EVALUATING ONLINE SAFETY

A Framework developed by UNICEF East Asia and Pacific Regional Office

(work in progress)

Contents

Definitions

Cyberbullying is a repeated series of aggressive, intentional acts, conducted through digital platforms and devices, that inflicts willful harm on an individual.

This definition builds on Olweus (1999) definition of bullying, while combining phrasings from both perpetrator-centered and victim-centered (Cyberbullying Research Center) definitions.

Online Grooming is a process where an adult uses digital media and platforms to befriend a minor and prepare them for a sexually abusive relationship.

This definition takes the lead from the Cambridge definition used by the International Centre for Missing & Exploited Children and several other studies, while also including some phrasing from Steffgen and Smith (2013).

Theories of Change

'Theories of change' differ from related ideas, like 'models of behaviour'(van der Linden 2013). While models are useful in understanding certain behaviours and the underlying factors determining them, a theory aims to provide a richer explanatory account of behaviour. Explanations can include any or all of: what leads to a certain behaviour, what its effects are, how it changes over time, and it might be influenced. Models are largely diagnostic, while theories of change focus on supporting interventions, predicting change and encouraging the adoption of new behaviour. Theories are also often more abstract than models or frameworks, and can extend from comparatively concrete – a theoretical account of why a specific intervention could be expected to work – to highly abstract – a theory of society or of mind, for example, that seeks to account for why a specific behaviour might appear. A theory of change also can illustrate the 'process of change' or the intermediate stages, and assumptions and linkages between them. It may be developed at the planning stage for an intervention or may adapt and respond to an ongoing intervention, taking into account emergent issues and decisions over time (Rogers 2014).

For example, a theory of change of cyberbullying victim behaviour might begin with an identifiable problematic pattern of that behaviour, such as why do victims continue to subject themselves to situations where bullying takes place? It might draw upon larger, more abstract theories, such as attachment theory, to suggest that victims are drawn to certain situations where they feel a sense of belonging and comfort – despite the pain they experience in those same situations due to bullying. It might suggest, also informed by this broader theory, that the desired behaviour change – removing the victim from the bullying situation – involves developing other sources that fulfil similar emotional needs. This would weaken the strength of this particular attachment, allowing the victim to remove themselves from the harmful situation, or at least experience it in a different way. To create this effect may mean exposing the victim to multiple other engaging social situations, offline and online, that act as alternative sources of belonging and comfort. This intervention can then be measured, at individual and aggregate levels, to determine whether it validates the underlying theory of change.

Applications and Contextual Considerations

Despite its origins, theories of change are usually deployed in operational or intervention contexts. Since a ToC is not purely sociological or psychological theory, but rather a pragmatic framework illustrating the precise linkages between intervention and change, it may deploy multiple theories within a singular intervention. A ToC can be supported by individual theories as the central links or assumptions in an intervention (De Silva et al. 2014).

A ToC can be used to design an intervention - *ex ante*, or before the fact – and can also be used to evaluate an intervention- *ex post*, or after the fact. Both types of ToCs can also be deployed as adaptive frameworks, accounting for emergent or unintended impacts. A ToC

can help identify specific evaluation queries and variables, as well as intermediate outcomes that can be evaluated during implementations to reveal individual steps to change, patterns as well as discrepancies. In order to generate a ToC (Aromatario et al. 2019), several types of explanatory theories, including classical theories, determinant theories and implementation theories (Nilsen 2015), can be selected to respond to the objectives, beneficiaries and context of the intervention.

A “classical” theory is usually one derived from earlier psychology or sociology, and refers often to holistic and encompassing theories of the self, of development, of learning, of mind, of society, or of some combination. Examples here could include Freudian, Jungian, Piagetian or Vygostian theories, as well as attachment theory referenced above. In the context of youth and childhood development, such theories rarely account for specific situational settings, such as cyberbullying, but rather examine in general if not universal terms how children develop. A determinant theory instead focuses on specific conditions or behaviours, along with their determinants, or causal factors. Such theories are often preferred in scientific contexts, since their specificity means they can be confirmed or denied empirically – through experiment, observation or self-reporting for instance. An example of a determinant theory might be Brofenbrenner’s (1981) ecological theory of the causal factors involved in violence. Implementation theories refer instead to ways or processes an intervention may be implemented, and how the details of these processes might in turn impact on outcomes. Such theories attend even more directly in the detail of specific situations; in the context of cyberbullying for example, an implementation theory might test whether online or offline modes of delivering content produce better outcomes.

It is tempting to view these three theoretical approaches along a scale, from big to small, or coarse to fine-grained. Along this scale, “classical” theories are sometimes called “Big-T” theories, while implementation and determinant theories are termed “Small-t” theories. While a useful guide, as we note above, theories of change do not need to be situated precisely along such a scale, and can be free to utilise both big T and small *t* theories. What matters, in our view, is the coherence between the levels of causation spelled out by the adopted theoretical approaches, the kinds of change predicted by them, and the extent to which the evaluation framework can register or measure these changes. For example, an open-ended and exploratory theory of change that suggests key determinants of cyberbullying in any concrete situation can never be known in advance would not align well with an evaluation framework that expects measurable and quantified change across a small number of pre-determined behavioural variables.

Mason and Barnes (2007) have argued that ToC can produce outcomes that are predetermined, rather than emergent and while it addresses a degree of complexity in interventions, as a framework it often ignores unintended outcomes and possibilities due to its overt emphasis on causal linkages. In order to address this limitation, feedback loops may be designed within a model integrating recursive modes of interventions allowing for transformations in communications as well as knowledge (Arensman et al. 2017). Mason and Barnes also argue that ToC could develop as narratives, as opposed to simple metric based evaluations, describing the context within which the programme is operating, the

nature of individual activities and services, partnering with stakeholders for a process of story-telling.

Furthermore, ToCs are implemented always within larger contextual frames, policies and legislation, so interventions are generally positioned within this landscape and behaviour. The Behaviour Change Wheel developed by Michie, van Stralen & West (2011) represents this ecosystem of functions as a possible layering of interventions of a ToC.

Behavioural Change Theories Applied to Cyberbullying and Cyber Grooming Interventions
Our review of candidate theories shows a range of approaches. Some, such as the Ecological Systems Theory of Change, emphasise the overlapping social layers that impact upon an individual's behaviour. Others, such as Nudge Theory, Theory of Reasoned Action and Empowerment Theory, focus on individuals exclusively.

Below we present several candidate theories, adapting general approaches from the literature to the specific issue of online cyberbullying. With ToC option D, we present our preferred hybrid or composite model, which integrates features of, in particular, an Ecological Systems Theory of Change and Nudge Theory. Our rationale for doing so is that these embody two extremes. One focussed on the entire ecology that surrounds an individual, from family up to national (e.g. legal, health) and global (e.g. Facebook, UNICEF) institutions. The other examines the micro-determinants of behaviour, and applies especially well to an online environment, where quite specific interactions can be designed for, observed and measured. Both have a strong evidence base, as noted above. This approach also leaves open the possibility for incorporating other theoretical models, such as General Strain Theory and Empowerment Theory.

Ecological Systems Theory
Brofenbrenner (1981) described five systems that act as ecological systems of interaction and development from childhood. The microsystem refers to the most immediate institutions and groups- such as the family and school while the mesosystem describes the interrelationships within these. The exosystem exerts indirect influence on the child and the macrosystem is the larger cultural context. Finally, the chronosystem describes the events and transitions over the time. Children interact with each of these ecological systems through their development, accommodating influences within their behaviours and knowledge. Thus any intervention needs to account for children within their 'systems' to carry out change.

Though varying across cultural, age and gender cohorts, cyberbullying is an unfortunately common online experience. Informed by *Ecological Systems Theory* and its adaptation to anti-bullying interventions in design, implementation and evaluation (Ortega-Baron et al. 2019, Swearer & Doll 2001), we consider the effects of cyberbullying to be produced through a combination of *personal*, *familial* and *school-based systems*. Each of these systems – a combination of individual and microsystems – needs to be adjusted to reduce prevalence and experience of bullying. An education program targeting young people (both perpetrators and victims) must be supplemented by adult-oriented materials distributed to parents and

teachers. Behavioural change, for both victim and perpetrator, depends upon continued reinforcement by supporting microsystems – family and school – for sustained success.

An evaluation framework for this ToC would identify measures of change at individual and aggregate (i.e. peer group) levels, as well as at family and school levels. Indicators would include changes in the knowledge and attitude, as well as behaviour, at each of these system levels.

"Nudge" Theory of Change

Nudge theory, in its purest form, suggests behaviour responds to a combination of cognitive shifts and "nudges" – incentives, prods or suggestions introduced into an environment. Applied to cyberbullying, an online environment could reward three kinds of anti-bullying behaviour – the cessation of bullying by perpetrators, more resilient responses by victims, and victim support offered by peers. Examples of incentives or nudges could include: warning prompts when aggressive or offensive text or emojis are being typed into chat text fields; congratulatory messages when those messages are re-keyed during typing or deleted after submission; praise when offensive material is reported, or when bullying behaviour is called out; and offers of support when victims block other users. Such examples rely upon increasingly prevalent automated systems to identify them, though social media policies and informal peer group standards also can reinforce them. Perversely, nudge theory also suggests that the most important causal influences on cyberbullying are online platforms themselves: a comprehensive "nudge" theory of change would therefore advocate for change within the (typically corporate) platform environments, including program managers, developers, UX (user experience) designers and content moderators.

Given its strictly behavioural focus, an evaluation framework for a nudge theory-driven ToC would focus on directly observable measures of change: less prevalence of cyberbullying "symptoms"; less reported incidents of cyberbullying (over time – in the short-term, a successful program may result in higher rates due to increased ease of reporting); and less tolerance within online communities for bullying types of behaviour.

General Strain Theory of Change

A general strain theory of change for cyberbullying would emphasise the perpetrator rather than victim (Paez 2016). Within this theoretical approach, violence is the outcome of diverse strains, which may be psychological, institutional, environmental, and so on. Change can only be brought about by addressing the root causes of these strains, and would need firstly to identify what these are. It would then need to consider strategies for addressing and minimising these strains, or alternatively find ways to re-channel the negative or violent emotions of perpetrators towards other, less destructive ends.

An evaluation framework for a general strain theory of change would similarly focus on perpetrator rather than victim behaviour. It might use a combination of qualitative – ie. verbal, or in an online context, written – and quantitative – ie. self-reported survey, as well as numbers of online reports of cyberbullying – measures. Distinct from nudge theory approaches, measurement would focus on cognitive and environmental as well as behavioural change. Unlike ecological systems theory, evaluation is less concerned here

with measuring change at levels beyond the individual (though these may come up as examples of causes of strain).

Empowerment Theory
An empowerment theory would focus on victims but may also include potential perpetrators as well as bystanders. According to this theory, individual, group, and community resources need to be made stronger in order to allow victims to exert a greater degree of control in various virtual environments and social media platforms. Here changes in behaviour are brought about by working on broader issues of self-esteem and building resilience among potential victims, equipping them with the tools to identify, report and cope with cyberbullying.

Evaluation of empowerment would focus primarily on victims and occasionally on peer groups who may be empowered to take an active role in reducing cyberbullying. Similar to evaluation approaches in strain theory, a combination of qualitative and quantitative measures may be used.

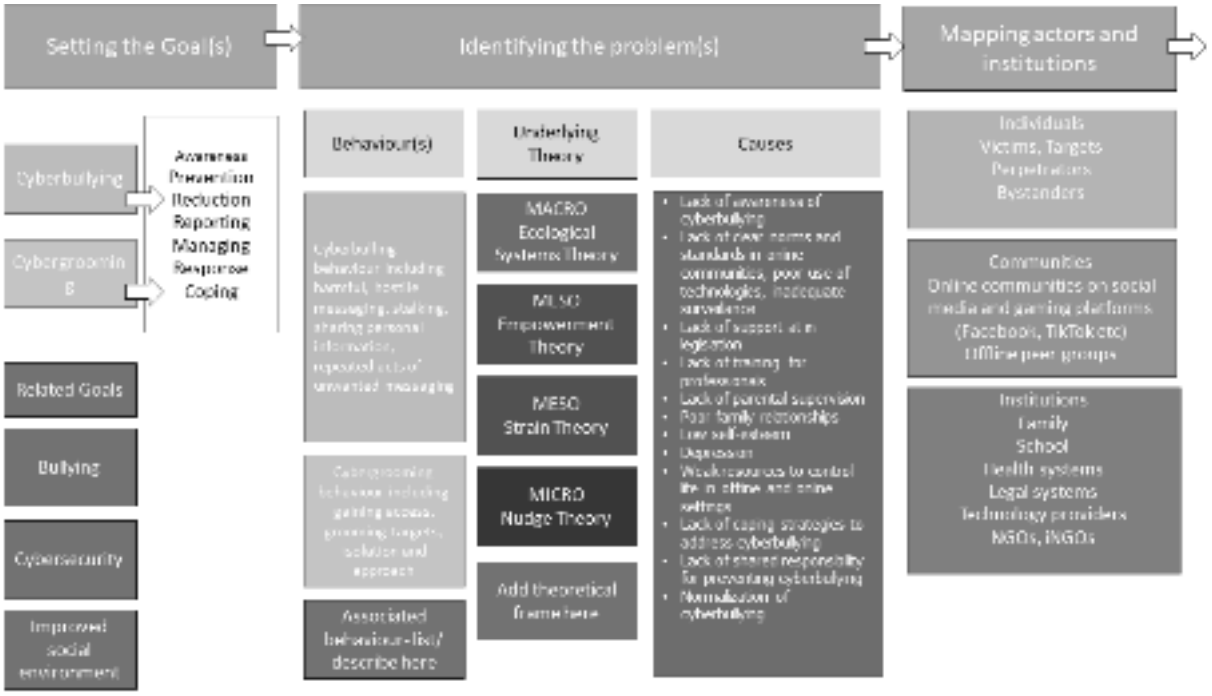
Composite Theory of Change
As noted above, theories of change do not need to be singular in their use of other theoretical models. Our preferred approach is to use a composite, informed by a combination of *Ecological Systems Theory*, *Nudge Theory* which may be supported by *General Strain Theory* and *Empowerment Theory* depending upon the actors being engaged within an intervention. A composite ToC may then, based upon the Ecological Systems Theory offer a vantage point for considering behaviour change at multiple levels; the macro, the meso and the micro levels spanning different causal pathways. In relation to cyberbullying and cybergrooming, we identify the following systems:

a. Individuals - both **victims** and **perpetrators** of online bullying
b. **Online communities**, including those on social media and gaming platforms (Facebook, TikTok etc)
c. Offline peer groups
d. Family
e. School
f. Health systems
g. Legal systems
h. Technology providers (here it is important to distinguish the different roles and degrees of control exercised by **social media and gaming companies**; operating system and device vendors; and telecommunications companies)
i. Multinational organisations focussed on childhood development and wellbeing (e.g. UNICEF)

Each of these systems has a role to play in impacting cyberbullying. At an individual level – for **victims**, **perpetrators** and other members of **online communities** – *nudge theory* can be used to design experimental conditions under which a specific

intervention can work. In a cyberbullying context, such experiments in turn depend upon **social media and gaming companies** to an unusual degree.

Developing a ToC which is a composite of two or more of the behavioural change theories listed would vary on the nature and scale of the intervention, the actors and institutions involved, the pathways and activities to map change and a series of indicators to evaluate change. A broad sequence of developing a ToC ex ante is illustrated. A similar sequence may be followed for developing a ToC post ante as well



Best Practices

Designing a Holistic Framework

Throughout the review, we have attempted to draw out the strengths and weaknesses of different evaluative approaches. Self-reporting, for instance, can provide highly personal indicators from the participant herself, yet is also prone to particular biases and tendencies. Pre and post tests are practical responses to evaluate the effectiveness of a programme, yet may also provide a "snapshot" that excludes long-term effects. The aim is not to dismiss any particular evaluative approach, but rather to develop an understanding of what they offer, what their limits are, and how they might best be integrated into a broader evaluative framework.

From the dozens of articles and case studies reviewed here, it is clear that there is no "silver bullet" for evaluation, no single indicator that reveals the presence or absence of behavioral change. Instead, the strongest case studies within this space developed a multimodal, multiscaled evaluation framework by combining many different measures into a cohesive model. When putting together an evaluative framework, organizations should aim for a holistic mix of indicators. Data collected at the "micro level" of the individual might be usefully supplemented with indicators that measure the "meso level" of a platform or social context. Harder quantitative indicators such as view counts, session durations, and sharing statistics could be mixed with "softer" qualitative forms of feedback such as narratives or interviews.

When considering how to implement an evaluation framework, one potentially helpful concept is that of calibration. Calibration suggests that forms of measurement are not fixed in place, but rather should be adapted to a particular time, place, and person. Calibration retains a core evaluative concept but adjusts it appropriately to a certain context. A small NGO with a dozen staff will have very different capabilities from a multinational corporation with a dozen offices. Participants in Thailand will inhabit a very different sociocultural milieu compared to participants in Japan. A programme for street children in Jakarta has considerations that a programme for international students in Singapore will not have. Calibration provides flexibility as a programme moves across contexts. If peer-based feedback, for instance, is critical for an evaluation framework, it should certainly be integrated into any programme, but its manifestation might be adjusted depending on time, budget, privacy considerations, technical expertise, or any number of other considerations. In this sense, calibration stresses that the key intents and concept of a framework should be retained, but recognises that it will also need to be tailored to an agency's particular needs.

Collecting and Analysing Data

asdf

Ethics in Evaluating
asdf

My Indicators

Your dynamically generated indicators are included on the next three pages. We've included a short description of each, some strengths and weaknesses, and the source literature where you can find out more about this indicator. If you'd like to add or change these indicators, just return to the "Evaluate It" web tool and generate another PDF.