), e1489–e1496.
[17] Eric G Benotsch, Daniel J Snipes, Aaron M Martin, and Sheana S Bull. 2013. Sexting, substance use, and sexual risk behavior in young adults. *Journal of adolescent health* 52, 3 (2013), 307–313.
[18] Asia S Bishop, Christopher M Fleming, and Paula S Nurius. 2020. Substance use profiles among gang-involved youth: social ecology implications for service approaches. *Children and youth services review* 119 (2020), 105600.
[19] Katja Boersma and Karin Lindblom. 2009. Stability and change in burnout profiles over time: A prospective study in the working population. *Work & Stress* 23, 3 (2009), 264–283.
[20] Rebecca C Brown, Tony Fischer, A David Goldwich, Frieder Keller, Robert Young, and Paul L Plener. 2018. # cutting: Non-suicidal self-injury (NSSI) on Instagram. *Psychological medicine* 48, 2 (2018), 337–346.
[21] Adina Bucur, Sorin Ursoniu, Constantin Caraion-Buzdea, Virgil Ciobanu, Silvia Florescu, and Cristian Vladescu. 2020. Aggressive behaviors among 15–16-Year-old Romanian High School Students: results from two consecutive surveys related to alcohol and other drug use at the European Level. *International journal of environmental research and public health* 17, 10 (2020), 3670.
[22] Xavier V Caddle, Afsaneh Razi, Seunghyun Kim, Shiza Ali, Temi Popo, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 2021. MOSafely: Building an Open-Source HCAI Community to Make the Internet a Safer Place for Youth. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 315–318.
[23] Esther Calvete, Izaskun Orue, Liria Fernández-González, and Angel Prieto-Fidalgo. 2019. Effects of an incremental theory of personality intervention on the reciprocity between bullying and cyberbullying victimization and perpetration

in adolescents. *PLoS one* 14, 11 (2019), e0224755.

[24] JJ Carballo, C Llorente, L Kehrmann, Itziar Flamarique, A Zuddas, D Purper-Ouakil, PJ Hoekstra, D Coghill, UME Schulze, RW Dittmann, et al. 2020. Psychosocial risk factors for suicidality in children and adolescents. *European child & adolescent psychiatry* 29 (2020), 759–776.

[25] Jennifer L Carey, Stephanie Carreiro, Brittany Chapman, Nathalie Nader, Peter R Chai, Sherry Pagoto, and Danielle E Jake-Schoffman. 2018. SoMe and Self Harm: The use of social media in depressed and suicidal youth. In *Proceedings of the... Annual Hawaii International Conference on System Sciences. Annual Hawaii International Conference on System Sciences*, Vol. 2018. NIH Public Access, 3314.

[26] Marina Carvalho, Cátia Branquinho, and Margarida Gaspar de Matos. 2021. Cyberbullying and bullying: Impact on psychological symptoms and well-being. *Child Indicators Research* 14 (2021), 435–452.

[27] Natalie Castellanos-Ryan, Jean-Baptiste Pingault, Sophie Parent, Frank Vitaro, Richard E Tremblay, and Jean R Seguin. 2017. Adolescent cannabis use, change in neurocognitive function, and high-school graduation: A longitudinal study from early adolescence to young adulthood. *Development and psychopathology* 29, 4 (2017), 1253–1266.

[28] Giulio Castelpietra, Ann Kristin Skrindo Knudsen, Emilie E Agardh, Benedetta Armocida, Massimiliano Beghi, Kim Moesgaard Iburg, Giancarlo Logroscino, Rui Ma, Fabrizio Starace, Nicholas Steel, et al. 2022. The burden of mental disorders, substance use disorders and self-harm among young people in Europe, 1990–2019: Findings from the Global Burden of Disease Study 2019. *The Lancet Regional Health-Europe* 16 (2022), 100341.

[29] Eunseong Cho and Seonghoon Kim. 2015. Cronbach's coefficient alpha: Well known but poorly understood. *Organizational research methods* 18, 2 (2015), 207–230.

[30] Shaunna L Clark, Bengt Muthén, Jaakko Kaprio, Brian M D'Onofrio, Richard Viken, and Richard J Rose. 2013. Models and strategies for factor mixture analysis: An example concerning the structure underlying psychological disorders. *Structural equation modeling: a multidisciplinary journal* 20, 4 (2013), 681–703.

[31] Jennifer E Copp, Elizabeth A Mumford, and Bruce G Taylor. 2021. Online sexual harassment and cyberbullying in a nationally representative sample of teens: Prevalence, predictors, and consequences. *Journal of Adolescence* 93 (2021), 202–211.

[32] Joran Cornelisse and Reshmi Gopalakrishna Pillai. 2020. Age Inference on Twitter using SAGE and TF-IGM. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*. 24–30.

[33] Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin* 52, 4 (1955), 281.

[34] Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. 2004. An anova test for functional data. *Computational statistics & data analysis* 47, 1 (2004), 111–122.

[35] Corinne David-Ferdon, Alana M Vivolo-Kantor, Linda L Dahlberg, Khiya J Marshall, Neil Rainford, and Jeffery E Hall. 2016. A comprehensive technical package for the prevention of youth violence and associated risk behaviors. (2016).

[36] Jordan P Davis, Eric R Pedersen, Joan S Tucker, Michael S Dunbar, Rachana Seelam, Regina Shih, and Elizabeth J D'Amico. 2019. Long-term associations between substance use-related media exposure, descriptive norms, and alcohol use from adolescence to young adulthood. *Journal of youth and adolescence* 48, 7 (2019), 1311–1326.

[37] Caoimhe Doyle, Ellen Douglas, and Gary O'Reilly. 2021. The outcomes of sexting for children and adolescents: A systematic review of the literature. *Journal of Adolescence* 92 (2021), 86–113.

[38] Rebecca Dredge, John Gleeson, and Xochitl De la Piedad Garcia. 2014. Cyberbullying in social networking sites: An adolescent victim's perspective. *Computers in human behavior* 36 (2014), 13–20.

[39] Rebecca Dredge and Lara Schreurs. 2020. Social media use and offline interpersonal outcomes during youth: A systematic literature review. *Mass Communication and Society* 23, 6 (2020), 885–911.

[40] Suqian Duan, Zhizhou Duan, Ronghua Li, Amanda Wilson, Yuanyuan Wang, Qiufang Jia, Yong Yang, Mengqing Xia, Guosheng Wang, Tingting Jin, et al. 2020. Bullying victimization, bullying witnessing, bullying perpetration and suicide risk among adolescents: A serial mediation analysis. *Journal of affective disorders* 273 (2020), 274–279.

[41] Hanie Edalati and Patricia J Conrod. 2019. A review of personality-targeted interventions for prevention of substance misuse and related harm in community samples of adolescents. *Frontiers in psychiatry* 9 (2019), 770.

[42] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 1041–1048.

[43] Julius Erdmann. 2020. The Aberrant Medial Body Visual Representations of Self-Harming Behavior on Social Network Sites. *Medial Bodies between Fiction and Faction: Reinventing Corporeality* (2020), 245.

[44] Myriam Forster, Chris J Rogers, Stephanie M Benjamin, Timothy Grigsby, Katherine Lust, and Marla E Eisenberg. 2019. Adverse childhood experiences, ethnicity, and substance use among college students: Findings from a two-state sample. *Substance use & misuse* 54, 14 (2019), 2368–2379.

[45] Isaac Chun-Hai Fung, Elizabeth B Blankenship, Jennifer O Ahweyevu, Lacey K Cooper, Carmen H Duke, Stacy L Carswell, Ashley M Jackson, Jimmy C Jenkins III, Emily A Duncan, Hai Liang, et al. 2020. Public health implications of image-based social media: a systematic review of Instagram, Pinterest, Tumblr, and Flickr. *The Permanente Journal*

24 (2020).
[46] Martin I Gallegos, Brittany Zaring-Hinkle, Nan Wang, and James H Bray. 2021. Detachment, peer pressure, and age of first substance use as gateways to later substance use. *Drug and Alcohol Dependence* 218 (2021), 108352.
[47] Manuel Gámez-Guadix and Estibaliz Mateos-Pérez. 2019. Longitudinal and reciprocal relationships between sexting, online sexual solicitations, and cyberbullying among minors. *Computers in Human Behavior* 94 (2019), 70–76.
[48] Aina M. Gassó, Bianca Klettke, José R. Agustina, and Irene Montiel. 2019. Sexting, Mental Health, and Victimization Among Adolescents: A Literature Review. *International Journal of Environmental Research and Public Health* 16, 13 (Jan. 2019), 2364. https://doi.org/10.3390/ijerph16132364 Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.
[49] General Data Protection Regulation (GDPR). 2021. Art. 20 GDPR – Right to data portability | General Data Protection Regulation (GDPR). https://gdpr-info.eu/art-20-gdpr/
[50] Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*. 229–233.
[51] Anke Görzig. 2016. Adolescents' experience of offline and online risks: Separate and joint propensities. *Computers in Human Behavior* 56 (2016), 9–13.
[52] Petra Gradinger, Dagmar Strohmeier, and Christiane Spiel. 2009. Traditional bullying and cyberbullying: Identification of risk groups for adjustment problems. *Zeitschrift für Psychologie/Journal of Psychology* 217, 4 (2009), 205–213.
[53] Alexa Guy, Kirsty Lee, and Dieter Wolke. 2019. Comparisons between adolescent bullies, victims, and bully-victims on perceived popularity, social impact, and social preference. *Frontiers in psychiatry* 10 (2019), 868.
[54] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. Safe sexting: The advice and support adolescents receive from peers regarding online sexual risks. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31.
[55] Heidi Hartikainen, Afsaneh Razi, and Pamela Wisniewski. 2021. 'If You Care About Me, You'll Send Me a Pic'-Examining the Role of Peer Pressure in Adolescent Sexting. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 67–71.
[56] Martine Hébert, Laetitia Mélissande Amédée, Valérie Théorêt, and Marie-Pier Petit. 2021. Diversity of adaptation profiles in youth victims of child sexual abuse. *Psychological trauma: theory, research, practice, and policy* (2021).
[57] Angela K Henneberger, Dawnsha R Mushonga, and Alison M Preston. 2021. Peer influence and adolescent substance use: A systematic review of dynamic social network research. *Adolescent Research Review* 6, 1 (2021), 57–73.
[58] Mireille Hildebrandt. 2008. Defining profiling: a new type of knowledge? In *Profiling the European citizen*. Springer, 17–45.
[59] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.
[60] Kana Imuta, Sumin Song, Julie D Henry, Ted Ruffman, Candida Peterson, and Virginia Slaughter. 2022. A meta-analytic review on the social–emotional intelligence correlates of the six bullying roles: Bullies, followers, victims, bully-victims, defenders, and outsiders. *Psychological Bulletin* 148, 3-4 (2022), 199.
[61] Haiyan Jia, Pamela J Wisniewski, Heng Xu, Mary Beth Rosson, and John M Carroll. 2015. Risk-taking as a learning process for shaping teen's online information privacy behaviors. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 583–599.
[62] Bu Kyung Kim, Jisu Park, Hi Jae Jung, and Yoonsun Han. 2020. Latent profiles of offline/cyber bullying experiences among Korean students and its relationship with peer conformity. *Children and Youth Services Review* 118 (2020), 105349.
[63] Teresa Kirchner, Ernesto Magallón-Neri, Maria Forns, Dámaris Muñoz, Anna Segura, Laia Soler, and Irina Planellas. 2020. Facing interpersonal violence: identifying the coping profile of poly-victimized resilient adolescents. *Journal of interpersonal violence* 35, 9-10 (2020), 1934–1957.
[64] E David Klonsky and Catherine R Glenn. 2009. Assessing the functions of non-suicidal self-injury: Psychometric properties of the Inventory of Statements About Self-injury (ISAS). *Journal of psychopathology and behavioral assessment* 31, 3 (2009), 215–219.
[65] Bart P Knijnenburg, Alfred Kobsa, and Hongxia Jin. 2013. Dimensionality of information disclosure behavior. *International Journal of Human-Computer Studies* 71, 12 (2013), 1144–1162.
[66] Eva M Kunst, Marianne van Woerkom, Geert H van Kollenburg, and Rob F Poell. 2018. Stability and change in teachers' goal orientation profiles over time: Managerial coaching behavior as a predictor of profile change. *Journal of Vocational Behavior* 104 (2018), 115–127.
[67] Hannah M Layman, Ingibjorg Eva Thorisdottir, Thorhildur Halldorsdottir, Inga Dora Sigfusdottir, John P Allegrante, and Alfgeir Logi Kristjansson. 2022. Substance use among youth during the COVID-19 pandemic: a systematic review. *Current psychiatry reports* 24, 6 (2022), 307–324.

[68] Craig Winston LeCroy, Skyler Milligan-LeCroy, and Darlene Lopez. 2022. Guy talk: a gender-specific sexual education program to reduce sexual risk behaviors with high school males. *Health Education & Behavior* 49, 4 (2022), 593–602.
[69] Mi-Ting Lin. 2016. Risk factors associated with cyberbullying victimization and perpetration among Taiwanese children. (2016).
[70] Gitta H Lubke and Bengt Muthén. 2005. Investigating population heterogeneity with factor mixture models. *Psychological methods* 10, 1 (2005), 21.
[71] Xi Luo, James J Yang, Anne Buu, Elisa M Trucco, and Chiang-Shan R Li. 2022. Alcohol and cannabis co-use and longitudinal gray matter volumetric changes in early and late adolescence. *Addiction biology* 27, 5 (2022), e13208.
[72] Xiao Ma, Jeff Hancock, and Mor Naaman. 2016. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 3857–3869.
[73] Sheri Madigan, Daphne J Korczak, Tracy Vaillancourt, Nicole Racine, Will G Hopkins, Paolo Pador, Jackson MA Hewitt, Batool AlMousawi, Sheila McDonald, and Ross D Neville. 2023. Comparison of paediatric emergency department visits for attempted suicide, self-harm, and suicidal ideation before and during the COVID-19 pandemic: a systematic review and meta-analysis. *The Lancet Psychiatry* (2023).
[74] Sheri Madigan, Vanessa Villani, Corry Azzopardi, Danae Laut, Tanya Smith, Jeff R Temple, Dillon Browne, and Gina Dimitropoulos. 2018. The prevalence of unwanted online sexual exposure and solicitation among youth: a meta-analysis. *Journal of Adolescent Health* 63, 2 (2018), 133–141.
[75] Srinagesh Mannekote Thippaiah, Muralidhara Shankarapura Nanjappa, Jayasudha G Gude, Emanuel Voyiaziakis, Sohum Patwa, Badari Birur, and Ananda Pandurangi. 2021. Non-suicidal self-injury in developing countries: A review. *International journal of social psychiatry* 67, 5 (2021), 472–482.
[76] Ameera Mansour and Helena Francke. 2021. Collective privacy management practices: A study of privacy strategies and risks in a private facebook group. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.
[77] Becky Mars, Jon Heron, E David Klonsky, Paul Moran, Rory C O'Connor, Kate Tilling, Paul Wilkinson, and David Gunnell. 2019. Predictors of future suicide attempt among adolescents with suicidal thoughts or non-suicidal self-harm: a population-based birth cohort study. *The Lancet Psychiatry* 6, 4 (2019), 327–337.
[78] Alice E Marwick and danah boyd. 2014. Networked privacy: How teenagers negotiate context in social media. *New media & society* 16, 7 (2014), 1051–1067.
[79] Gloria Mbokota and Alison Reid. 2022. The role of group coaching in developing leadership effectiveness in a business school leadership development programme. *South African Journal of Business Management* 53, 1 (2022), 10.
[80] Mary L McHugh. 2011. Multiple comparison analysis testing in ANOVA. *Biochemia medica* 21, 3 (2011), 203–209.
[81] R Kathryn McHugh and Roger D Weiss. 2019. Alcohol use disorder and depressive disorders. *Alcohol research: current reviews* 40, 1 (2019).
[82] Aksha M Memon, Shiva G Sharma, Satyajit S Mohite, and Shailesh Jain. 2018. The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature. *Indian journal of psychiatry* 60, 4 (2018), 384.
[83] Nathalie Meurens. 2022. Child safety by design that works against online sexual exploitation of children. https://www.academia.edu/84089223/Child_safety_by_design_that_works_against_online_sexual_exploitation_of_children
[84] Kimberly J Mitchell and Lisa M Jones. 2011. Youth Internet Safety Study (YISS): Methodology Report. (2011).
[85] Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.
[86] Camille Mori, Jeff R Temple, Dillon Browne, and Sheri Madigan. 2019. Association of sexting with sexual behaviors and mental health among adolescents: A systematic review and meta-analysis. *JAMA pediatrics* 173, 8 (2019), 770–779.
[87] John C Moskop and Kenneth V Iserson. 2007. Triage in medicine, part II: Underlying values and principles. *Annals of emergency medicine* 49, 3 (2007), 282–287.
[88] Bengt Muthén and Bengt O Muthén. 2009. *Statistical analysis with latent variables*. Wiley New York, NY.
[89] Aditi Nath, Sonali G Choudhari, Sarika U Dakhode, Asmita Rannaware, and Abhay M Gaidhane. 2022. Substance abuse amongst adolescents: an issue of public health significance. *Cureus* 14, 11 (2022).
[90] Naresh Nebhinani, Pranshu Singh, and Mamta. 2022. Substance Use Disorders in Children and Adolescents. *Journal of Indian Association for Child and Adolescent Mental Health* 18, 2 (2022), 128–136.
[91] Egil Nygaard, Kari Slinning, Vibeke Moe, Anders Fjell, and Kristine B Walhovd. 2020. Mental health in youth prenatally exposed to opioids and poly-drugs and raised in permanent foster/adoptive homes: A prospective longitudinal study. *Early human development* 140 (2020), 104910.
[92] Karen L Nylund, Tihomir Asparouhov, and Bengt O Muthén. 2007. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal* 14, 4 (2007), 535–569.

[93] Sara Pabian, Sara Erreygers, Heidi Vandebosch, Kathleen Van Royen, Julie Dare, Leesa Costello, Lelia Green, Dianne Hawk, and Donna Cross. 2018. "Arguments online, but in school we always act normal": The embeddedness of early adolescent negative peer interactions within the whole of their offline and online peer interactions. *Children and youth services review* 86 (2018), 1–13.

[94] Jinkyung Park, Mamtaj Akter, Naima Samreen Ali, Zainab Agha, Ashwaq Alsoubai, and Pamela Wisniewski. 2023. Towards Resilience and Autonomy-Based Approaches for Adolescents Online Safety. *Available at SSRN 4608406* (2023).

[95] Frank W Paulus, Jens Joas, Ida Gerstner, Anna Kühn, Markus Wenning, Thomas Gehrke, Holger Burckhart, Ulf Richter, Alexandra Nonnenmacher, Michael Zemlin, et al. 2022. Problematic Internet Use among Adolescents 18 Months after the Onset of the COVID-19 Pandemic. *Children* 9, 11 (2022), 1724.

[96] Rafael Pichel, Sandra Feijóo, Manuel Isorna, Jesús Varela, and Antonio Rial. 2022. Analysis of the relationship between school bullying, cyberbullying, and substance use. *Children and Youth Services Review* 134 (2022), 106369.

[97] Anthony T Pinter, Pamela J Wisniewski, Heng Xu, Mary Beth Rosson, and Jack M Caroll. 2017. Adolescent online safety: Moving beyond formative evaluations to designing solutions for the future. In *Proceedings of the 2017 Conference on Interaction Design and Children*. 352–357.

[98] Lynn E Ponton and Samuel Judice. 2004. Typical adolescent sexual development. *Child and Adolescent Psychiatric Clinics* 13, 3 (2004), 497–511.

[99] Megan L Ranney, Sarah K Pittman, Alison Riese, Christopher Koehler, Michele L Ybarra, Rebecca M Cunningham, Anthony Spirito, and Rochelle K Rosen. 2020. What counts?: A qualitative study of adolescents' lived experience with online victimization and cyberbullying. *Academic pediatrics* 20, 4 (2020), 485–492.

[100] Jiaming Rao, Haiqing Wang, Minhui Pang, Jianwei Yang, Jiayi Zhang, Yunfeng Ye, Xiongfei Chen, Shengyong Wang, and Xiaomei Dong. 2019. Cyberbullying perpetration and victimisation among junior and senior high school students in Guangzhou, China. *Injury prevention* 25, 1 (2019), 13–19.

[101] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2023. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 89 (apr 2023), 29 pages. https://doi.org/10.1145/3579522

[102] Afsaneh Razi, Karla Badillo-Urquiola, and Pamela J Wisniewski. 2020. Let's Talk about Sext: How Adolescents Seek Support and Advice about Their Online Sexual Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[103] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Xavier Caddle, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela Wisniewski. 2021. Teens at the Margin: Artificially Intelligent Technology for Promoting Adolescent Online Safety. In *ACM Conference on Human Factors in Computing Systems (CHI 2021)/Artificially Intelligent Technology for the Margins: A Multidisciplinary Design Agenda Workshop*.

[104] Torkel Richert, Mats Anderberg, and Mikael Dahlberg. 2020. Mental health problems among young people in substance abuse treatment in Sweden. *Substance abuse treatment, prevention, and policy* 15, 1 (2020), 1–10.

[105] Juan A Ríos Vega. 2020. School to deportation pipeline: Latino youth counter-storytelling narratives. *Journal of Latinos and Education* (2020), 1–13.

[106] Mónica Rodríguez-Enríquez, Miquel Bennasar-Veny, Alfonso Leiva, and Aina M Yañez. 2019. Alcohol and tobacco consumption, personality, and cybervictimization among adolescents. *International journal of environmental research and public health* 16, 17 (2019), 3123.

[107] Julie C Rusby, John M Light, Ryann Crowley, and Erika Westling. 2018. Influence of parent–youth relationship, parental monitoring, and parent substance use on adolescent substance use onset. *Journal of family psychology* 32, 3 (2018), 310.

[108] Christopher P Salas-Wright, Michael G Vaughn, David R Hodge, and Brian E Perron. 2012. Religiosity profiles of American youth in relation to substance use, violence, and delinquency. *Journal of youth and adolescence* 41 (2012), 1560–1575.

[109] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*. Vol. 11700. Springer Nature.

[110] Kanika Samuels-Wortley. 2021. To serve and protect whom? Using composite counter-storytelling to explore Black and Indigenous youth experiences and perceptions of the police in Canada. *Crime & Delinquency* 67, 8 (2021), 1137–1164.

[111] Rocío Sánchez-Salmerón, José L Gómez-Urquiza, Luis Albendín-García, María Correa-Rodríguez, María Begoña Martos-Cabrera, Almudena Velando-Soriano, and Nora Suleiman-Martos. 2022. Machine learning methods applied to triage in emergency services: A systematic review. *International Emergency Nursing* 60 (2022), 101109.

[112] Sebastian Scherr, Florian Arendt, Thomas Frissen, and José Oramas M. 2020. Detecting intentional self-harm on Instagram: development, testing, and validation of an automatic image-recognition algorithm to discover cutting-related posts. *Social science computer review* 38, 6 (2020), 673–685.

[113] Gianluca Serafini, Andrea Aguglia, Andrea Amerio, Giovanna Canepa, Giulia Adavastro, Claudia Conigliaro, Jacopo Nebbia, Larissa Franchi, Eirini Flouri, and Mario Amore. 2021. The relationship between bullying victimization and perpetration and non-suicidal self-injury: a systematic review. *Child Psychiatry & Human Development* (2021), 1–22.

[114] Gianluca Serafini, Andrea Aguglia, Andrea Amerio, Giovanna Canepa, Giulia Adavastro, Claudia Conigliaro, Jacopo Nebbia, Larissa Franchi, Eirini Flouri, and Mario Amore. 2023. The relationship between bullying victimization and perpetration and non-suicidal self-injury: A systematic review. *Child Psychiatry & Human Development* 54, 1 (2023), 154–175.

[115] Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review* 5, 1 (2001), 3–55.

[116] Jennifer D. Shapka and Rose Maghsoudi. 2017. Examining the validity and reliability of the cyber-aggression and cyber-victimization scale. *Computers in Human Behavior* 69 (April 2017), 10–17. https://doi.org/10.1016/j.chb.2016.12.015

[117] Donald Sharpe. 2015. Chi-square test is statistically significant: Now what? *Practical Assessment, Research, and Evaluation* 20, 1 (2015), 8.

[118] Daniel TL Shek, Xiaoqin Zhu, Diya Dou, and Wenyu Chai. 2020. Influence of family factors on substance use in early adolescents: A longitudinal study in Hong Kong. *Journal of Psychoactive Drugs* 52, 1 (2020), 66–76.

[119] Crystal Lederhos Smith, Brittany Rhoades Cooper, Andre Miguel, John Roll, Laura Hill, Michael Cleveland, and Sterling McPherson. 2022. Youth risk profiles and their prediction of distal cannabis and tobacco co-use in the Population Assessment of Tobacco Health (PATH). *Substance Abuse* 43, 1 (2022), 733–741.

[120] Michael A Tarrant, Michael J Manfredo, Peter B Bayley, and Richard Hess. 1993. Effects of recall bias and nonresponse bias on self-report estimates of angling participation. *North American Journal of Fisheries Management* 13, 2 (1993), 217–222.

[121] Jean M Twenge. 2020. Increases in depression, self-harm, and suicide among US adolescents after 2012 and links to technology use: possible mechanisms. *Psychiatric Research and Clinical Practice* 2, 1 (2020), 19–25.

[122] Sarah E Victor, Alison E Hipwell, Stephanie D Stepp, and Lori N Scott. 2019. Parent and peer relationships as longitudinal predictors of adolescent non-suicidal self-injury onset. *Child and adolescent psychiatry and mental health* 13, 1 (2019), 1–13.

[123] Sebastian Wachs. [n. d.]. Hate Speech and Bullying: Two sides of the same coin? ([n. d.]).

[124] Sebastian Wachs, Michelle F Wright, Manuel Gámez-Guadix, and Nicola Döring. 2021. How are consensual, non-consensual, and pressured sexting linked to depression and self-harm? The moderating effects of demographic variables. *International journal of environmental research and public health* 18, 5 (2021), 2597.

[125] Sebastian Wachs, Michelle F Wright, and Alexander T Vazsonyi. 2019. Understanding the overlap between cyberbullying and cyberhate perpetration: Moderating effects of toxic online disinhibition. *Criminal Behaviour and Mental Health* 29, 3 (2019), 179–188.

[126] James P Walsh. 2020. Social media and moral panics: Assessing the effects of technological change on societal reaction. *International Journal of Cultural Studies* 23, 6 (2020), 840–859.

[127] Glenn D Walters and Dorothy L Espelage. 2018. From victim to victimizer: Hostility, anger, and depression as mediators of the bullying victimization–bullying perpetration association. *Journal of school psychology* 68 (2018), 73–83.

[128] Aliza Werner-Seidler, Yael Perry, Alison L Calear, Jill M Newby, and Helen Christensen. 2017. School-based depression and anxiety prevention programs for young people: A systematic review and meta-analysis. *Clinical psychology review* 51 (2017), 30–47.

[129] Stuart F White, Joel L Voss, Jessica J Chiang, Lei Wang, Katie A McLaughlin, and Gregory E Miller. 2019. Exposure to violence and low family income are associated with heightened amygdala responsiveness to threat among adolescents. *Developmental cognitive neuroscience* 40 (2019), 100709.

[130] Pamela Wisniewski. 2018. The Privacy Paradox of Adolescent Online Safety: A Matter of Risk Prevention or Risk Resilience? *IEEE Security Privacy* 16, 2 (2018), 86–90. https://doi.org/10.1109/MSP.2018.1870874

[131] Pamela Wisniewski, Haiyan Jia, Na Wang, Saijing Zheng, Heng Xu, Mary Beth Rosson, and John M Carroll. 2015. Resilience mitigates the negative effects of adolescent internet addiction and online risk exposure. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 4029–4038.

[132] Pamela Wisniewski, Heng Xu, Jack Carroll, and Mary Beth Rosson. 2013. Grand challenges of researching adolescent online safety: a family systems approach. (2013).

[133] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F Perkins, and John M Carroll. 2016. Dear diary: Teens reflect on their weekly online risk experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3919–3930.

[134] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F. Perkins, and John M. Carroll. 2016. Dear Diary: Teens Reflect on Their Weekly Online Risk Experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3919–3930. https://doi.org/10.1145/2858036.2858317 event-place: San Jose, California, USA.

[135] Chia-Yi Wu, Robert Stewart, Hui-Chun Huang, Martin Prince, and Shen-Ing Liu. 2011. The impact of quality and quantity of social support on help-seeking behavior prior to deliberate self-harm. *General hospital psychiatry* 33, 1 (2011), 37–44.

[136] Jiping Yang, Shuang Li, Ling Gao, and Xingchao Wang. 2022. Longitudinal associations among peer pressure, moral disengagement and cyberbullying perpetration in adolescents. *Computers in Human Behavior* 137 (2022), 107420.

[137] Yoewon Yoon, Jungeun Olivia Lee, Junhan Cho, Mariel S Bello, Rubin Khoddam, Nathaniel R Riggs, and Adam M Leventhal. 2019. Association of cyberbullying involvement with subsequent substance use among adolescents. *Journal of Adolescent Health* 65, 5 (2019), 613–620.

[138] Milan Zarchev, Astrid Kamperman, Hanan El Marroun, Anthony Bloemendaal, Cornelis Mulder, Witte Hoogendijk, and Nina Grootendorst van Mil. 2022. Timing and severity of adverse life events: impact on substance use among high-risk adolescents. (2022).

[139] Yongcheng Zhan, Zhu Zhang, Janet M Okamoto, Daniel D Zeng, Scott J Leischow, et al. 2019. Underage JUUL use patterns: content analysis of Reddit messages. *Journal of medical Internet research* 21, 9 (2019), e13038.

[140] Lian Zhu, Nicholas J Westers, Sarah E Horton, Jessica D King, Andrew Diederich, Sunita M Stewart, and Betsy D Kennard. 2016. Frequency of exposure to and engagement in nonsuicidal self-injury among inpatient adolescents. *Archives of suicide research* 20, 4 (2016), 580–590.

[141] Izabela Zych, Carmen Viejo, Elena Vila, and David P Farrington. 2021. School bullying and dating violence in adolescents: A systematic review and meta-analysis. *Trauma, Violence, & Abuse* 22, 2 (2021), 397–412.

A SURVEY QUESTIONS

A.1 Online and Offline Risk Experiences Survey Questions

The following self-reported constructs were measured on a 5-point Likert Scale: Never - All the time.

Construct	Questions
Risky Behavior Questionnaire [14]	1. Have you destroyed property (other than your own)?
	2. Have you been unfaithful to your boyfriend or girlfriend?
	3. Have you been in a physical fight?
	4. Have you bullied, threatened, or intimidated a peer(s)?
	5. Have you been binge drinking and/or drinking to get drunk?
	6. Have you used illegal drugs?
	7. Have you sold illegal drugs?
	8. Have you skipped class (or entire days of school)?
	9. Have you cheated or plagiarized?
	10. Have you shoplifted?
	11. Have you stolen money?
	12. Have you had unsafe sex?
	13. Have you verbally harassed someone?
	14. Have you made attempts to cut or burn yourself?
	15. Have you purged or binged?
	16. Have you gambled?
	17. Have you lied to your family members (e.g., grandparents, parents, siblings)?
	18. Have you driven (a bicycle, a moped, and/or a car) recklessly (e.g., at fast speeds, under the influence of a substance)?
	19. Have you used cigarettes?
	20. Have you engaged in acts of revenge?
Inventory of Statements About Self-harm [64]	1. Cutting
	2. Severe Scratching
	3. Biting
	4. Banging or Hitting Self
	5. Burning
	”6. Interfering w/ Wound Healing (e.g., picking scabs)”
	7. Carving
	8. Rubbing Skin Against Rough Surface
	9. Pinching
	10. Sticking Self w/ Needles
	11. Pulling Hair
	12. Swallowing Dangerous Substances
Cyber-Aggression Victimization [116]	1. Had something embarrassing or mean posted or re-posted about you on Instagram.
	2. Received a hurtful message from someone on Instagram.
	3. Had an embarrassing photo or video of you posted or re-posted on Instagram that you didn’t want others to see.

Construct	Questions
	4. Had hurtful comments made on Instagram about an online photo or video of you.
	5. Been purposely excluded by others on Instagram.
	6. Had something personal posted or re-posted about you on Instagram that you didn't want others to know.
	7. Had gossip or rumors spread about you on Instagram.
	8. Received hurtful comments or messages about your race or ethnicity on Instagram.
	9. Received hurtful comments or messages about your perceived sexual orientation on Instagram.
	10. Received hurtful comments about your perceived sexual behaviors on Instagram.
	11. Received a sexual message from somebody on Instagram who was trying to be mean to you or to embarrass you.
	12. Had sexual content (photos or jokes) sent to you from somebody on Instagram who was trying to be mean to you or embarrass you.
Cyber-Aggression Perpetration [116]	1. Posted or re-posted something embarrassing or mean about another person on Instagram.
	2. Sent or forwarded a hurtful message to someone on Instagram.
	3. Posted or re-posted an embarrassing photo or video of someone on Instagram that he or she did not want others to see.
	4. Posted or texted a hurtful comment on Instagram about a photo or video of somebody else.
	5. Posted or sent messages on Instagram to purposely exclude a certain person or group of people.
	6. Posted or re-posted something private on Instagram about another person that he or she did not want others to know.
	7. Used Instagram to spread rumors or gossip about someone.
	8. Made hurtful comments about somebody's race or ethnicity on Instagram.
	9. Made hurtful comments about somebody's perceived sexual orientation on Instagram.
	10. Made hurtful comments about somebody's perceived sexual behaviors on Instagram.
	11. Said something sexual to somebody else on Instagram to embarrass them or to be mean.
	12. Sent sexual content (photos or jokes) to somebody else on Instagram to embarrass them or to be mean.
Unwanted Sexual Solicitations and Approaches [84]	1. Someone tried to get me to talk on Instagram about sex when I did not want to.
	2. Someone on Instagram asked me for sexual information about myself when I did not want to answer such questions. (Very personal questions, like what your body looks like or sexual things you have done)

Construct	Questions
	3. Someone on Instagram asked me to do something sexual that I did not want to do.
	4. My Instagram feed showed me pictures of naked people or people having sex when I did not want to see such content.
	5. Someone on Instagram sent me a direct message of naked people or people having sex that I did not want to see.
Youth Produced Sexual Images (Sexting) [84]	1. I have shared a nude or nearly nude picture or video of myself on Instagram.
	2. Someone else shared a nude or nearly nude picture or video of me on Instagram.

A.2 Demographic Information

Questions	Possible Responses
1. Please select your gender	a. Male
	b. Female
	c. Non-Binary
	d. Prefer to self-identify
2. Please select your current age.	a. 13
	b. 14
	c. 15
	d. 16
	e. 17
	f. 18
	g. 19
	h. 20
	i. 21
3. Please select your race. Check all that apply.	a. White/Caucasian
	b. Black/African-American
	c. Hispanic/Latino
	d. Asian or Pacific Islander
	e. American Indian/Alaska Native
	f. Prefer to Self-Identify
4. What is your sexual orientation?	a. Heterosexual or straight
	b. Homosexual or gay
	c. Bisexual
	d. Prefer to self-identify
5. Please select the city and state in which you live.	[Drop down of U.S. states and territories]

Received: January 2023; Revised: July 2023; Accepted: November 2023